Acceleration Method for Super-Resolution Based on Diffusion Models by Intermediate Step Prediction

Jichen Ma
Graduate School of FSE, Waseda University
Tokyo, Japan
majichen@ruri.waseda.jp

Hiroshi Watanabe
Graduate School of FSE,Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*— In this paper, we propose a new method to improve the generation speed of single-image super-resolution models based on the diffusion model. We address the generation speed problem of super-resolution models based on the diffusion model and propose an acceleration method by predicting intermediate steps. The proposed method is highly compatible with other sampling acceleration methods while maintaining high image quality and improving the efficiency and quality of the super-resolution task.**

Keywords*— *diffusion model, super-resolution, text-to-image, variational autoencoders, clip

I. Introduction

The widespread of social networking services (SNS) has led to an increased focus on photography. However, the degradation of image quality due to data compression has become an issue. To solve this problem, single-image super-resolution technology has become increasingly important, and neural network-based super-resolution methods in particular have attracted attention. Yet, there are problems with the possibility of low-resolution images corresponding to multiple high-resolution images, as well as deficiencies in the processing of high-frequency portions of conventional methods. Meanwhile, diffusion models have made remarkable progress in image generation, but their slow generation speed is still an issue.

In this study, we propose a new method to improve the generation speed of single-image super-resolution models based on the diffusion model. Stable diffusion, developed by Ludwig Maximilian University of Munich and released by Stability AI, is a publicly available text-to-image generation tool. Our method is highly compatible with other sampling acceleration methods, improving the efficiency and quality of the super-resolution task while maintaining high image quality.

II. Related Work

Denoising Diffusion Probabilistic Models (DDPM) [1] is a neural network model that has made significant progress in the image generation task and employs an architecture based on the U-Net model [2]. DDPM consists of two stages: the diffusion process and the inverse diffusion process. In the diffusion process, gaussian noise is added to an input image x_0 repeatedly T times to generate a series of images x_1, x_2, ..., x_T, where the final x_T is the gaussian noise image. In the inverse diffusion process, starting from x_T, the inverse diffusion process is performed through the diffusion model to finally recover the original image x_0.

SR3 [3] is a super-resolution technique based on the diffusion model. The network structure is based on a U-Net. In the training process, a low-resolution image is interpolated to a higher resolution and concatenated with a high-resolution image to which noise is added to serve as input to the model.

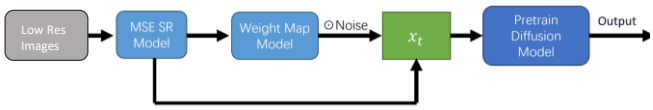

Fig. 1. Processing diagram of the proposed method.

The model predicts the noise added at the previous time, thus achieving super-resolution. In the generation stage, gaussian noise and the interpolated low-resolution image are concatenated and fed into the model, which then undergoes an inverse diffusion process to produce the final high-resolution image. This process allows SR3 to effectively achieve super-resolution.

III. Proposed Method

We propose a new acceleration method to improve the efficiency of super-resolution tasks based on the diffusion model. The method focuses on the prediction of the intermediate step x_t and aims to reduce the time of the generation process. We do not start the generation process from the gaussian noise in the diffusion model, instead begin it from the predicted intermediate step x_t. This reduces the generation time significantly. Figure 1 shows the processing diagram of the method proposed.

The prediction model for intermediate x_t consists of two parts. One is the MSE SR model, which is trained on the mean squared error (MSE) loss function, and the other is weight map model. The MSE SR model is used to generate super-resolution images. The weight map model is responsible for predicting the difference between the output of the MSE SR model and the real image. Let x_{sr} be the super-resolution image output by the MSE SR model. Let W be the output of the weight map model, where β is a matrix of the same shape and element values as W. Equation (1) shows how to compute the prediction x_t.

$$x_t = \sqrt{\bar{\alpha}_t}x_{sr} + clamp(\beta + W)\sqrt{1-\bar{\alpha}_t}\epsilon,\ \ \epsilon \sim \mathcal{N}(0, I),$$
$$clamp(\beta + W) = \max(0, \min(1, \beta + W)). \quad (1)$$

IV. Experiment

We train the MSE SR model and the weight map model on the Flickr-Faces-HQ (FFHQ) [4] dataset. The weight map model is a neural network with 97M parameters and is trained by the FFHQ dataset. For obtaining the difference between high resolution image and the output from MSE SR model, we use the contrast enhancement technique. To evaluate the effectiveness of the proposed method, super-resolution images are generated on the test set in combination with the DDPM sampling technique and the DDIM (Denoising

979-8-3503-5506-2/24/$31.00 ©2024 IEEE

Diffusion Implicit Models), which is an accelerated sampling technique. FID score is used as the evaluation metric. In our experiments, we use a trained SR3 model to directly predict an intermediate step image with x_t=300 and perform 300 iterations. Here, the original pre-trained DDIM requires 2000 iterations. Thus, the ratio of speedup is approximately 6.67 (≈2000/300) times. We investigate the impact of different values of the hyper parameter β on the quality and FID of the generated images. Further, we test the output of the MSE SR model without using the Weight Map and directly add noise to the output of the MSE SR model. The evaluation results are shown in Table I.

An acceleration experiment combined with the DDIM sampling method [5] is also conducted to compare the FID scores of the proposed method and DDIM. In this experiment, β is set to 0.8, x_t is predicted at time t=300. Number of Function Evaluations (NFE) refers to the total number of neural network computations to evaluate the advantage of the proposed method. The results are shown in Table II.

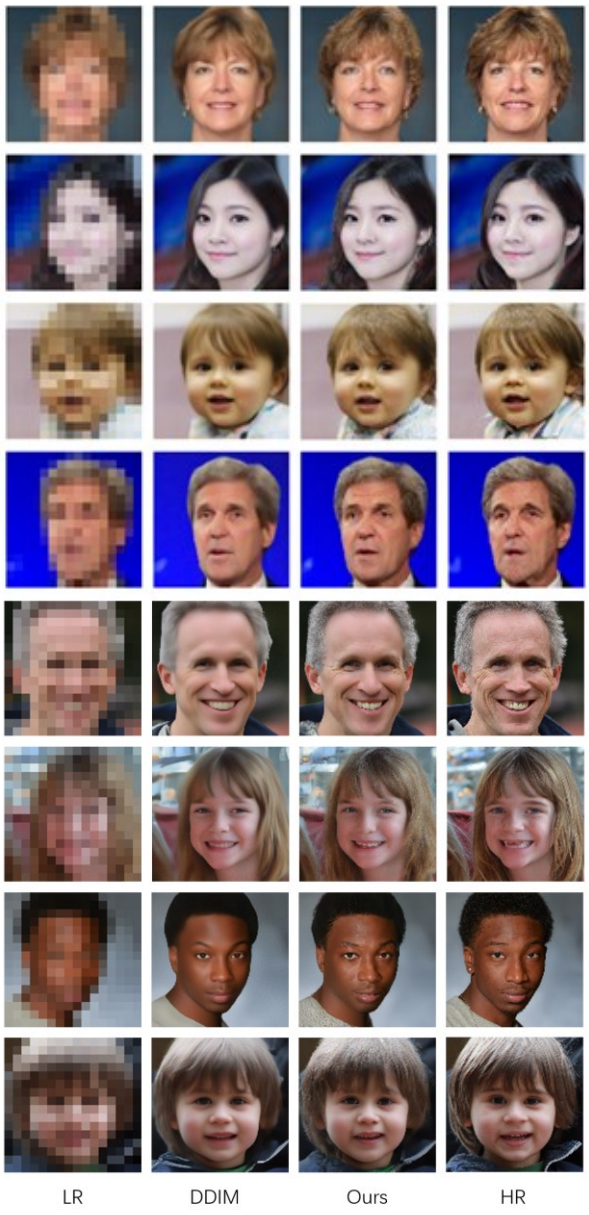

Fig. 2. Generated images by the proposed method compared with DDIM (Flickr-Faces-HQ dataset).

TABLE I EVALUATION RESULTS BY FID (↓)

Model	Hyper parameter β					
	0.0	0.2	0.4	0.6	0.8	1.0
w/ Weight Map	15.70	14.28	13.39	**13.22**	13.56	14.74
w/o Weight Map	17.20	17.14	16.80	16.38	15.64	14.74

TABLE II FID (↓) SCORE COMPARISON WITH DDIM

Model	NFE					
	8	10	15	20	25	50
DDIM	34.27	33.42	30.39	28.25	26.12	19.26
Ours(t=200)	18.10	21.92	20.61	21.59	17.76	18.84
Ours(t=300)	**16.27**	19.27	18.65	18.68	22.76	19.68

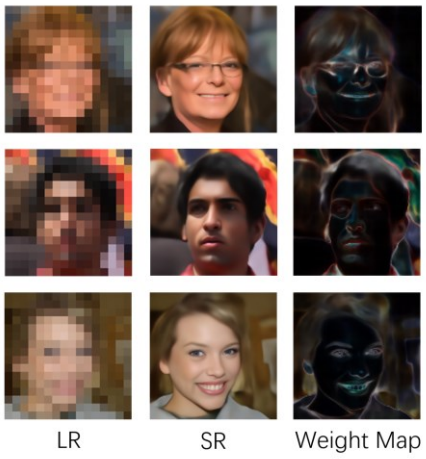

Fig. 3. Visualization examples of Weight Map. (left: low resolution, middle: super-resolution image by MSE SR model, right: Weight Map).

PSNR is not used as an evaluation metric in the super-resolution task because blurred images tend to show better PSNR values. Figure 2 presents the super-resolution image generated by the proposed method when the NFE is 8, alongside the image generated by DDIM when the prediction time t is 300. Figure 3 shows an example of a visualization of the weight map.

V. CONCLUSION

We address the generation speed problem of super-resolution models based on the diffusion model and propose an acceleration method by predicting intermediate steps. Evaluation experiments confirm that the proposed method can be used in combination with the DDPM and DDIM sampling methods. The combination with DDIM achieves excellent FID scores at low NFE values.

REFERENCES

[1] J. Ho, A. Jain, P. Abbeel, "Denoising Diffusion Probabilistic Models," arXiv preprint, arXiv:2006.11239, Jun. 2020.

[2] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," arXiv preprint, arXiv:1505.04597, May 2015.

[3] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, M. Norouzi, "Image Super-Resolution via Iterative Refinement," arXiv preprint, arXiv:2104.07636, Apr. 2021.

[4] T. Karras, S. Laine, T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4217-4228, 2021.

[5] J. Song, C. Meng, S. Ermon, "Denoising Diffusion Implicit Models," arXiv preprint, arXiv:2010.02502v4, Oct. 2020.