High-Resolution and Quality Settings with Latent Consistency Models

Steven Chen
Old Dominion University

Junrui Zhang
Old Dominion University

Rui Ning
Old Dominion University

COVA CCI Spring 2024 Undergraduate Cybersecurity Research Program

High-Resolution and Quality Settings with Latent Consistency Models

Steven Chen[1], Junrui Zhang[2], and Rui Ning[3]

[1]Department of Computer Science, Old Dominion University
[2]Commonwealth Cyber Initiative Coastal Virginia

April 13, 2024

Abstract

Diffusion Models have become powerful generative models which is capable of synthesizing high-quality images across various domains. This paper explores Stable Diffusion and mostly focuses on Latent Diffusion Models. Latent Consistency Models can enhance the inference with minimal iterations. It demonstrates the performance in image in-painting and class-conditional synthesis tasks. Throughout the experiment different datasets and parameter configurations, the paper highlights the image quality, processing time, and parameter. It also discussed the future directions including adding trigger-based implementation and emotional-based themes to replace the prompt.

Keywords Diffusion, enhance, Models, quality, inference, synthesis, configurations

1 Introduction

Diffusion Models are generative models, that are used to generate data similar to the data on which they are trained. In recent years, it has gained a lot of attention and remarkable results in different domains. Particularly Stable Diffusion which is a text-to-image model[4]. Stable Diffusion uses something called the Latent Diffusion Model on 512x512 images from a subset of datasets. These diffusion models are probabilistic models that aim to learn the reverse process of a fixed Markov Chain of Length or to learn a data distribution by gradually denoising a normally distributed variable[4]. Stable Diffusion also has something called image-to-image[3], which this paper will talk about how to utilize image-to-image and constantly generate them, so the user can see every generated image and movement.

2 Related Work

Diffusion Models have achieved significant success in high-quality image synthesis tasks [1]. It leverages many inferences and the reverse diffusion process to generate high-quality image samples by gradually adding noise in the reverse directions until the original signal is reconstructed. This model has a strong performance and offers a promising approach to generating realistic images [1].

Latent Consistency Models enable swift inference with minimal steps on any pre-trained LDMs including Stable Diffusion. It is designed to predict the solution of such ODE in latent space, mitigating the need for numerous iterations and allowing rapid, high-fidelity sampling[2].

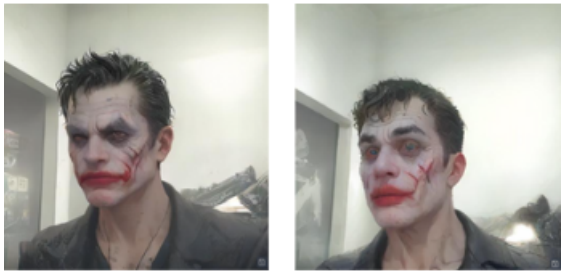
Latent Diffusion Models is done by decomposing the image formation process into the sequential application of denoising autoencoder. It achieves a new state-of-the-art score for image inpainting and class-conditional image synthesis and highly competitive performance on various tasks[4].

Image-to-Image The context of SDEdit is to transform an input image into an output image based on the prompt-given tasks. This process involves using a model to understand the content of the prompt for the given image and generate a corresponding output based on the default prompt and negative prompt[3].

3 Results

We will use different datasets to identify the output we want. In the LCMs they have already trained their dataset which is the LCM Dreamshaperv7 which is the fine-tuned version of Stable-Diffusion v1-5 with only 4,000 train-

ing iterations[2]. In the LCMs they used the ControlNet-canny checkpoint. This may be a good checkpoint because it is a monochrome image with white edges on a black background[5]. However, the output image is not detailed enough. The ControlNet scribble is a similar checkpoint as the canny but has better quality because it is a hand-drawn monochrome image with a white outline on a black background[5]. The below image shows the two checkpoints of the same diffusion setting and prompt. But the output seems a little different.



Scribble Checkpoint
Step:4
Strength: 50
controlnet: 32-100

Canny Checkpoint
Step:4
Strength: 50
controlnet: 32-100

Figure 1: Scribble Checkpoint vs Canny Checkpoint

As you can see, there are various variations between the two checkpoints for the Joker prompt. The one with the Joker-like nature and the scrawl checkpoint appears to be more detailed. Conversely, the other one appears to record an excessive amount of data and exclude certain Joker-related elements from the checkpoints.
The steps, strength, and ControlNet Start-End are set at those values because the more realistic the results will be, the lower the steps and strength. If the value is too high, on the other hand, the output will generate slowly and its quality will resemble a plain image of your prompt rather than a combination of the input image and prompt. Figure 2 shows the configuration of steps and strength from low to high.

As you can see, it grows more joker-like as the Step and Strength increase. But it also loses some of the original input's detail. As you can see, as the steps and strengths increase, the bottle that the dataset is holding gradually vanishes. Additionally, when the step and strength are high, it generates more slowly. This has an impact on image production time as well. The image production time can reach a maximum of 0.45 seconds at high steps and low strengths, and a minimum of 0.08 to 0.1 seconds at low steps and strengths.

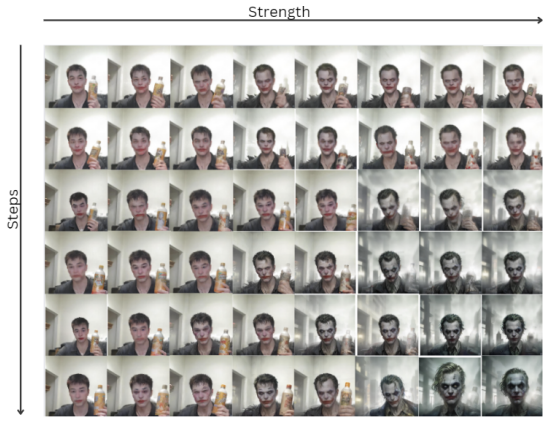


Figure 2: Difference low to high in steps and strengths

Regarding the start and finish controlnet scales. Anytime you shorten the processing time and compromise quality between 0.1 and 1.0, the processing time is shortened. Therefore, we set it to 1.0 for the controlnet end and 0.32 for the controlnet start.

4 Future Work

We will delve into the area of adding a trigger to the image in later work. For instance, the diffusion output image's character will vanish and just the input image's background will be visible if a particular yellow glass acts as the trigger for the diffusion. However, things will carry on normally if the user is not wearing the glasses. Furthermore, we will record the user's feelings and facial expressions in order to deduce the theme of the photographs rather than relying on a prompt. In this manner, the user needs just alter the phrase to alter the theme of the output image they don't need to alter the prompt.

5 Conclusions

Diffusion models have gained a lot of success, particularly Stable Diffusion, which has led to remarkable strides in various domains, especially in high-quality image synthesis tasks[4]. Latent Consistency Models (LCMs) enhance the efficiency of inference and enable high-fidelity sampling with minimal steps[2]. As we look toward the future, the exploration of additional features such as trigger-based image alterations and emotion-based theme deduction promises exciting avenues for further research and development. By incorporating user interactions and feedback, future iterations of diffusion models can evolve to provide more intuitive and personalized image generation experiences, ultimately

pushing the boundaries of what is achievable in synthetic image creation.

Diffusion models have gained a lot of success in recent years, especially Stable Diffusion, which led to high-quality image synthesis tasks[4]. Latent Consistency Models (LCMs) improve the diffusion models by enhancing the efficiency of inference and enabling high-fidelity sampling with minimal steps[2]. As we continue to explore the future with additional features such as trigger-based images and emotional-based themes for further research and development. By implementing user interaction, the diffusion models can become a more intuitive and personalized image-generation experience.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[2] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[3] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[5] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.