Post-processing Based Image Coding via Stable Diffusion

Luoxu Jin† Taiga Hayami† Hiroshi Watanabe† Takeshi Chujoh‡
Tomoko Aono‡ Yukinobu Yasugi‡ Sujun Hong‡ Zheming Fan‡
Tomohiro Ikai‡
† Waseda University ‡ Sharp Corporation

Abstract: Diffusion model has made remarkable results in the text-to-image field in recent years. In this work, we propose a post-processing and pre-trained diffusion model based image coding method to compress 512x512 size images to 3.4KB data size without any fine-tuning by VAE[10] encoder and quantized latent vector method, and by DDIM Inversion[9] post-processing, the images improve in LPIPS[7] metrics and outperform JPEG and WEBP compression methods with similar data size.

1 Introduction

Diffusion models have made state-of-the-art achievements in the field of generative models, and more and more researches have been using diffusion models to compress images in the field of image compression. Existing studies typically use a prior neural network to encode an image into a vector in the latent space, and then use a conditional diffusion model to generate image using the vector or reconstruction image $\hat{x}$ as a condition, which usually requires optimal training on a specific datasets[11, 5, 2]. Since Stable Diffusion[8] has been trained on millions of images, we believe that it has the generalisation ability to generate the majority of images and has the ability to zero-shot compress images without fine-tuning. In this work we use the VAE[10] module of the pre-trained Stable Diffusion model to encode the image x into vectors in the latent space, which are then quantised and transmitted with lossless encoding. The reconstructed images $\hat{x}$ from the VAE are usually artifacts and noisy, we use the DDIM Inversion method to regenerate the image $\hat{x}$. This results in a regeneration image that approximates the ground truth at the time of sampling, Additionally, regenerated image able to removes noise and artefacts from the decode image $\hat{x}$.

2 Related Work

2.1 Stable Diffusion

In Stable Diffusion model, through the VAE encoder $E_\theta(X)$ image x_0 will be $z_0 = E_\theta(x_0)$ encoded to the latent space. The forward process of z_0 randomly adds Gaussian noise ϵ to the distribution z_t, after which the model learns the denoising process by predicting the noise using $\bar{\epsilon} = \epsilon_\theta(z_t, t)$. The inference starts from z_T to the z_0 distribution using the denoising model $\epsilon_\theta(z_t, t)$. And it will be decoded back again by the VAE decoder $D_\theta(Z)$ through $\hat{x_0} = D_\theta(z_0)$.

2.2 Denoising Diffusion Implicit Models

DDIM is a deterministic sampling method, which removes random resampling compared to the DDPM[3] sampling method. By given a random Gaussian noise distribution z_T and conditional input c, it will be progressively denoised for T steps until z_0 using denoise model $\epsilon_\theta(z_t, t, c)$, and this process is reproducible deterministically.

3 Method

Taking an (3,512,512) RGB image with format (C, H, W) as an example, the downsampling factor of Stable Diffusion's VAE module is 8, which means that a (3,512,512) image will be compressed to (4,64,64) latent vector, which are usually stored in float16 format, and the amount of data is 64*64*4*16 = 32KB. With the quantization of the latent vecotr by paletizing and dithering[1], the latent vector will be compressed to 64*64*8 + 64*4*8 = 3.4KB.

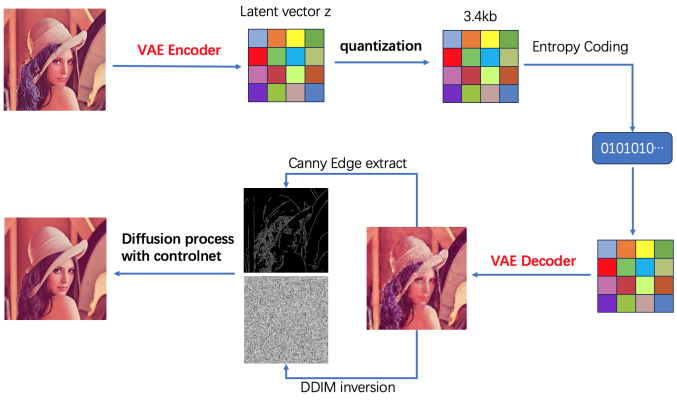

Figure 1: Overview of compression method.

Despite the fact that the image is compressed at a very low bit rate, the quality of the reconstructed image is very poor. We propose to regenerate the image using the DDIM Inversion method:

$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} z_t + \sqrt{\alpha_{t+1}} \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \varepsilon_\theta \left(z_t, t, c \right) \tag{1}$$

Specifically, we will renoise [Equation 1] the image back to Gaussian noise distribution z_T, and in order to reduce the effect of classifier-free guidance[4], we set the hyper-parameter w to 1 as follow:

$$\bar{\epsilon}_\theta(z_t, t, c) = \ w \cdot \epsilon_\theta(z_t, t, c) + (1 - w) \cdot \epsilon_\theta(z_t, t)$$

Similarly, we keep the same parameter settings when using DDIM sampling to regenerate the image. In order to reduce the perceptual loss of the generated image, The canny edge is extracted from the quantised decoded image $\overline{x}$ and is

used as the Controlnet[6] plug-in conditional bootstrap image generation, noting that the coefficient of the conditional input canny is set to a small value.

4 Experiment

We test the performance on several images and evaluate their SSIM metrics and LPIPS metrics. From the experimental results, It could be observed that JPEG images have many artefacts in the case of quality of 1 and low bit rate, but WEBP and ours propose images still have relatively great subjective perception. This subjective perception is also reflected in the LPIPS metric.

Figure 2: Example of Pepper Generation

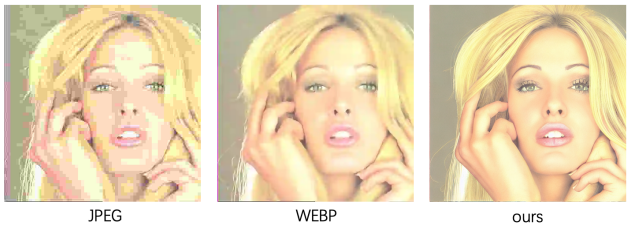

Figure 3: Example of Tiffany Generation

We use **bold** font to mark the best results and the second best results are underlined. From the results in the table, it can be seen that the LPIPS metrics are improved after post-processing the images using the DDIM inversion method and the LPIPS metrics are better than the JPEG and WEBP compression methods and the SSIM metrics are better than the JPEG compression method.

	SSIM ↑	LPIPS ↓	Size
Pepper-JPEG	0.482	0.644	3.4kb
Pepper-WEBP	**0.645**	0.428	4.1kb
Pepper-w/o Inversion	0.562	0.338	3.4kb
Pepper-w/ Inversion	0.606	**0.27**	3.4kb
Tiffany-JPEG	0.635	0.668	3.6kb
Tiffany-WEBP	**0.744**	0.45	3.6kb
Tiffany-w/o Inversion	0.652	0.343	3.4kb
Tiffany-w/ Inversion	0.668	**0.276**	3.4kb

Table 1: Testing performance metrics on sample images

5 Conclusion

In this study, we tried to use Pre-trained Stable Diffusion to compress images, and used a post-processing method based on DDIM Inversion to improve the quality of compressed low-bitrate images. It is true that Diffusion has powerful decoding capabilities, but the decompression speed is relatively slow due to the characteristics of the diffusion model. In future research, improving decoding speed and smaller bpp metric is a promising research topic.

References

[1] Matthias Bühlmann. Stable Diffusion Based Image Compression. In *Blog*, 2022.

[2] Noor Fathima Ghouse, Jens Petersen, Auke Wiggers, Tianlin Xu, and Guillaume Sautière. A residual diffusion model for high perceptual quality codec augmentation, 2023.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[5] Emiel Hoogeboom, Eirikur Agustsson, Fabian Mentzer, Luca Versari, George Toderici, and Lucas Theis. High-fidelity image compression with score-based generative models, 2023.

[6] Anyi Rao Lvmin Zhang and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023.

[7] Alexei A. Efros Eli Shechtman and Oliver Wang Richard Zhang, Phillip Isola. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 2018.

[8] Dominik Lorenz1 Patrick Esser Robin Rombach, Andreas Blattmann1 and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022.

[9] Kfir Aberman Yael Pritch and Daniel Cohen-Or Ron Mokady, Amir Hertz. Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *CVPR*, 2023.

[10] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In *axXiv*, 2018.

[11] Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models, 2023.

School of Fundamental Science and Engineering
Waseda University, Shillman Hall 401, 3-14-9 Okubo
Shinjuku Tokyo, 169-0072 Phone: 03-5286-2509
Fax:03-5286-3488