2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR) | 979-8-3503-5142-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/MIPR62202.2024.00107

Perceptual Image Compression via Stable Diffusion at Low Bitrate

Luoxu Jin
CSCE, Graduate School of FSE
Waseda University
Tokyo, Japan
kinryokyoku@akane.waseda.jp

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

***Abstract*—With the development of text-to-image model in the field of image generation, it is possible to generate high fidelity images using only short text. We investigate the ability to use these pre-trained text-to-image models applying to image compression tasks without any fine-tuning. We extract caption, spatial structure sketch, and a special embedding information from the image. The bit rate of these components can be reduced using difference compression techniques. We can reconstruct them back to the image using the pre-trained generative models at decoding process. We show this compression algorithm outperforms the learned image compression method in perceptual metrics FID and KID at very low bit rate.**

***Index Terms*—Diffusion model, Image compression**

I. Introduction

Image compression is an important task in the field of image processing. The most widely used compression algorithms today are still hand-crafted. Recently, promising results have been achieved using neural networks to learn end-to-end image compression models. Neural network-based image compression model can obtain a compact latent representation for the ground truth. It is shown that neural network-based image compression outperforms traditional compression algorithms in terms of rate-distortion (R-D) performance [1], [2], [3]. Generative model-based image compression algorithm solves the artifacts and blurring problems associated with previous image compression method. In particular, Mentzer et al. [4] enhanced image compression performance by introducing adversarial loss, thus improving the R-D characteristic compared to previous methods. Yang et al. [5] utilized conditional diffusion model as a decoder, the encoded image is treated as conditional information into the diffusion model. This approach achieves superior performance compared to previous GAN-based models in lossy image compression tasks.

Although these image compression models can provide high-fidelity image compression, they usually require end-to-end training using R-D loss functions on datasets containing a large number of images. In the field of text-to-image generation, diffusion models have been shown to generate photo realistic images guided by text descriptions. A natural question is that if these powerful image generation models can generate a wide variety of high-quality images, then they should also be able to apply to other image processing tasks. In Stable Diffusion [6], as long as the conditional information such as prompt, seed, model version are kept the same, then only these conditional information need to be transmitted. The decoder can surely reconstruct back to the input image by utilizing these conditional information. Therefore, we believe that this property can be exploited to apply Stable Diffusion to image compression task. Our approach is very close to the one of PICS [7]. Both of them aim to compress images using pre-trained text-to-image model. PICS demonstrates the capability to compress images at very low bit rate. However, PICS cannot faithfully recover the color and style information of the ground truth, as shown in Fig. 1.

(a) PICS reconstruct (b) Ours reconstruct (c) Ground Truth

Fig. 1: Comparison of reconstructed images, (a) PICS, (b) Ours, (c) Ground Truth.

Inspired by previous works [7], [8], we propose a special embedding process to recover faithful texture information and use existing image compression models to compress this special embedding. In this paper, we try to extract three different components, prompt, sketch, embedding from images and compress them with different encoding methods respectively. At the decoding phase, the three expanded components are fed into pre-trained text-to-image model to reconstruct the image. In summary, our contributions are as follows.

1) We use a pre-trained text-to-image model for the image compression task, preserving input image texture and style by optimizing a special embedding. This compression method maintains the fidelity of the compressed image even at very low bit rate.

2) We propose to use edge information as a constraint to accelerate null text inversion optimization and improve the generation quality.

2770-4319/24/$31.00 ©2024 IEEE
DOI 10.1109/MIPR62202.2024.00107

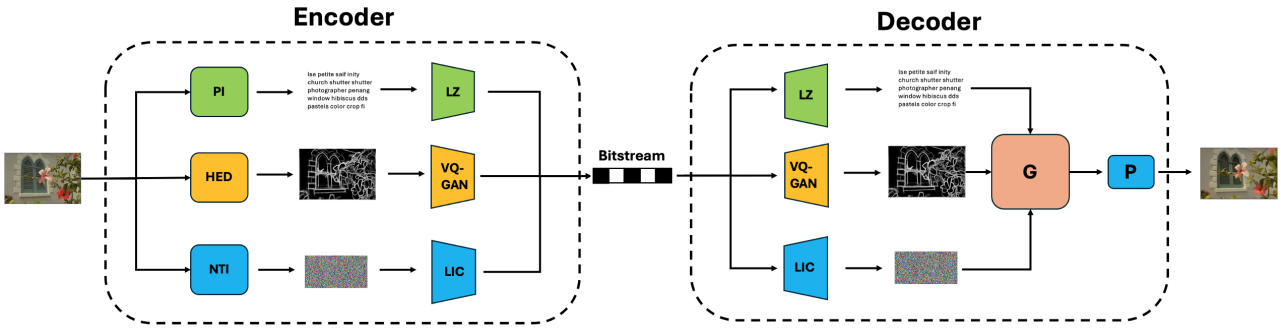


Fig. 2: The overall coding architecture that extracts prompt, edge, and optimized special embedding from ground truth.

II. Related Work

A. Diffusion Models

The diffusion model was first successfully used in unconditional image generation by Ho et al. [9]. Later Dhariwal et al. [10] propose a classifier guidance to enable diffusion model to conditionally generate images. This scheme outperforms the GAN [11] model in fidelity. Stable Diffusion is trained on a dataset of millions of images [6], it can generate semantically compatible images from the text. ControlNet [12] is a plug-in for the Stable Diffusion that utilizes edge maps, key-point maps, depth maps, etc., as conditional inputs. It ensures that the structure of the image generated by Stable Diffusion is consistent with the conditional inputs.

B. Null-text Inversion for Editing Real Images

Mokady et al. [8] propose a method to address the inconsistency caused by classifier-free guidance, which arises when reconstructing images after DDIM inversion [10], [13]. They suggest optimizing a special embedding to resolve this problem, which is then used for image editing tasks. We propose to transfer this idea to the image compression task. The null text optimization requires ground truth for the DDIM Inversion process, which is not reasonable for the image compression task. We investigate the use of an arbitrary image generated by the text-to-image model as a starting point and leveraging edge information to constrain the optimization process.

III. Proposed Method

A. Overall Coding Scheme

In this section, we introduce the overall framework and strategies for coding the different components shown in Fig. 2. First, we use the Prompt Inversion (PI) [14] method to extract the text corresponding to the image and apply lossless compression. This method searches the CLIP latent space of the image to find the text embedding with the highest cosine similarity, and then projects the text embedding back into the prompt, capturing semantic information better than human captions.

Reconstructing the image using only the prompt fails to restore the structural information of the ground truth. By adding the ControlNet [12], Stable Diffusion can constrain the structure of the generated image [7]. We extract Holistically-nested Edge Detection (HED) map from the ground truth and later compress the HED map using the VQ Compression [15] method. VQ Compression uses a pre-trained VQGAN [16] model, which compresses the image into a feature map consisting of codebook vectors. Each vector can be indexed to represent a particular vector in the codebook, then be recovered by transmitting the index value of the codebook and decoded to the image by the VQGAN decoder. For the pre-trained VQGAN model with the codebook size of 1024, the codebook is clustered to 255 using K-means algorithm so that it can be represented by a value of type uint8 to reduce bit-rate.

Although the semantic and structural information of the generated image is controlled by prompt and HED map, the MSE loss between the pixels of the generated image and ground truth is still large that similar colors cannot be generated. Specifying the color prompt or improving the bit rate of the HED map will not solve the issue. To address this problem, we propose to use the Null Text Inversion method [8] to generate color-consistent images. For each trajectory in the diffusion backward process, we use a special null text embedding as an optimization object to make the trajectory of the generated image as close as possible to the one of the ground truth. A null text will be encoded by CLIP text encoder into null text embedding in (N, D) shape. In order to compress this embedding, we use the existing learned image compression method [3] to reshape the embedding into an RGB image format $(3, H, W)$ and then compress it. Since the learned image compression model cannot be applied to floating-point compression directly, we optimize the compressed embedding combining compression and optimization together. Moreover, to improve the quality of the image, the compressed prompt and HED map are used as a condition for further post-processing of the generated image using ControlNet Tile method [12].

B. Null Text Inversion

For an image x_0, it is first encoded into the latent vector z_0, then artificial noise ϵ is added to $\{z_1, ..., z_T\}$. The noise

predictor ϵ_θ learns to remove the added artificial noise by training with the following equation:

$$L = E_{\epsilon \sim N(0,I), t \sim Uniform(1,T)} ||\epsilon - \epsilon_\theta(z_t, t, C)||. \quad (1)$$

Using DDIM [17] sampling formula, the noise predictor gradually remove the noise until z_0. We have the sampling formula:

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \epsilon_\theta(z_t, t, C). \quad (2)$$

Based on the assumptions of the ODE process, we can also obtain the reversal formula for DDIM Inversion [10] get z_t to z_{t+1}.

$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \epsilon_\theta(z_t, t, C). \quad (3)$$

The noise predictor $\epsilon_\theta(z_t, t, C)$ is computed by the classifier-free guidance [13], where the model predicts the unconditional and conditional noise, after which it performs linear combination of the two predicted noises. Here ω is the hyper-parameter of the computation.

$$\overline{\epsilon_\theta}(z_t, t, C) = \omega \cdot \epsilon_\theta(z_t, t, C) - (1 - \omega) \cdot \epsilon_\theta(z_t, t, \emptyset), \quad (4)$$

where the empty set $\emptyset = E_{clip}(\text{“”})$ as the embedding of null text is encoded by the pre-trained CLIP [18] encoder. The null text embedding value will be influenced by the noise predictor $\overline{\epsilon_\theta}(z_t, t, C)$, which is noise prediction result through classifier-free guidance formula [13]. According to the DDIM sampling formulation, the value of z_{t-1} is calculated from the value of z_t with the $\overline{\epsilon_\theta}(z_t, t, C)$ prediction noise [17]. Thus, the value of z_{t-1} can be optimized by adjusting the prediction noise, which is in turn optimized by modifying the value of the null text embedding.

The overall embedding optimization process is shown in Fig. 3. First we use a compressed version of prompt and edge to generate an image via Stable Diffusion, and then use DDIM Inversion [10] to add deterministic noise to get $\{z_1^*, ..., z_T^*\}$ trajectories. We record the added noise and copy it to the ground truth to get the $\{z_1, ..., z_T\}$ trajectory as well. Taking the trajectory $\{z_0, z_1, ..., z_T\}$ as the optimization target, starting from z_T^* given the timestamp $t = T, ..., 1$ our optimization formula is shown as follows:

$$\mathcal{L}_{RD} = R(y) + \lambda \cdot D(z_{t-1}, z_{t-1}^*(\epsilon_\theta(z_t^*, t, \hat{\emptyset}, C))). \quad (5)$$

We optimize z_{t-1} and z_{t-1}^* feature map using MSE and Kullback–Leibler divergence as a loss function.

$$D = D_{KL}(z_{t-1} || z_{t-1}^*) + D_{MSE}(z_{t-1}, z_{t-1}^*), \quad (6)$$

where $\hat{\emptyset}$ represents the null text embedding of the compressed version by the LIC model [3] and C denotes the conditions prompt and edge, while $R(y)$ represents the bit rate required by the LIC model to compress the null text embedding [3], hyper-parameter λ is used to control the bit rate. For each z_t^* to z_{t-1}^* denoising process, we combine additional ControlNet to predict the next feature map z_{t-1}^* as shown in Fig. 3.

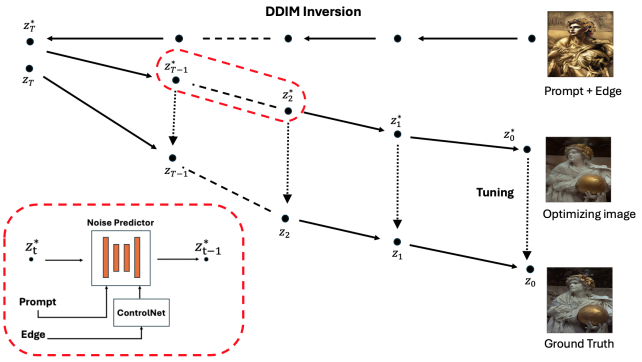


Fig. 3: DDIM Inversion obtains z_T^* and then optimizes the target trajectory $\{z_0, ..., z_T\}$ with z_T^*. ControlNet is leveraged with an edge component to enhance the optimized results.

IV. Experiment

A. Experiment Settings

1) Implementation details: We use the open-source Stable Diffusion version 1.5 [6] as the pre-trained model, and the official HED checkpoint by ControlNet [12]. For training, we use Adam with a learning rate of 0.01 as the optimizer and set the DDIM sampling step to 25. It takes 1 hour and 30 minutes to optimize an image on an RTX A5000 graphics card.

2) Datasets and Evaluation: We use the Kodak [19] and the CLIC2021 [20] test set for evaluation. We evaluate the compression performance using five metrics, PSNR, SSIM, LPIPS [21], FID [22], and KID [23] at 0.05∼0.06bpp.

B. Experimental Results

1) Qualitative results: We give the visualization examples of reconstructed images with different compression algorithms shown in Fig. 4. Cheng et al. [3] brings blur at low bit rate and the sharp details of the image are lost. PICS [7] is capable of reconstructing sharp image details and keeping the high-level spatial information unchanged. However, the texture and

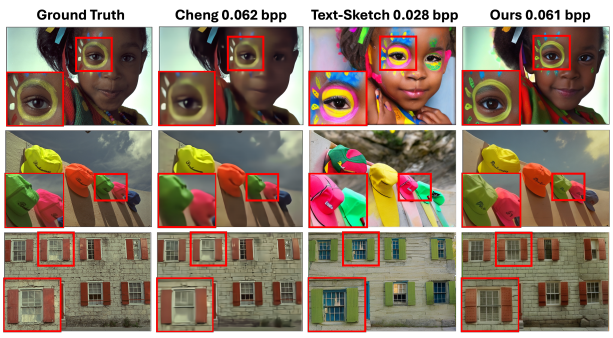


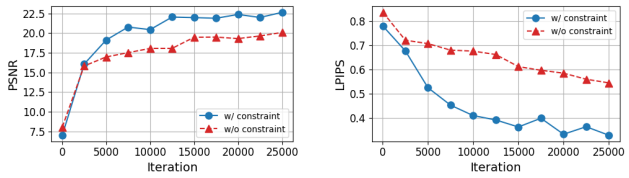
Fig. 4: Qualitative results for the Kodak dataset using different compression algorithms.

TABLE I: Evaluation Result of Kodak Images

	Bpp	MS-SSIM↑	PSNR↑	LPIPS↓	FID↓	KID↓
Ours	0.061	0.618	18.85	**0.468**	**118**	**0.148**
Cheng [3]	0.062	**0.881**	**26.58**	0.502	180	0.150
PICS [7]	0.028	0.307	11.54	0.644	143	0.158

TABLE II: Evaluation Result of CLIC Images

	Bpp	MS-SSIM↑	PSNR↑	LPIPS↓	FID↓	KID↓
Ours	0.058	0.639	17.96	0.442	**136**	**0.167**
Cheng [3]	0.058	**0.904**	**26.34**	**0.435**	179	0.190
PICS [7]	0.026	0.296	10.07	0.654	181	0.177



(a) PSNR (b) LPIPS

Fig. 5: Ablation experiments demonstrate learning convergence with and without the use of edges as constraints.

color information of the image are completely modified. Our method generates images with sharp details and maintains the texture and color information of the ground truth at low bit rate. Another property of this compression method is that due to random sampling, each generated image may have slight differences in details. This property allows us to generate images with different details based on the ground truth.

2) Quantitative Results: We first compare the performance of the PICS as baseline model. The experimental results are shown in Table I and Table II, the **bold** results show the best and the underline one show the second. From Table I and Table II, we observe that our method improves on all metrics by introducing an optimized embedding of 0.03 bits per pixel (bpp) compared to PICS. Further our method outperforms existing neural network image compression model by Cheng at the same bit rate in terms of the image fidelity metrics FID and KID.

3) Ablation Study: We used the PSNR and LPIPS metrics to evaluate ablation study experiment, as shown in Fig. 5. Experiments demonstrate that using edges as a constraint significantly improves the results.

V. Conclusion

In this paper, we propose a novel perceptual image compression method. We use a pre-trained text-to-image model for the image compression task and optimize a special embedding to control texture and color. Experimental results show this perceptual compression model maintains good fidelity score even in the case of less than 0.1 bpp. However, the image encoding process is time-consuming, which limits its practicality in scenarios with constrained computational power. Consequently, it is more suitable for environments with sufficient computational resources and limited bandwidth.

References

[1] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *International Conference on Learning Representations*, Jul. 2017.

[2] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, Feb. 2018.

[3] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.

[4] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," in *Advances in Neural Information Processing Systems*, Dec. 2020.

[5] R. Yang and S. Mandt, "Lossy image compression with conditional diffusion models," in *Advances in Neural Information Processing Systems*, Dec. 2023.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022.

[7] E. Lei, Y. B. Uslu, H. Hassani, and S. S. Bidokhti, "Text+ sketch: Image compression at ultra low rates," in *ICML 2023 Workshop on Neural Compression: From Information Theory to Applications*, Jul. 2023.

[8] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, Dec. 2020.

[10] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems*, Dec. 2021.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Dec. 2014.

[12] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *International Conference on Computer Vision (ICCV)*, Oct. 2023.

[13] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, Dec. 2021.

[14] Y. Wen, N. Jain, J. Kirchenbauer, M. Goldblum, J. Geiping, and T. Goldstein, "Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery," in *Thirty-seventh Conference on Neural Information Processing Systems*, Dec 2023.

[15] Q. Mao, T. Yang, Y. Zhang, S. Pan, M. Wang, S. Wang, and S. Ma, "Extreme image compression using fine-tuned vqgan models," *arXiv preprint arXiv:2307.08265*, 2023.

[16] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021.

[17] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, Jan. 2021.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, Jul. 2021.

[19] "Franzen, r. w. kodak lossless true color image suite. url https://r0k.us/graphics/kodak/."

[20] "Clic 2021: Challenge on learned image compression. url https://clic.compression.cc/2021/index.html."

[21] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[22] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, Dec. 2017.

[23] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *International Conference on Learning Representations (ICLR)*, Feb. 2018.