Extreme Image Compression Using Fine-tuned VQGANs

Qi Mao*, Tinghan Yang*, Yinuo Zhang*, Zijian Wang*,
Meng Wang†, Shiqi Wang†, Libiao Jin* and Siwei Ma‡

*Communication University of China †City University of Hong Kong ‡Peking University
{qimao, yangtinghan, yinuozhang, wangzijian, libiao}@cuc.edu.cn
{ mwang98-c, shiqwang}@cityu.edu.hk
swma@pku.edu.cn

2024 Data Compression Conference (DCC) | 979-8-3503-8587-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/DCC58796.2024.00028

Abstract

Recent advances in generative compression methods have demonstrated remarkable progress in enhancing the perceptual quality of compressed data, especially in scenarios with low bitrates. However, their efficacy and applicability to achieve extreme compression ratios (< 0.05 bpp) remain constrained. In this work, we propose a simple yet effective coding framework by introducing vector quantization (VQ)–based generative models into the image compression domain. The main insight is that the codebook learned by the VQGAN model yields a strong expressive capacity, facilitating efficient compression of continuous information in the latent space while maintaining reconstruction quality. Specifically, an image can be represented as VQ-indices by finding the nearest codeword, which can be encoded using lossless compression methods into bitstreams. We propose clustering a pre-trained large-scale codebook into smaller codebooks through the K-means algorithm, yielding variable bitrates and different levels of reconstruction quality within the coding framework. Furthermore, we introduce a transformer to predict lost indices and restore images in unstable environments. Extensive qualitative and quantitative experiments on various benchmark datasets demonstrate that the proposed framework outperforms state-of-the-art codecs in terms of perceptual quality-oriented metrics and human perception at extremely low bitrates (≤ 0.04 bpp). Remarkably, even with the loss of up to 20% of indices, the images can be effectively restored with minimal perceptual loss.

1 Introduction

With the ever-increasing amount of visual data being generated at an unprecedented pace, the demand for highly efficient and effective compression algorithms has become increasingly crucial. However, under minimal network bandwidth, the signal-oriented traditional image/video compression codecs (*e.g.*, BPG [1], and the latest video coding standard VVC [2]) inevitably adopt large scalar quantization steps, resulting in a significant loss of texture information with unacceptable blurring and blocking artifacts.

To bridge the gap of shallow bitrate scenarios, recent image coding methods [3–12] leverage the power of generative models [13, 14] to reconstruct the human-favored decoded image/video. There are currently two main perspectives in such generative compression: the first [3,5,7,8] involves using a conditional GAN [15] as an additional distortion term to optimize deep learning-based end-to-end neural codecs. This category of methods enhances the reconstruction of texture details in the decoded image

2375-0359/24/$31.00 ©2024 IEEE
DOI 10.1109/DCC58796.2024.00028

through adversarial training. However, their effectiveness in achieving high compression ratios, especially below 0.05 bpp (bits per pixel), remains limited. Another line [4, 9–12, 16] aims to compress images into compact feature representations at the encoding end and generate decoded images with the aid of GANs, achieving visual pleasing reconstruction even at extremely low compression ratios. However, without additional training, such approaches [9–12, 16] have difficulty reconstructing the original image with large semantic information gaps against the training dataset. For instance, codecs optimized for face images may not perform well on natural scenario images. As such, their practical use in extremely low-bitrate scenarios is hindered by poor generalization ability.

Recently, vector quantization (VQ)–based generative models [17, 18], which utilize discrete image representations, have been well used in image generation tasks. Despite their success in other generation fields, VQ-based generative models have received relatively less attention in the image compression domain. In this work, we reveal a surprising finding: The learned codebook of the VQGAN model [17], which has been trained on a large-scale dataset, exhibits a powerful and robust representational capacity. Accordingly, we propose a simple yet effective coding framework by directly integrating VQ-indices compression into the VQGAN model, which yields a significant advancement in the capability of extreme image compression and generalizability across various semantics and resolutions of images. In particular, the VQ-indices map is obtained by identifying the nearest sample in the learned codebook, which is then encoded into a bitstream using arithmetic compression. To enable variable bitrates, we propose clustering a pre-trained large-scale codebook using the K-means algorithm, resulting in a series of smaller codebooks. As such, the image can be represented by various VQ-indices maps, enabling variable bitrates and different levels of reconstruction quality. Moreover, given the potential impact of bitstream loss on the decoding process within an unstable transmission environment, we propose to leverage the second-stage transformer of the VQGAN model to predict missing indices. It can be effectively predicted based on context indices that adhere to the underlying discrete distribution, thereby effectively circumventing image reconstruction failure due to the loss of bitstreams.

The main contributions in this paper can be summarized as follows:

- We present a simple yet effective coding framework that utilizes the VQGAN model in developing a novel extreme image compression framework.
- We propose a K-means clustering approach to compress the large-scale codebook into a smaller new codebook, which enables variable bitrates and levels of reconstruction quality within our framework. Furthermore, a second-stage transformer is leveraged to predict missing indices, enhancing the resilience of the proposed framework in unstable transmission.
- Both qualitative and quantitative results demonstrate that the proposed framework achieves a significant improvement compared to state-of-the-art codecs about both perceptual quality metrics and human-viewed under extremely low bitrates ($\leq$ 0.04bpp). Notably, even in cases where up to 20% of indices are lost, the images can be efficiently restored with minimal perceptual loss, thanks to the collaboration with the generative transformer model.

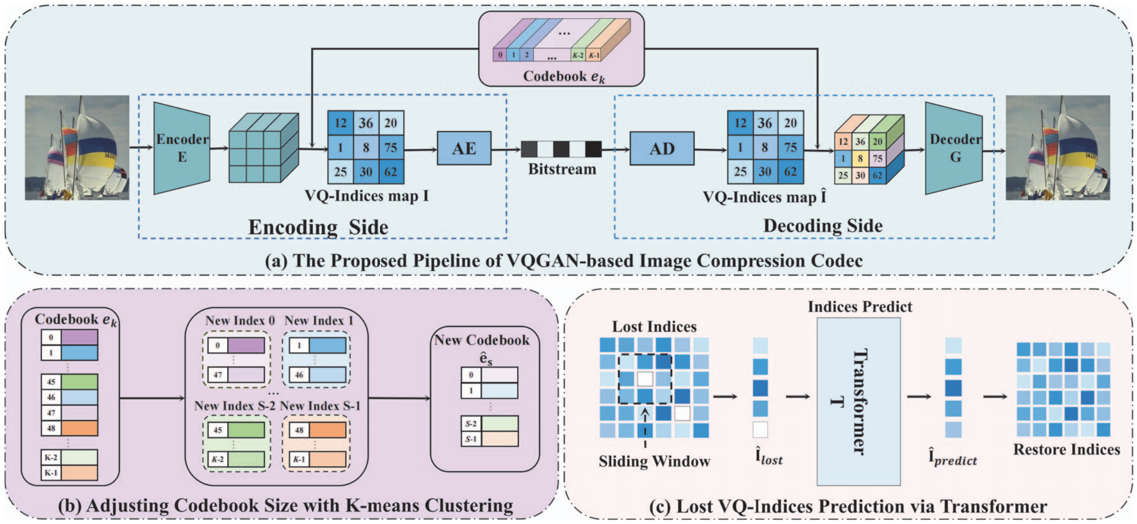

Figure 1: Overview of the proposed VQGAN-based image coding framework.

2 Proposed VQGAN-based Extreme Image Codec

In this work, we aim to compress images at ultra-low bit rates while maintaining the high-perceived quality of the reconstructed images. Our framework is built upon the discrete representations and the generative capacity of VQGAN models. As illustrated in Fig. 1, the entire framework consists of four key components:

- The encoder E: extract the input image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ into a latent representation $\mathbf{z} \in \mathbb{R}^{\frac{H}{M}\times\frac{W}{M}\times n_z}$.
- The codebook $\mathbf{e}_k \in \mathbb{R}^{n_z}, k \in 1,2,\cdots,K$: map the latent representation $\mathbf{z}$ into a sequence of VQ-indices and invert it back to quantized latent representation $\mathbf{z}_q \in \mathbb{R}^{\frac{H}{M}\times\frac{W}{M}\times n_z}$ through the nearest neighbor lookup. The proposed K-means clustering algorithm method compresses the large-scale codebook into a smaller one, enabling variable bitrates and varying reconstruction quality.
- The decoder G: synthesize the quantized latent representation $\mathbf{z}_q$ into a reconstructed image $\hat{\mathbf{x}} \in \mathbb{R}^{H\times W\times 3}$.
- The transformer T: predict the missing index based on the context indices.

On the encoding side, VQ-indices are compressed into the final bitstream using lossless compression techniques. On the decoding side, we decode the index sequences from the bitstream and convert the decoded VQ indices back into the quantized latent representation $\mathbf{z}_q$ by searching the codeword from the codebook. We detail the key components of the proposed framework below.

2.1 VQ-indices Compression

Unlike existing image compression methods [19,20] that adopt scalar-quantization, we leverage the power of VQ to construct a codebook of representative vectors for latent representations $\mathbf{z}$. In particular, the codebook $\mathbf{e}_k$ is then used to encode the latent representation by replacing each position of a vector with the **index** of the closest representative vector by Euclidean distance, which results in a highly compressed version of the latent representation with minimal loss of quality:

$$\mathbf{I}_{ij} = \underset{k \in 1,2,\cdots,K}{\operatorname{argmin}} \left\|\mathbf{z}_{ij} - \mathbf{e}_k\right\|^2, \tag{1}$$

where i and j denote the position of vector $\mathbf{z}_{ij}$ in latent representation $\mathbf{z}$, $\mathbf{I}_{ij}$ represents its corresponding index, and K indicates the size of codebook. As demonstrated in Fig. 1(a), **an input image $\mathbf{x}$ can be efficiently represented by VQ-indices map $\mathbf{I} \in \mathbb{R}^{\frac{H}{M}\times\frac{W}{M}}$, which significantly reduces the data amount.** Then, we adopt the widely used arithmetic coding to compress VQ-indices into bitstreams, further reducing the data size. After decoding the compressed bitstream, the reconstructed latent vectors $\mathbf{z}_q$ are generated by searching for their corresponding code words based on their indices. Finally, the reconstructed image $\hat{\mathbf{x}}$ is synthesized by the decoder G.

2.2 Adjusting Variable Bitrates

The quality of the codebook in our proposed framework plays a crucial role in determining the compression performance of the entire system. As such, we develop a rate control strategy that utilizes the K-means algorithm to **cluster the pre-trained large-scale trained codebook into the smaller one**, thereby enabling the flexible adjustment of compression bitrates by altering the size of the codebook. Subsequently, the newly generated codebook $\hat{\mathbf{e}}_s$ derived from the K-means clustering algorithm can serve as a starting point for further fine-tuning, enabling faster convergence during subsequent optimization while ensuring codebook quality. Finally, we can obtain new codebooks of varying sizes, each with a corresponding set of VQ-indices to represent the compressed image, as illustrated in Fig. 1(b).

2.3 Lost VQ-indices Prediction

In an unreliable transmission environment, lost packets may result in the loss of indices, which can cause incorrect decoding of bitstreams. However, our proposed image codec framework, which utilizes the VQ-indices transformer, can **accurately predict lost indices at the decoder end**. This approach enhances the robustness of the codec in dealing with unreliable network transmission and ensures that the decoded bitstreams remain accurate. We flatten the VQ-indices map $\hat{\mathbf{I}}$, and denote it as $[\hat{\mathbf{I}}_i]_{i=1}^N$, where N indicates the total length. Subsequently, the VQGAN's second-stage transformer is trained to predict the probability distribution of the next possible indices $p(\hat{\mathbf{I}}_i|\hat{\mathbf{I}}_{<i})$. The objective is to maximize the log-likelihood of the data representation, which can be expressed as follows:

$$\mathcal{L}_T = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\big[-\log p(\hat{\mathbf{I}})\big], \tag{2}$$

where $p(\hat{\mathbf{I}}) = \prod_i p(\hat{\mathbf{I}}_i|\hat{\mathbf{I}}_{<i})$. To simulate the potential loss of indices during transmission, we incorporate a **masking** procedure. Specifically, we apply a binary mask $M = [m_i]_{i=1}^N$ as follows: For $m_i = 1$, the corresponding index $\hat{\mathbf{I}}_i$ is replaced by a special $[mask]$ token to indicate that it has been lost. Conversely, if $m_i = 0$, then $\hat{\mathbf{I}}_i$ is left unchanged. The mask process is controlled by a mask ratio ($\alpha \in [0, 1]$), which determines the number of missing indices as $\alpha \cdot N$, denoted as $\hat{\mathbf{I}}_{lost}$. During the restoration stage, as depicted in Fig. 1(c), a sliding window input strategy is employed

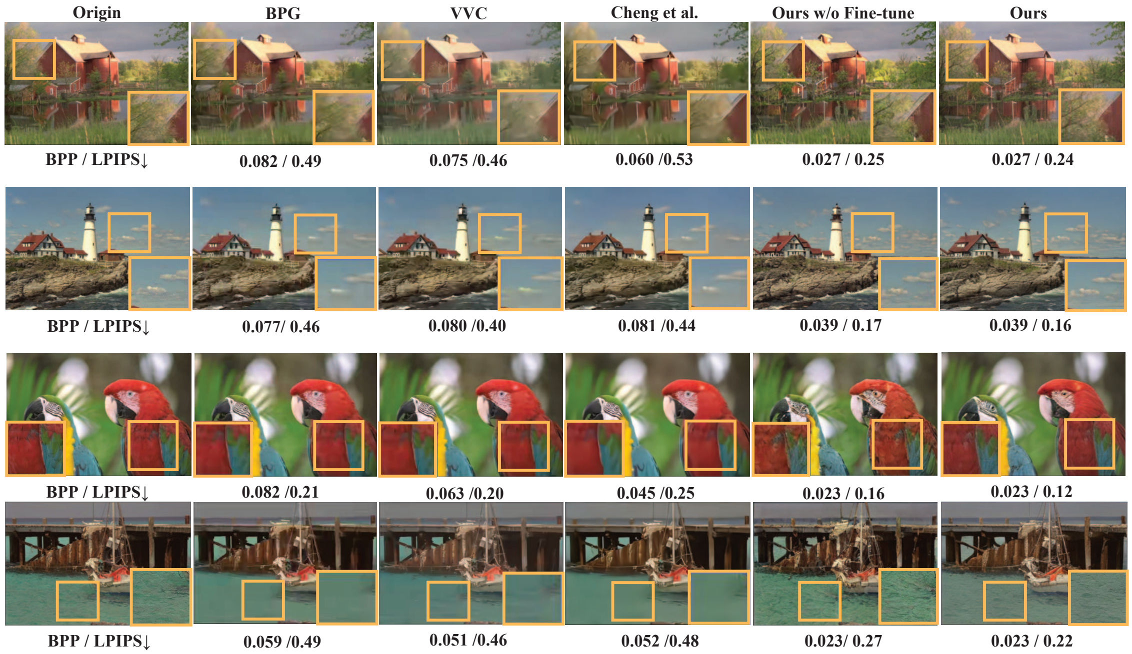

Figure 2: The qualitative comparison results of BPG, VVC, Cheng *et al.*, ours w/o fine-tune and our method on the Kodak dataset.

to input a 16×16 window centered on $\hat{\mathbf{I}}_i$ at each prediction. Only the indices before i in the 16×16 window are included as input due to the autoregressive nature of the transformer, which is consistent with the decoding process. Finally, the restored image $\tilde{\mathbf{x}}$ is generated by feeding the predicted index sequence $\hat{\mathbf{I}}_{predict}$ into the decoder G.

3 Experimental Results

3.1 Experimental Settings

DataSets. The proposed framework is trained on the ImageNet dataset [21], which contains 1.2 million images distributed across $1,000$ distinct categories. To assess the performance of the proposed model, we evaluate two commonly used datasets in image compression: the Kodak dataset [22], which includes 24 natural uncompressed 512×768 or 768×512 images; and the CLIC2020 dataset [23], comprising 250 images that exhibit varying lighting conditions and dynamic range, with resolutions ranging from 320×240 to 4032×3024.

Implementation Details. We adopt $M = 16$ to downsample the image $\mathbf{x}$ into the latent representation $\mathbf{z}$. First, the weights of encoder E, decoder G, and the discriminator D are initialized by the officially provided pre-trained VQ-GAN model [1]. To enable the proper bitrate range, we perform K-means clustering on the size of 16384 codebook of the pre-trained model, reducing the new codebooks into the size, ranging from $\{2048, 1024, 512, 256, 128, 64, 32, 16, 8\}$. We then fine-tune the entire framework

[1] https://github.com/CompVis/taming-transformers

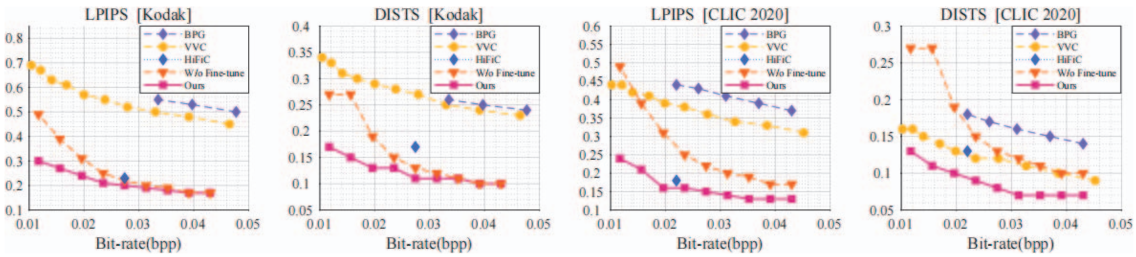

Figure 3: The R-D performance of BPG, VVC, Cheng *et al.*, FCC, HiFiC, ours w/o fine-tune and the proposed method on the Kodak dataset, and the CLIC2020 dataset.

Table 1: BD-rate and BD-metric relative to the BPG, VVC, Cheng *et al.*, and ours w/o fine-tuning respectively, where LPIPS is used as the distortion metric in BD-metric.

Baseline	Kodak		CLIC	
	BD-rate	BD-LPIPS	BD-rate	BD-LPIPS
VVC	-98.55%	-0.33	-93.02%	-0.21
BPG	-99.99%	-0.36	-99.74%	-0.27
HiFiC	-11.62%	-0.02	-8.23%	-0.01
Cheng et. al	-97.93%	-0.37	-98.90%	-0.25
Ours W/o Fine-tune	-21.57%	-0.06	-25.59%	-0.06

using the default settings and training losses as [17], which takes 13 hours on two NVIDIA Tesla-A100 GPUs.

Evaluation Metrics. Traditional objective quality assessment methods, such as PSNR and SSIM, have been primarily devised for the calculation of pixel-level distortions. However, these methods may not be suitable for assessing perceptual quality. Therefore, we integrate recent approaches *i.e.* the learned perceptual image patch similarity (LPIPS) [24] and the deep image structure and texture similarity (DISTS) [25] to assess perceptual quality, which better aligns with the way humans perceive images. All of the above metrics are the smaller, the better. Furthermore, we utilize bpp to evaluate the rate performance.

3.2 Compression Performance Comparison

Compared Methods. To evaluate the efficacy of our proposed framework, we compare the proposed method with both traditional standard and neural-based typical compression frameworks: First, we compare with classic image codec BPG [1] and the latest video coding codec VVC [2]. For the typical end-to-end (E2E) codec based on deep learning, we compare with Cheng *et al.* [19] and retrain the model implemented by CompressAI[2] to cover a bitrate range similar to ours. Furthermore, we develop

[2]https://github.com/InterDigitalInc/CompressAI

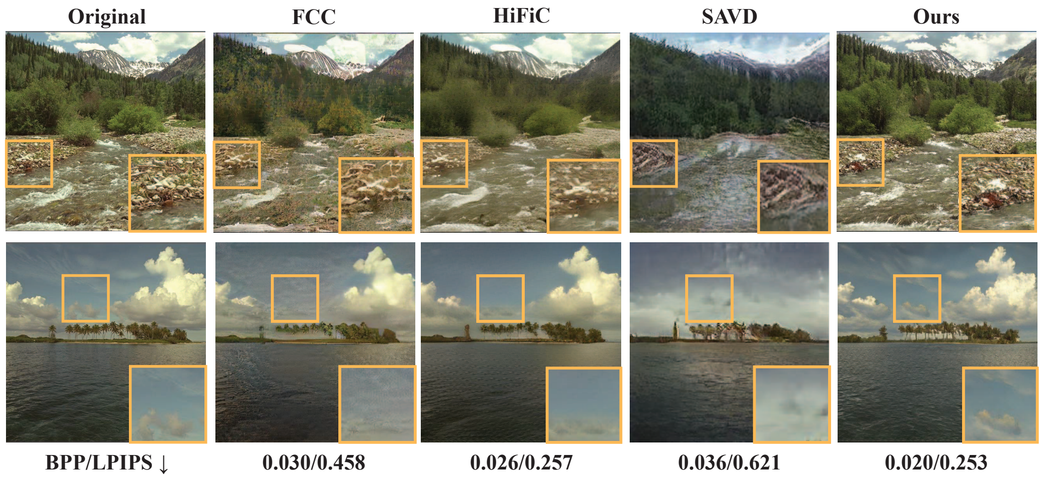

Figure 4: Qualitative comparisons of three typical generative compression codecs (*i.e.* FCC, HiFiC, SAVD) and the proposed method over the Kodak dataset.

an additional baseline that directly adopts the K-means clustering algorithm without any fine-tuning, denoted as "Ours w/o Fine-tune".

Qualitative Evaluation. Fig. 2 shows the reconstruction results of various methods on different images on the Kodak dataset [22], as well as the corresponding bpp and LPIPS. It is evident that at extremely low bitrates, VVC [2], BPG [1], and Cheng *et al.* [19] exhibit varying degrees of blurring and missing texture details. In contrast, our proposed method reconstructs more texture details, such as trees, ripples, clouds, and feathers, resulting in the lowest LPIPS value. Moreover, models obtained using the K-means clustering algorithm can produce relatively rich details even without fine-tuning training, thanks to the inherent expressive power of the pre-trained codebook. Nevertheless, upon fine-tuning, our proposed method outperforms the former by achieving more color and texture reconstruction consistency.

Quantitative Evaluation. Our proposed framework is capable of achieving extremely low bitrate compression from **0.01** to **0.04** bpp while maintaining perceptual quality for a wide range of images with diverse semantics and resolutions across various datasets. As illustrated in Fig. 3, the experimental results demonstrate the proposed approach exhibits remarkable improvements in perceptual-oriented metrics compared to other compression methods. To better evaluate the rate-distortion (R-D) performance improvement of our proposed method, we utilize the Bjontegaard metric [26], as demonstrated in Table 1. For instance, with the same reconstruction quality of the LPIPS metric, our approach achieves approximately 97.93% $\sim$ 99.99%, 93.02% $\sim$ 99.74% bitrates saving compared to VVC [2], BPG [1] and Cheng *et al.* [19] over the two datasets, respectively. These results demonstrate the remarkable advantage of our proposed method achieving high perceptual quality for extremely low bitrate coding. The performance of models generated by utilizing the K-means clustering algorithm without fine-tuning can yield promising results on the R-D curve when the size of the codebook is relatively large. Nonetheless, a significant degradation in performance is observed when the codebook size is reduced to below 64.

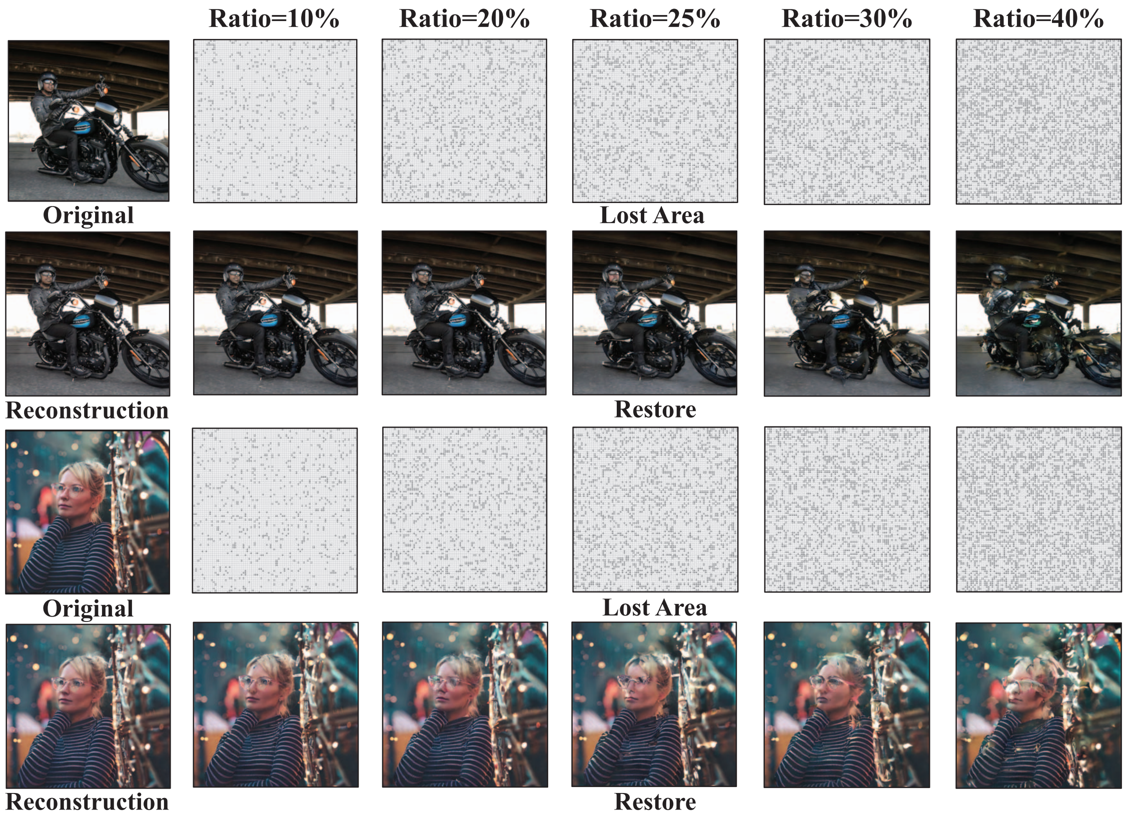

Figure 5: Image restoration via lost VQ-indices prediction ranging from 10% to 40%.

Fine-tuning can effectively enhance the performance of the model under such conditions, with 21.57% $\sim$ 25.59% bitrates saving, as shown in Table 1.

3.3 Comparisons with Generative Image Compression Codecs

To comprehensively assess the efficiency and applicability of the proposed approach vis-a-vis state-of-the-art generative compression codecs, we perform comparative analyses against three representative codecs, as illustrated in Fig. 3 and Fig. 4. FCC [8] and HiFiC [7] are among the first categorical approaches that utilize adversarial loss to optimize the end-to-end framework. Nevertheless, all of them exhibit suboptimal performance in the presence of artifacts resembling colored and circle dots under extreme bitrates ($\leq$ 0.03 bpp). Furthermore, the proposed method achieves bitrate savings ranging from 8.23% to 11.62% when compared to HiFiC [7], as indicated in Table 1. As a second line of approach, the SAVD method [16] encodes images into semantic maps and corresponding texture features. We compare our method with the model trained on the ADE20K outdoor dataset. It can be observed that it suffers from poor generalization ability in the presence of a semantic gap between the training and testing domains, resulting in inadequate texture generation. In contrast, our method exhibits superior performance and generalization in generating reconstructed images with both subjective perception and objective metrics, with much lower bitrates.

3.4 Lost VQ-indices Prediction

Fig. 5 presents a visualization of the lost indices map ranging from 10% to 40% loss ratios, as well as the corresponding restored images, generated via the utilization of indices predicted by the transformer model. It can be observed that the restored images display a similar degree of fidelity to the reconstructed image without missing indices when the levels of missing data are not particularly significant ($\leq 20\%$). Large ratio loss can also preserve the rough structure of the original image, which verifies the robustness and resilience of the proposed coding framework.

4 Conclusions

In this work, we propose a novel scheme that utilizes the VQ-indices maps obtained from VQGANs as compact visual data representations to achieve extremely high image compression ratios while maintaining perceptual quality. Our proposed scheme uniquely adjusts the quantization step by varying the size of codebooks through the K-means clustering and recovers missing indices by the transformer, enabling reliable compression with variable bit rates and varying levels of reconstruction quality. Qualitative and quantitative results demonstrate the superiority of our proposed scheme in perceptual quality at extremely low bit rates (≤ 0.04 bpp) compared to state-of-the-art codecs. Even when up to 20% of indices are lost, the images can be successfully restored with minimal perceptual loss. Overall, our work advances image/video coding research by demonstrating the potential of VQ-based generative models for research in ultra-low bitrate compression.

5 Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 62201526, 62025101, and 62022002; the Fundamental Research Funds for the Central Universities (CUC23GZ007); and the Public Computing Cloud at CUC, all of which are gratefully acknowledged.

References

[1] Fabrice Bellard, "Bpg image format," `https://bellard.org/bpg/`, 2018.

[2] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[3] Oren Rippel and Lubomir Bourdev, "Real-time adaptive image compression," in *ICML*. PMLR, 2017, pp. 2922–2930.

[4] Shibani Santurkar, David Budden, and Nir Shavit, "Generative compression," in *2018 Picture Coding Symposium (PCS)*. IEEE, 2018, pp. 258–262.

[5] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool, "Generative adversarial networks for extreme learned image compression," in *ICCV*, 2019, pp. 221–231.

[6] Jooyoung Lee, Donghyun Kim, Younhee Kim, Hyoungjin Kwon, Jongho Kim, and Taejin Lee, "A training method for image compression networks to improve perceptual quality of reconstructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 144–145.

[7] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson, “High-fidelity generative image compression,” *NIPS*, vol. 33, pp. 11913–11924, 2020.
[8] Shoma Iwai, Tomo Miyazaki, Yoshihiro Sugaya, and Shinichiro Omachi, “Fidelity-controllable extreme image compression with generative adversarial networks,” in *ICPR*. IEEE, 2021, pp. 8235–8242.
[9] Jianhui Chang, Qi Mao, Zhenghui Zhao, Shanshe Wang, Shiqi Wang, Hong Zhu, and Siwei Ma, “Layered conceptual image compression via deep semantic synthesis,” in *ICIP*. IEEE, 2019, pp. 694–698.
[10] Jianhui Chang, Zhenghui Zhao, Lingbo Yang, Chuanmin Jia, Jian Zhang, and Siwei Ma, “Thousand to one: Semantic prior modeling for conceptual coding,” in *IEEE International Conference on Multimedia and Expo*. IEEE, 2021, pp. 1–6.
[11] Jianhui Chang, Zhenghui Zhao, Chuanmin Jia, Shiqi Wang, Lingbo Yang, Qi Mao, Jian Zhang, and Siwei Ma, “Conceptual compression via deep structure and texture synthesis,” *TIP*, vol. 31, pp. 2809–2823, 2022.
[12] Jianhui Chang, Jian Zhang, Youmin Xu, Jiguo Li, Siwei Ma, and Wen Gao, “Consistency-contrast learning for conceptual coding,” in *ACM MM*, 2022, pp. 2681–2690.
[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[14] Diederik P Kingma and Max Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
[15] Mehdi Mirza and Simon Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014.
[16] Jianhui Chang, Jian Zhang, Jiguo Li, Shiqi Wang, Qi Mao, Chuan Jia, Siwei Ma, and Wen Gao, “Semantic-aware visual decomposition for image coding,” *IJCV*, vol. 131, pp. 2333–2355, 2023.
[17] Patrick Esser, Robin Rombach, and Bjorn Ommer, “Taming transformers for high-resolution image synthesis,” in *CVPR*, 2021, pp. 12873–12883.
[18] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman, “Maskgit: Masked generative image transformer,” in *CVPR*, 2022, pp. 11315–11325.
[19] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto, “Learned image compression with discretized gaussian mixture likelihoods and attention modules,” in *CVPR*, 2020, pp. 7939–7948.
[20] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár, “Lossy image compression with compressive autoencoders,” *arXiv preprint arXiv:1703.00395*, 2017.
[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*. Ieee, 2009, pp. 248–255.
[22] Eastman Kodak, “Kodak photocd dataset,” `http://r0k.us/graphics/kodak/`, 2013.
[23] Toderici George, Theis Lucas, Johnston Nick, Agustsson Eirikur, Mentzer Fabian, Balle Johannes, Shi Wenzhe, and Timofte. Radu, “Clic 2020: Challenge on learned image compression, 2020,” `https://www.tensorflow.org/datasets/catalog/clic`, 2020.
[24] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *CVPR*, 2018, pp. 586–595.
[25] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli, “Image quality assessment: Unifying structure and texture similarity,” *ICME*, vol. 44, no. 5, pp. 2567–2581, 2020.
[26] Gisle Bjontegaard, “Calculation of average psnr differences between rd-curves,” *ITU SG16 Doc. VCEG-M33*, 2001.