

# Generative Latent Coding for Ultra-Low Bitrate Image Compression

Zhaoyang Jia<sup>1\*</sup> Jiahao Li<sup>2</sup> Bin Li<sup>2</sup> Houqiang Li<sup>1</sup> Yan Lu<sup>2</sup>  
<sup>1</sup> University of Science and Technology of China <sup>2</sup> Microsoft Research Asia

jzy\_ustc@mail.ustc.edu.cn, lihq@ustc.edu.cn, {li.jiahao, libin, yanlu}@microsoft.com

## Abstract

Most existing image compression approaches perform transform coding in the pixel space to reduce its spatial redundancy. However, they encounter difficulties in achieving both high-realism and high-fidelity at low bitrate, as the pixel-space distortion may not align with human perception. To address this issue, we introduce a **Generative Latent Coding (GLC)** architecture, which performs transform coding in the latent space of a generative vector-quantized variational auto-encoder (VQ-VAE), instead of in the pixel space. The generative latent space is characterized by greater sparsity, richer semantic and better alignment with human perception, rendering it advantageous for achieving high-realism and high-fidelity compression. Additionally, we introduce a categorical hyper module to reduce the bit cost of hyper-information, and a code-prediction-based supervision to enhance the semantic consistency. Experiments demonstrate that our GLC maintains high visual quality with less than 0.04 bpp on natural images and less than 0.01 bpp on facial images. On the CLIC2020 test set, we achieve the same FID as MS-ILLM with 45% fewer bits. Furthermore, the powerful generative latent space enables various applications built on our GLC pipeline, such as image restoration and style transfer.

## 1. Introduction

Amid the ongoing surge of digital visual data, the importance of achieving high-efficiency image compression becomes increasingly paramount. From the traditional compression standards [54, 55] to the emerging learned image compression models [4, 5, 11, 20, 41, 45], most compression algorithms follow the pixel-space transform coding [4, 19] paradigm. Specifically, they convert pixels into compact representations through a transform module, which eliminates the redundancy to reduce the bit cost in the subsequent entropy coding process.

However, we observe a common inherent limitation in

\*This work was done when Zhaoyang Jia was an intern at Microsoft Research Asia.

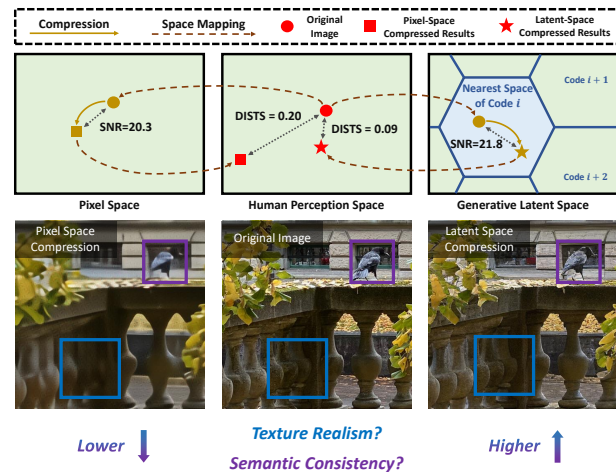


Figure 1. Generative latent space of VQ-VAE exhibits better alignment with human perception than pixel space for ultra-low bitrate compression. Under comparable distortion levels (measured by signal-to-noise ratio, SNR), latent-space compression produces reconstructions with superior perceptual quality (measured by DISTs [15]) than pixel-space generative codec MS-ILLM [46], as the compressed latents remain in the same latent code space.

these methods: the pixel-space distortion is not always consistent with the human perception, especially at low bitrate. In practice, human observers prioritize the semantic consistency and the texture realism of an image, but this information is difficult to be adequately exploited solely by a pixel-space transform module. As shown in the left of Figure 1, pixel-space generative image codec MS-ILLM [46] struggles to guarantee visual quality at low bitrate, even after it incorporates perceptual supervision [29] and adversarial supervision [18] within the pixel space.

Based on such observation, a natural problem arises: *how can we compress images in a way that aligns with human perception?* To address this challenge, we introduce a **Generative Latent Coding (GLC)** paradigm. In GLC, we first encode images into a generative latent space that aligns with human perception, and subsequently perform transform coding in this latent space, instead of in pixel space. To concretize this concept, we adopt a generative vector-quantized variational auto-encoder (VQ-VAE) [17, 53] to

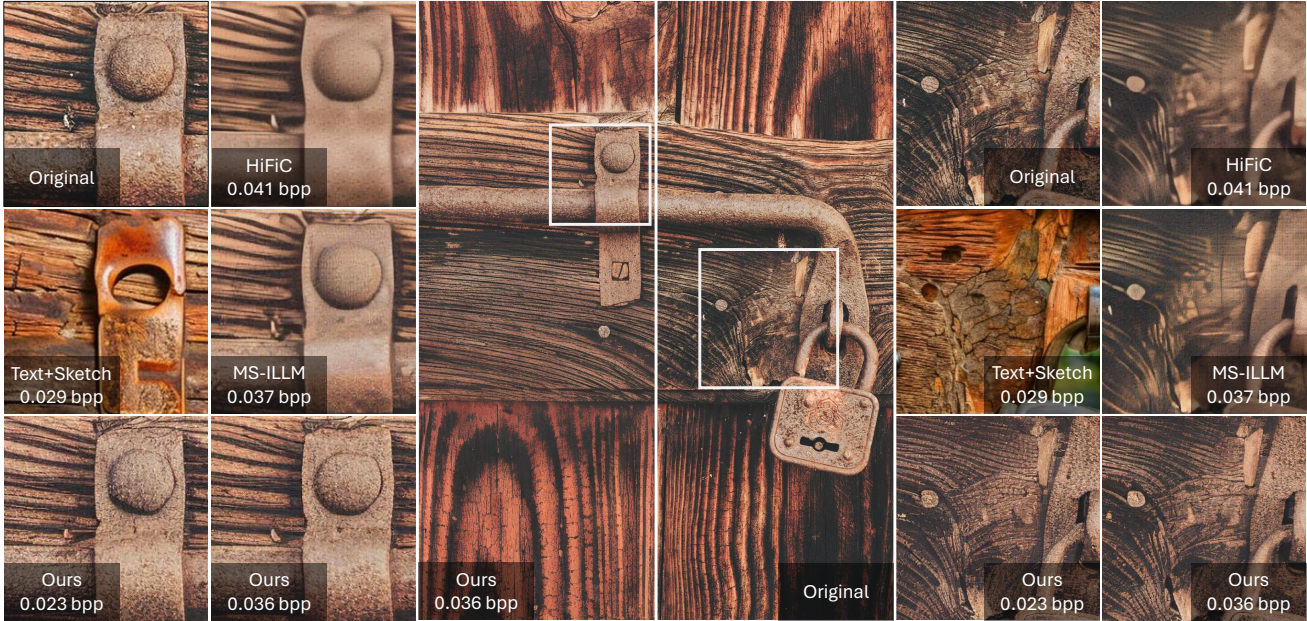


Figure 2. A qualitative comparison between HiFiC [44], MS-ILLM [46], Text+Sketch [35] and the proposed GLC. GLC produces images with high visual quality, even in regions with complex texture. In contrast, HiFiC and MS-ILLM exhibit noticeable artifacts, and Text+Sketch generates results that deviate significantly from the input. *Best viewed when zoomed in.*

produce the latent space, which offers three significant advantages: 1) The discrete codes of VQ-VAE encapsulate semantic visual components [53], allowing GLC to focus on compressing the semantic content to guarantee fidelity. 2) Generative VQ-VAE exhibits remarkable generative capabilities [17] for high-realism texture reconstruction. 3) The discrete variational bottleneck naturally brings a low-entropy and distortion-robust latent space for compression. Thanks to such characteristics, GLC is more aligned with human perception to achieve enhanced visual quality, as demonstrated in the right of Figure 1.

When implementing GLC, two crucial questions persist: *How to effectively compress the generative latents?* And *how to supervise the generative latent coding?* A straightforward approach to compress VQ-VAE latents is indices-map coding [27, 28, 43], but it is limited by the ineffective redundancy reduction between indices and the lack of rate-variable coding support. In this paper, we propose a novel generative-latent-space transform coding approach, where an effective rate-variable structure is adopted to reduce latent redundancy for higher compression ratio. In addition, a categorical hyper module is introduced to model the distribution of  $z$  with a discrete codebook, which significantly reduces the bitrate of  $z$  when compared to the factorized hyper module [5]. As for the supervision of GLC, inspired by recent code prediction transformers [27, 28, 60], a code-prediction-based supervision is proposed. It serves as an auxiliary supervision employed solely in the training process to greatly enhance the semantic consistency.

Benefited from these advanced designs, our GLC achieves excellent performance on both natural and facial images. In the CLIC 2020 test set [51], GLC attains a bitrate less than 0.04 bpp while delivering high visual quality. It obtains 45% bit savings compared to MS-ILLM at the same FID. In the CelebAHQ [30] dataset, GLC achieves an even lower bitrate of less than 0.01 bpp. As shown in Figure 2, compared with recent advanced generative image compression approaches MS-ILLM [46] and Text+Sketch [35], GLC produces more visually appealing compression results with a lower bit cost.

Furthermore, leveraging the representative generative latent space, GLC supports various vision applications such as image restoration and style transfer. By replacing the compression encoder with a restoration encoder, the proposed restoration application surpasses the performance of cascading a restoration model and a neural codec. We hope such versatility of generative latent space will help connect image compression with other vision tasks in the future.

In summary, our main contributions are :

- We present a generative latent coding (GLC) scheme, which performs transform coding in the generative latent space of a VQ-VAE to achieve high-fidelity and high-realism compression at ultra-low bitrate.
- We introduce a categorical hyper module to significantly reduce the bit cost of hyper information. Additionally, a code-prediction-based supervision is adopted to enhance the perceptual quality.
- GLC obtains 45% bit reduction on CLIC2020 with the

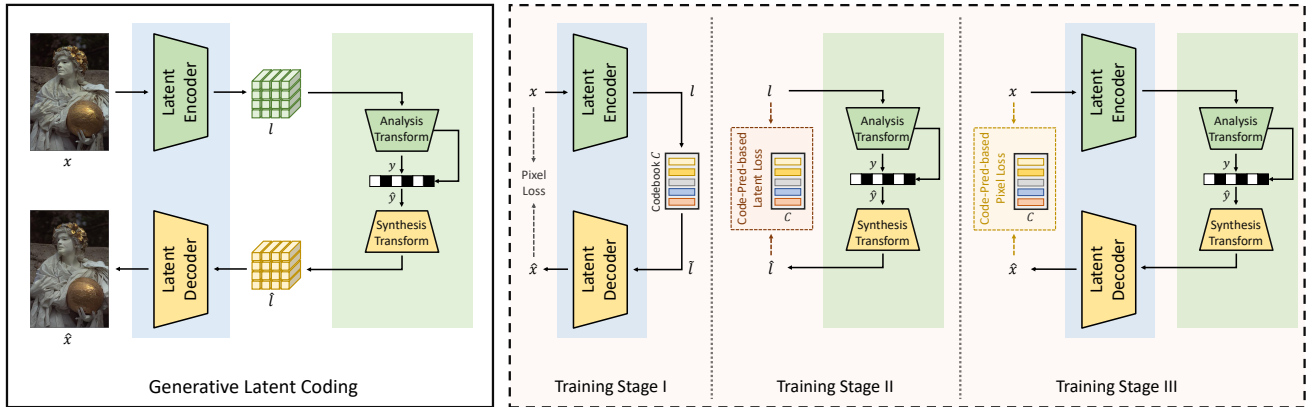


Figure 3. Illustration of the proposed **Generative Latent Coding (GLC)** framework. (Left) GLC firstly encodes the image into a generative latent representation (Section 3.2), then compresses the latent with transform coding (Section 3.3), and finally reconstructs image from the compressed latent. (Right) We progressively train GLC in three stages (Section 4): In stage I, we train a generative VQ-VAE to obtain a human-perception-aligned latent space. In stage II, the transform coding module learns to compress the latent with a code-prediction-based latent supervision (Figure 6). Finally, in stage III, the entire network is fine-tuned jointly with a code-prediction-based pixel supervision to further enhance the compression performance.

same FID as the previous SOTA. Furthermore, GLC enables various application within its latent space.

## 2. Related Works

### 2.1. Learned Image Compression

Lossy image compression is grounded on Shannon’s rate-distortion theory [12]. Ballé et al. [4] first proposed to utilize neural networks for pixel-space transform coding [19], employing analysis and synthesis transform modules to convert images into compact representations for entropy coding. Subsequently, some researches make strides in improving the probability model [4, 5, 11, 34, 38, 45] for more accurate estimation, while others explore the network structure [11, 41], optimization algorithm [58, 59] and rate-variable coding [13, 20] for improved compression performance and practicality.

A recently raised critical challenge in image compression is how to improve the perceptual quality of the reconstruction. Agustsson et al. [2] first introduced the concept of *generative compression*, which compresses essential image features and generate distorted details using GAN. Some subsequent works [9, 25] extract image sketches and latent codes to ensure geometry-consistency. Text+Sketch [35] utilizes a conditional-diffusion model to generate image based on image captions and sketches, achieving superior perceptual quality. While these schemes produce visual appealing results, they often deviate significantly from the input and cannot guarantee the semantic consistency.

To achieve generative compression with high-fidelity, Mentzer et al. [44] further studied the network structure and generative adversarial loss to enable high-fidelity compression. Subsequent researches further enhance the trans-

form coding [16, 22], generative post-processing [24] or focus on controlling the trade-off between fidelity and realism [3, 26]. Recently, MS-ILLM [46] introduces a no-binary discriminator which is conditioned on quantized local image representations to greatly enhance the statistical fidelity.

### 2.2. Latent Space Modeling

Latent space image modeling means modeling the distribution of image within the latent space of a neural network. This technique has been primarily developed for image generation. Chen et al. [10] and Oord et al. [53] introduced PixelCNN [52] in the latent space of VAE and VQ-VAE for image generation. Esser et al. [17] took it a step further by incorporating transformers into VQ-VAE latent space for high-quality generation. More recently, Rombach et al. [48] employed diffusion models to model the latent space of VQ-VAE, achieving remarkable results in high-resolution image generation. These studies underscore the potential of image processing within the generative latent space, particularly in the latent space of VQ-VAE.

Recently, the concept of latent space modeling has been extended to other tasks. CodeFormer [60] introduces a code prediction transformer that takes distorted latents as input and predicts the high-quality VQ-VAE index for facial restoration. Building upon this, Jiang et al. [27, 28] proposed transmitting the predicted indices to achieve restoration-based facial conferencing. In this paper, we explore the characteristics of latent space modeling in the realm of generative image compression. We specially design a transform coding paradigm in the latent space, which demonstrates superior effectiveness.

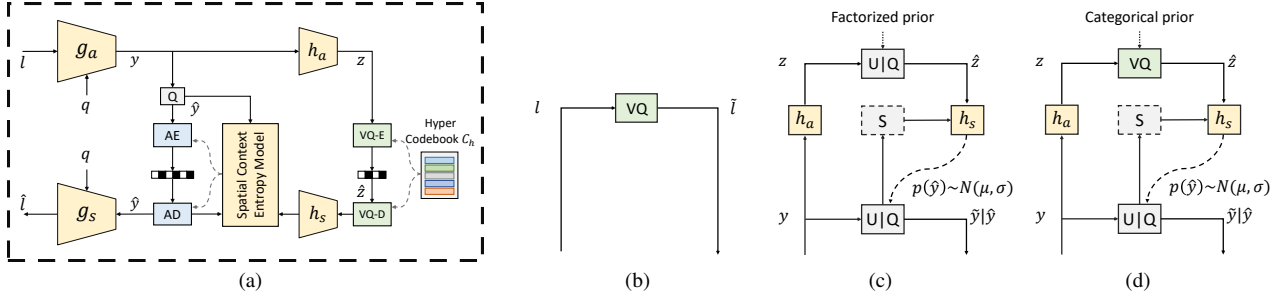


Figure 4. Illustration of the transform coding in latent space. (a) The model structure of the transform coding module. We further compare it with other coding schemes in operational diagrams : (b) indices-map coding [27, 28, 43], (c) transform coding with factorized hyper module [5, 21, 38] and (d) proposed transform coding with categorical hyper module. Here AE and AD denote arithmetic encoding and decoding, VQ-E and VQ-D stand for VQ-indices-map encoding and decoding, Q refers to scalar quantization, U signifies the addition of uniform noise as a differential simulation of Q, and S denotes the spatial context entropy module.

### 3. Method

#### 3.1. Overview

In this section, we introduce the details of the proposed **Generative Latent Coding (GLC)** architecture. To achieve high-perceptual-quality compression, GLC encodes image into a perception-aligned latent space through a generative latent auto-encoder, and perform transform coding on the latent representations for lower bitrate. As depicted in the left of Figure 3, the input image  $x$  is firstly encoded into the latent  $l$  using the latent encoder  $E$ . Then  $l$  undergoes an analysis transform  $g_a$  to produce the code  $y$ , which is further scalar-quantized to  $\hat{y}$  for entropy coding. Following that, a synthesis transform  $g_s$  is employed to transform  $\hat{y}$  back to  $\hat{l}$ , and finally, the reconstruction  $\hat{x}$  is generated by the latent decoder  $D$ . This entire process is formulated as:

$$\begin{aligned}
 l &= E(x), \quad y = g_a(l) \\
 \hat{y} &= Q(y) \\
 \hat{l} &= g_s(\hat{y}), \quad \hat{x} = D(\hat{l})
 \end{aligned} \tag{1}$$

#### 3.2. Generative Latent Auto-Encoder

To achieve high-quality generative latent coding, *how to obtain a human-perception-aligned latent space* is a crucial challenge. In GLC, we address it by employing the generative VQ-VAE [17] as the latent auto-encoder ( $E$  and  $D$ ). By mapping images into visual semantic elements within a codebook  $C$  and incorporating a generative image decoding process, both semantic consistency and texture realism can be well guaranteed. Additionally, it contributes to the compression process through a sparse yet robust latent space, which is achieved by training with the discrete codebook  $C$  as a variational bottleneck.

#### 3.3. Transform Coding in Latent Space

To compress the latent representations  $l$ , a direct approach is VQ-indices-map coding [27, 28, 43] (Figure 4b). How-

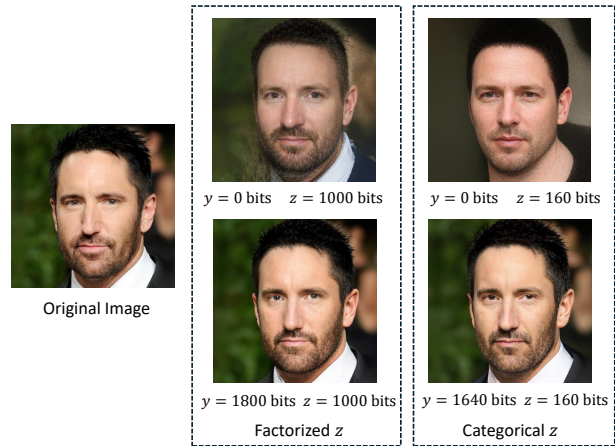


Figure 5. Example of comparison between factorized and categorical hyper modules. The proposed categorical  $z$  encodes essential semantic and structural information with much less bits.

ever, these methods often lack a careful consideration of the correlation among the latents, resulting in an insufficient redundancy reduction and consequently a high bit cost. In GLC, we introduce a transform coding module to compress the latent, replacing the vector-quantization step for more effective reduction of latent redundancy. As shown in Figure 4a, the latents are transformed into code  $y$  using transformations  $g_a, g_s$  and then quantized to  $\hat{y}$ . Entropy coding is applied to  $\hat{y}$  based on a probability  $p(\hat{y})$ , which is estimated by a categorical hyper module (Section 3.3.1) and a quadtree-partition-based spatial context module [38].

##### 3.3.1 Categorical Hyper Module

Factorized hyper module [5] (Figure 4c) is commonly employed in recent image compression schemes. However, at ultra-low bitrate, we notice that the factorized  $z$  tends to encode low-level information such as color and texture, incur-

ring a high bit cost, as illustrated in Figure 5. To address it, we propose a categorical hyper module (Figure 4d), which utilizes a hyper codebook to store the basic semantic elements rather than the low-level information. This module comprises a hyper analysis transform  $h_a$ , a hyper synthesis transform  $h_s$  and a hyper codebook  $C_h$ . The transformations are formulated as:

$$z = h_a(y), \hat{z} = VQ(z, C_h), \text{prior}_z = h_s(\hat{z}) \quad (2)$$

where  $z$  and  $\hat{z}$  denote hyper-codes.  $VQ(\cdot, C_h)$  represents vector-quantization by nearest lookup in  $C_h$ . As shown in the right of Figure 5, the categorical  $z$  is more inclined to capture high-level semantic information, which can be encoded with significantly fewer bits.

### 3.3.2 Rate-Variable Transformation

A notable advantage of transform coding over VQ-indices-map coding is its capability for rate-variable compression, which is a core functionality for a practical image codec. Indices-map coding is limited since the codebook can only model one specified distribution, but different rates naturally need different distributions. In contrast, transform coding converts latent into a unified Gaussian distribution, and variable-rate can be achieved by variable parameters (e.g., means and scales) of Gaussian. In GLC, we follow DCVC series [36–39, 47, 49] to incorporate a scaler  $q$  in the transform coding to achieve rate-variable compression.

## 4. Progressive Training

As depicted in the right of Figure 3, we adopt a three-stage progressive training manner to fully leverage the potential of GLC. We initially learn a human-perception-aligned latent space to guarantee the perceptual quality, subsequently learn to perform transform coding on this latent space to achieve low bitrate, and finally fine-tune the entire network for superior compression performance. At each stage, distinct loss functions are adopted to guide different modules.

### 4.1. Stage I : Auto-Encoder Learning

To obtain a human-perception-aligned latent space for compression, we begin with training a generative VQ-VAE as the initialization of  $E$  and  $D$ . To ensure the sparsity of the latent space, an auxiliary codebook  $C$  is employed to perform nearest vector-quantization, transforming  $l$  to  $\tilde{l}$ . The supervision comprises reconstruction loss, perceptual loss [29], adversarial loss [18] and codebook loss [53]:

$$\begin{aligned} \mathcal{L}_{\text{Stage I}} = & \|x - \hat{x}\| + \mathcal{L}_{\text{per}}(x, \hat{x}) \\ & + \lambda_{\text{adv}} \cdot \mathcal{L}_{\text{adv}}(x, \hat{x}) + \mathcal{L}_{\text{codebook}} \end{aligned} \quad (3)$$

Here,  $\mathcal{L}_{\text{per}}$  corresponds to the LPIPS loss calculated using VGG [50] extracted features.  $\mathcal{L}_{\text{adv}}$  is the adaptive PatchGAN adversarial loss [17] with a weight of  $\lambda_{\text{adv}} = 0.8$ .

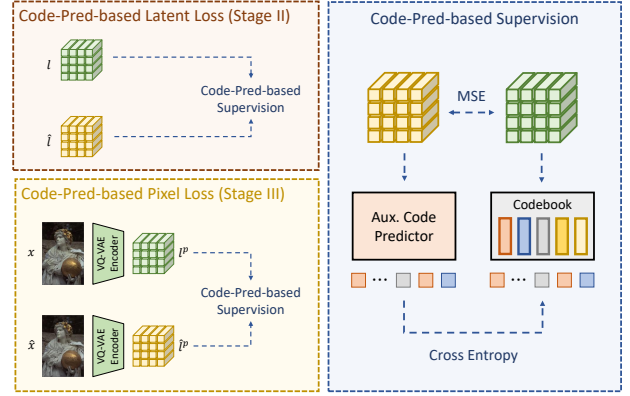


Figure 6. Illustration of code-prediction-based supervision.

The codebook loss is formulated as

$$\mathcal{L}_{\text{codebook}} = \|\text{sg}(l) - \tilde{l}\| + \beta \cdot \|\text{sg}(\tilde{l}) - l\| \quad (4)$$

where  $\text{sg}(\cdot)$  denotes the stop-gradient operator and  $\beta = 0.25$  controls the update rates of the  $E$  and  $C$ .

### 4.2. Stage II : Transform Coding Learning

Given the trained latent space, we further learn to perform transform coding to achieve low-bitrate latent compression, while fixing the auto-encoder  $E$  and  $D$ . We introduce an auxiliary code predictor  $CP$  to enhance the semantic consistency by necessitating the latent to possess the capability to predict the correct VQ-indices. As shown in Figure 6, we encode  $l$  into VQ-indices by  $M_l = VQ(l, C)$  and subsequently predict these indices by  $\hat{M}_l = CP(\hat{l})$ . So the code-prediction-based loss can be formulated by

$$\mathcal{D}_{\text{code}}(l, \hat{l}) = \alpha \cdot CE(M_l, \hat{M}_l) + \|l - \hat{l}\|_2^2 \quad (5)$$

where  $CE$  denotes the cross entropy loss and we set  $\alpha = 0.5$  by default. Then the transform coding module can be supervised by the rate-distortion trade-off

$$\mathcal{L}_{\text{Stage II}} = \mathbf{E}_{x \sim p_X} [\mathcal{R}(\hat{y}) + \lambda \cdot \mathcal{D}_{\text{code}}(l, \hat{l})] \quad (6)$$

where  $\mathcal{R}$  is the estimated rate and  $\lambda$  is used the control the trade-off. Note that a codebook loss (as formulated in Equation 4) is required to train the hyper codebook  $C_h$  in the categorical hyper module. We omit it from the loss functions of both stage II and stage III for the sake of conciseness.

### 4.3. Stage III : Joint Training

Finally, we fine-tune the entire network with the pixel space supervision to achieve better compression performance. As shown in Figure 6, we extend the code-prediction-based latent supervision into the pixel space. Specifically, we utilize the encoder  $E_{VQ}$  trained from stage I to encode  $x$  and

$\hat{x}$  into latent space by  $\hat{l}^p = E_{VQ}(\hat{x})$  and  $l^p = E_{VQ}(x)$ , so the code-prediction-based pixel loss can be calculated by  $\mathcal{D}_{code}(l^p, \hat{l}^p)$  in the same formulation with Equation 5. Here we use  $E_{VQ}$  since it can map the input data to a compatible latent space with the codebook  $C$  for code prediction. The overall pixel supervision is defined as :

$$\begin{aligned} \mathcal{D}_{\text{Stage III}} = & \|x - \hat{x}\| + \mathcal{L}_{per}(x, \hat{x}) \\ & + \lambda_{adv} \cdot \mathcal{L}_{adv}(x, \hat{x}) + \lambda_{code} \cdot \mathcal{D}_{code}(l^p, \hat{l}^p) \end{aligned} \quad (7)$$

where we set  $\lambda_{code} = 0.05$  by default. The rate-distortion trade-off supervision is :

$$\mathcal{L}_{\text{Stage III}} = \mathbf{E}_{x \sim p_X} [\mathcal{R}(\hat{y}) + \lambda \cdot \mathcal{D}_{\text{Stage III}}] \quad (8)$$

#### 4.4. Discussion of Code-Prediction-Based Loss

Code prediction transformers [27, 28, 60] have demonstrated their effectiveness in high-quality image reconstruction. They typically input the predicted latent directly into the decoder for reconstruction. Different from these methods, in GLC, we suggest to consider code prediction solely as an auxiliary supervision during training, but not used in the inference process of the compression pipeline.

This design is based on an observation: if a code prediction module is introduced before the decoder, the fine-tuning process in stage III cannot enhance compression performance further. It appears that the codebook becomes a performance bottleneck, restricting the decoder to receiving only the vector-quantized latent as input, which has already been well-trained in stage I. In GLC, by utilizing code prediction solely as auxiliary supervision during training, we eliminate this bottleneck, allowing the decoder to receive more flexible input for additional fine-tuning. We find that this code-prediction-based supervision effectively enhances the semantic consistency of the reconstructions, as shown in Figure 7. By necessitating the latent to possess the capability to predict the code index, the latent contains more semantic information, such as gestures and attributes.

## 5. Experiments

### 5.1. Implementation Details

**Training details.** We train GLC for both natural image compression and facial image compression. For natural images, we conduct stage I on ImageNet training set [14], stage II and III on OpenImage test set [33], using randomly cropped  $256 \times 256$  patches. For facial images, GLC is trained on FFHQ dataset [31] for all stages with a resolution of  $512 \times 512$ . Both models are optimized by AdamW [42] with a batch size of 8. For each batch, we train the model with different  $\lambda$  to achieve rate-variable compression.

**Evaluation dataset.** We evaluate GLC on CLIC 2020 test set [51] with original resolution for natural image compression, and evaluate on CelebAHQ [30] with a resolution of

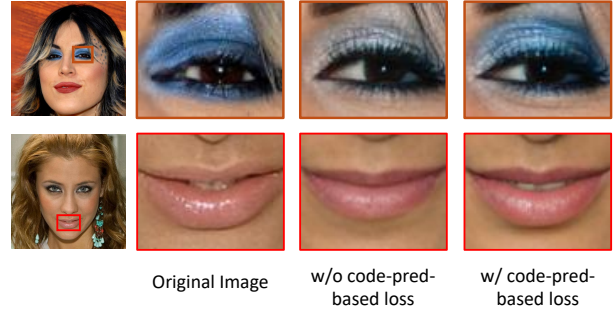


Figure 7. An example of using code-prediction-based supervision after stage II. Code prediction loss enhances the semantic consistency of the compressed latent, such as the color of eye shadow and the opening of the mouth.

$512 \times 512$  for facial image. We also show the results on Kodak [32], DIV2K [1] and MS-COCO 30K [40] in the supplementary material.

**Evaluation metrics.** We measure bit-stream size by bits per pixel (bpp), and measure visual quality by reference perceptual metrics LPIPS [29] and DISTS [15] and no-reference perceptual metrics FID [23] and KID [6]. We also provide PSNR and MS-SSIM [56] results in the supplementary material for completeness. Nevertheless, it is worth note that these pixel-level distortion metrics PSNR, MS-SSIM and LPIPS have strong limitations when evaluating image compression at ultra-low bitrate, which is also mentioned in other works [15, 35]. We provide a clearer demonstration of it in the supplementary material.

**Baseline methods.** We compare with traditional codec VVC [54], neural codec TCM [41], EVC [20], and generative codec FCC [26], Text+Sketch [35], HiFiC [44], MS-ILLM [46]. As some methods do not release models for ultra-low bitrate, we either retrain or fine-tune their models to suit such low bitrate. Text+Sketch is not evaluated on CLIC since it does not support compression in high resolution. In addition, we also compare with recent works HFD [24] and PerCo [8] in the supplementary material. For facial compression, we fine-tune EVC, TCM, HiFiC and MS-ILLM using FFHQ dataset for a fair comparison.

### 5.2. Main Results

Figure 8 shows the performance of the proposal and compared methods at ultra-low bitrate. On CLIC 2020, GLC demonstrates superiority in terms of DISTS, FID and KID than other methods. Specifically, GLC saves about 45% bits compared to previous SOTA method MS-ILLM while maintaining an equivalent FID. When comparing the pixel-level metric LPIPS, GLC also achieves comparable performance with high-fidelity generative codecs such as HiFiC and MS-ILLM. On CelebAHQ, GLC outperforms all other methods across all metrics by a large margin.

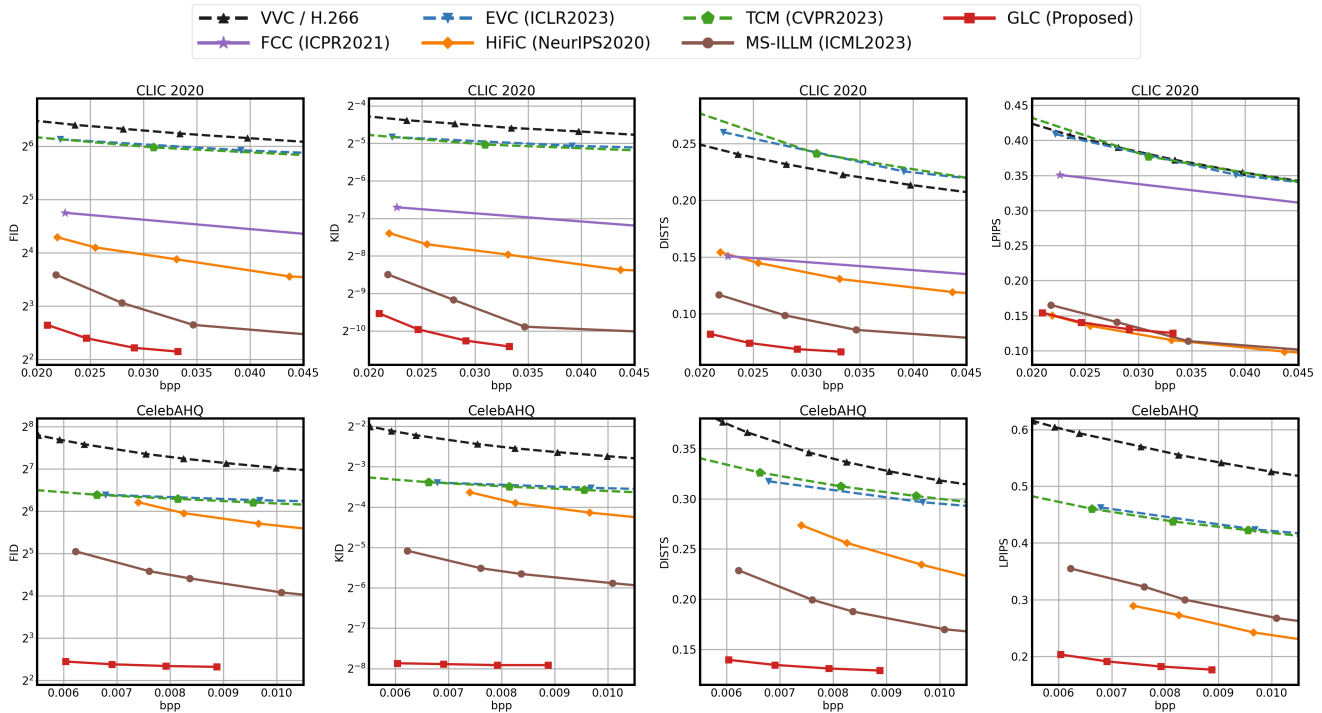


Figure 8. Comparison of methods on CLIC 2020 test set and CelebAHQ.

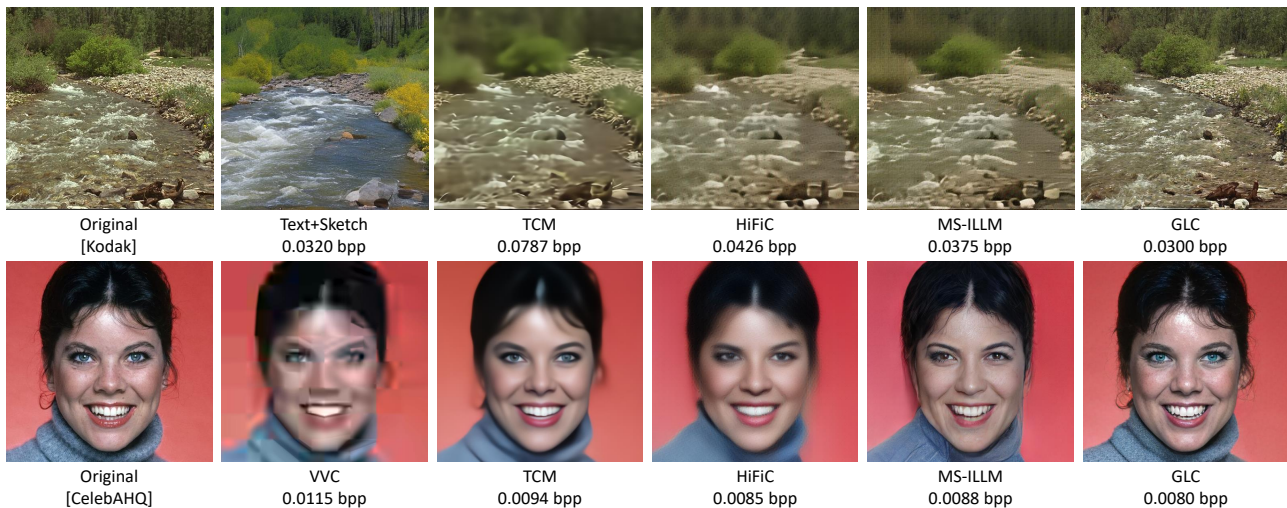


Figure 9. Qualitative examples of different methods on Kodak and CelebAHQ. More comparisons are in supplementary materials.

Figure 9 shows the qualitative comparison results. For natural image compression, Text+Sketch generates low-fidelity reconstructions, while TCM, HiFiC, and MS-ILLM produce blurry results at ultra low bitrate. In contrast, GLC achieves high-fidelity and high-realism results. In the case of facial image compression, we find existing methods cannot produce satisfactory reconstruction due to the severely distorted information at even more extreme bitrate limitation (e.g., 0.01 bpp or lower). TCM and HiFiC fall short in generating highly realistic results. Even though MS-ILLM

provides clearer details, it struggles to preserve correct facial attribute. Compared to them, GLC excels in both realism and fidelity at even lower bitrate.

### 5.3. Ablation Study

In this section, we conduct ablation studies to demonstrate the effectiveness of each proposed component. To provide a clearer comparison, we evaluate the BD-Rate [7] on the FID-BPP curve on the CLIC 2020 test set.

**Transform coding.** A straightforward approach to com-

Table 1. Ablation study on latent-space compression.

Latent coding scheme	Probability model of $z$	BD-Rate ↓
Indices-map coding	-	66.2%
<b>Transform coding</b>	<b>Factorized prior</b> <b>Categorical prior</b>	<b>17.7%</b> <b>0%</b>

Table 2. Ablation study on the code prediction module.

code prediction usage	BD-Rate ↓
w/o code pred.	13.1%
code pred. in network	60.7%
<b>code pred. as supervision</b>	<b>0%</b>

press the VQ-VAE latents is indices-map coding [27, 28, 43]. However, it causes 66.2% performance loss compared with transform coding, as shown in Table 1. It shows the effectiveness of transform coding on reducing redundancy.

**Categorical hyper module.** In Section 3.3, we illustrate the superiority of employing a categorical prior for  $z$  compared to the commonly used factorized prior. Table 1 further provides a quantitative comparison, demonstrating a significant improvement of 17.7% with such design.

**Code-prediction-based supervision.** In Section 4, we suggest employing the code prediction module as an auxiliary loss during training, instead of during the inference process of the model pipeline as in [27, 28, 60]. As shown in Table 2, incorporating the code prediction module directly into the network leads to a 60.7% performance drop. We further remove the code-prediction-based supervision for comparison, and results show that adopting the code-prediction-based supervision brings a 13.1% improvement.

## 6. Applications

Leveraging the potent latent space, our GLC pipeline opens avenues for exploring various vision applications. In this paper, we implement image restoration and style transfer as examples to show its potential. As depicted in Figure 10, for image restoration, we train a restoration encoder to map distorted images into clean latents, allowing users to directly compress a noisy image and decompress a clean one without additional cost. We compare our restoration application with a straightforward scheme of cascading an additional restoration network [57] with the GLC codec. The results in Table 3 indicate superior performance for our restoration application without the need for extra model parameters. Similarly, users can directly decode the latent into another style through a stylization decoder to achieve style transfer. We hope such versatility of the generative latent space will foster connections between image compression and other vision tasks in future research.

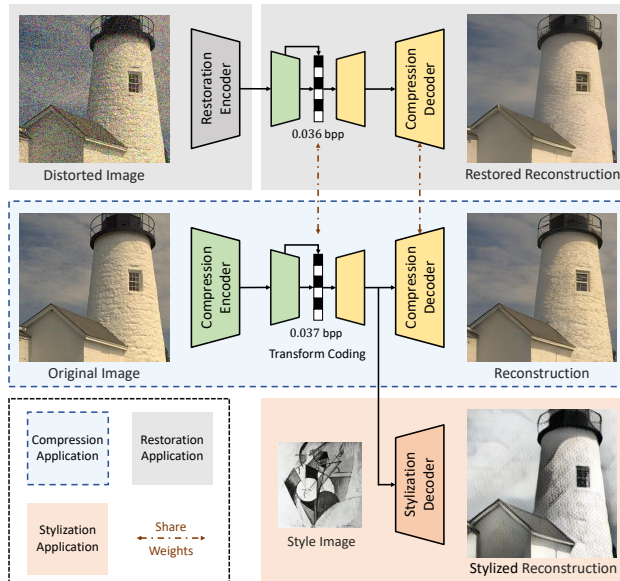


Figure 10. Generative latent applications built on GLC. In the practical image compression system, users can choose different encoders and decoders, to compress an image (medium), compress a distorted image and decompress a clear one (top), or decode the image into another style (bottom).

Table 3. Comparison for different joint restoration and compression schemes on CLIC 2020 test set.

Scheme	BPP ↓	FID ↓	DISTS ↓	Parameters
Restormer [57] + GLC Codec	0.0314	10.79	0.1174	25M + 109M
GLC Restoration Application	<b>0.0299</b>	<b>8.62</b>	<b>0.1081</b>	109M

## 7. Conclusion and Limitation

In this paper, we introduce a generative latent coding (GLC) scheme to achieve high-fidelity and high-realism generative compression at ultra-low bitrate. Unlike most existing pixel-domain codecs, GLC performs transform coding on the latent domain of a generative VQ-VAE. By incorporating a categorical hyper module and a code-prediction-based supervision, GLC demonstrates state-of-the-art performance on several benchmarks. We further develop several vision applications on the GLC pipeline to demonstrate its practical potential.

However, as a generative image codec trained on specified datasets, the generalization capability of GLC is not always satisfactory. For instance, GLC cannot guarantee a clear and accurate reconstruction of screen contents, as illustrated in the supplementary material. Future work will focus on addressing this limitation, to enhance the generalization ability of GLC by improving the model structure and the training strategy.



## References

- [1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 6
- [2] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 221–231, 2019. 3
- [3] Eirikur Agustsson, David Minnen, George Toderici, and Fabian Mentzer. Multi-realism image compression with a conditional generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22324–22333, 2023. 3
- [4] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *5th International Conference on Learning Representations, ICLR 2017*, 2017. 1, 3
- [5] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. 1, 2, 3, 4
- [6] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6
- [7] Gisle Bjontegaard. Calculation of average psnr differences between rd-curves. *ITU SG16 Doc. VCEG-M33*, 2001. 7
- [8] Marlène Careil, Matthew J Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. In *The Twelfth International Conference on Learning Representations*, 2023. 6
- [9] 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. *IEEE Transactions on Image Processing*, 31:2809–2823, 2022. 3
- [10] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prfulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016. 3
- [11] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7939–7948, 2020. 1, 3
- [12] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 3
- [13] Ze Cui, Jing Wang, Shangyin Gao, Tiansheng Guo, Yihui Feng, and Bo Bai. Asymmetric gained deep image compression with continuous rate adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10532–10541, 2021. 3
- [14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6
- [15] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 1, 6
- [16] Alaaeldin El-Nouby, Matthew J Muckley, Karen Ullrich, Ivan Laptev, Jakob Verbeek, and Herve Jegou. Image compression with product quantized masked image modeling. *Transactions on Machine Learning Research*, 2022. 3
- [17] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1, 2, 3, 4, 5
- [18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1, 5
- [19] Vivek K Goyal. Theoretical foundations of transform coding. *IEEE Signal Processing Magazine*, 18(5):9–21, 2001. 1, 3
- [20] Wang Guo-Hua, Jiahao Li, Bin Li, and Yan Lu. Evc: Towards real-time neural image compression with mask decay. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 3, 6
- [21] Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14771–14780, 2021. 4
- [22] Dailan He, Ziming Yang, Hongjiu Yu, Tongda Xu, Jixiang Luo, Yuan Chen, Chenjian Gao, Xinjie Shi, Hongwei Qin, and Yan Wang. Po-elic: Perception-oriented efficient learned image coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1764–1769, 2022. 3
- [23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6
- [24] Emiel Hoogeboom, Eirikur Agustsson, Fabian Mentzer, Luca Versari, George Toderici, and Lucas Theis. High-fidelity image compression with score-based generative models. *arXiv preprint arXiv:2305.18231*, 2023. 3, 6
- [25] Yueyu Hu, Shuai Yang, Wenhan Yang, Ling-Yu Duan, and Jiaying Liu. Towards coding for human and machine vision: A scalable image coding approach. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 3
- [26] Shoma Iwai, Tomo Miyazaki, Yoshihiro Sugaya, and Shinichiro Omachi. Fidelity-controllable extreme image compression with generative adversarial networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8235–8242. IEEE, 2021. 3, 6
- [27] Wei Jiang, Hyomin Choi, and Fabien Racadé. Adaptive human-centric video compression for humans and machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1121–1129, 2023. 2, 3, 4, 6, 8

- [28] Wei Jiang, Hyomin Choi, Fabien Racapé, Simon Feltman, and Fatih Kamisli. Face restoration-based scalable quality coding for video conferencing. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 206–211. IEEE, 2023. 2, 3, 4, 6, 8
- [29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 1, 5, 6
- [30] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 2, 6
- [31] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 6
- [32] Kodak Lossless True Color Image Suite. <http://r0k.us/graphics/kodak/>. 6
- [33] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 6
- [34] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. In *International Conference on Learning Representations*, 2018. 3
- [35] Eric Lei, Yigit Berkay Uslu, Hamed Hassani, and Shirin Saeedi Bidokhti. Text+ sketch: Image compression at ultra low rates. In *ICML 2023 Workshop Neural Compression: From Information Theory to Applications*, 2023. 2, 3, 6
- [36] Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. *Advances in Neural Information Processing Systems*, 34:18114–18125, 2021. 5
- [37] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1503–1511, 2022.
- [38] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22616–22626, 2023. 3, 4
- [39] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with feature modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 17–21, 2024*, 2024. 5
- [40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6
- [41] Jinming Liu, Heming Sun, and Jiro Katto. Learned image compression with mixed transformer-cnn architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14388–14397, 2023. 1, 3, 6
- [42] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 6
- [43] Qi Mao, Tinghan Yang, Yinuo Zhang, Shuyin Pan, Meng Wang, Shiqi Wang, and Siwei Ma. Extreme image compression using fine-tuned vqgan models. *arXiv preprint arXiv:2307.08265*, 2023. 2, 4, 8
- [44] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33:11913–11924, 2020. 2, 3, 6
- [45] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31, 2018. 1, 3
- [46] Matthew J. Muckley, Alaaeldin El-Nouby, Karen Ullrich, Hervé Jégou, and Jakob Verbeek. Improving statistical fidelity for neural image compression with implicit local likelihood models. In *International Conference on Machine Learning*, 2023. 1, 2, 3, 6
- [47] Linfeng Qi, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Motion information propagation for neural video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6111–6120, 2023. 5
- [48] 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. 3
- [49] Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu. Temporal context mining for learned video compression. *IEEE Transactions on Multimedia*, 2022. 5
- [50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5
- [51] George Toderici, Lucas Theis, Nick Johnston, Eirikur Agustsson, Fabian Mentzer, Johannes Ballé, Wenzhe Shi, and Radu Timofte. Clic 2020: Challenge on learned image compression, 2020, 2020. 2, 6
- [52] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. 3
- [53] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1, 2, 3, 5
- [54] VVC-21.2. [https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware\\_VTM/-/tree/VTM-21.2](https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tree/VTM-21.2). Accessed: 10/23/2023, 2023. 1, 6
- [55] Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991. 1
- [56] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In

*The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pages 1398–1402. Ieee, 2003. 6

- [57] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 8
- [58] Jing Zhao, Bin Li, Jiahao Li, Ruiqin Xiong, and Yan Lu. A universal encoder rate distortion optimization framework for learned compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1880–1884, 2021. 3
- [59] Jing Zhao, Bin Li, Jiahao Li, Ruiqin Xiong, and Yan Lu. A universal optimization framework for learning-based image codec. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(1):1–19, 2023. 3
- [60] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022. 2, 3, 6, 8