Generative Latent Coding for Ultra-Low Bitrate Image Compression

Supplementary Material

In this document, we provide the supplementary material for the proposed generative latent coding (GLC) scheme. This includes the detailed network structure, additional experimental results, discussion on limitations, and application details.

1. Network Structure

GLC comprises two components: a generative latent auto-encoder and a latent-space transform coding module. In this section, we will demonstrate their respective model designs.

1.1. Generative Latent Auto-Encoder

In this subsection, we introduce the model structure of the generative auto-encoder, and propose a latent patch attention mechanism for high-resolution image compression.

Auto-Encoder Structure. We employ generative VQ-VAE models [5, 23] as the generative latent auto-encoder due to their generative capabilities, reconstruction semantic consistency, and sparse latent space. For the natural image codec, we adopt the same structure as VQGAN [5], with a latent resolution of $f = \frac{1}{16}$ of the original images and a codebook size of $M = 16384$. In the case of the facial image codec, we utilize a modified version from CodeFormer [23] with $f = \frac{1}{32}$ and $M = 1024$.

Latent Patch Attention. The generative VQ-VAE models employ global attentions in the latent space to capture correlations within an image. However, we observe that global attention is less effective for compressing high-resolution images, where correlations between distant objects are relatively small. To address this issue, we divide the latent representations into patches and leverage patch attention instead of global attention. As illustrated in Table 1, latent patch attention brings significant performance improvement on the high-resolution CLIC 2020 test set [20]. In this paper, we use a patch size of 32×32 by default.

1.2. Transform Coding in Latent Space

In this subsection, we introduce the details of transform coding. As depicted in Figure 2, this process involves a latent transformation that converts latent l into code y, and an entropy model to estimate the probability of $\hat{y}$ for entropy coding.

Latent Transformation. Our model design is based on the image codec presented in [14], which employs cascaded depth-wise blocks for efficient compression. We configure the channel number to $N = 256$, aligning it with the channel number of the latent l generated by the latent auto-encoder. We incorporate learned scalers q_{enc} and q_{dec} as the feature modulators to enable rate-variable compression.

Figure 1. An example of comparison between pixel-level metrics PSNR (higher is better), MS-SSIM (higher is better), LPIPS (lower is better), and image-level metric DISTS (lower is better). For each metric, the superior result is highlighted in **brown**. From the comparision, we can see that DISTS is a better reference perceptual metric than LPIPS.

Table 1. Ablation study on patch attention on CLIC 2020 test set.

Patch size	BD-Rate ↓
Global	20.8%
64×64	8.4%
32×32	**0%**
16×16	1.8%

Entropy Model. It estimates the entropy of the quantized code $\hat{y}$ through a categorical hyper module and a spatial context module. In the categorical hyper module, the codebook number M_h in the hyper codebook C_h is the same as that in the auxiliary codebook C. During inference, the indices of the hyper information $\hat{z}$ are compressed using fixed-length coding, where each code index is encoded into $\log_2 M_h$ bits. For the spatial context module, we adopt the same structure as the quantree-partition-based context module [14], which predicts the probability using the hyper prior and the previously decoded parts of $\hat{y}$.

2. Experiments

2.1. Perceptual Metrics

We assess the visual quality using reference perceptual metrics LPIPS [10] and DISTS [4], along with no-reference perceptual metrics FID [7] and KID [2]. Additionally, we include PSNR and MS-SSIM [22] for completeness.

Limitations of Pixel-Wise Metrics. It is worth noting that the pixel-level distortion metrics such as PSNR, MS-SSIM, and LPIPS have inherent limitations when evaluating image compression at ultra-low bitrates. These metrics prioritize pixel accuracy over the semantic consistency or

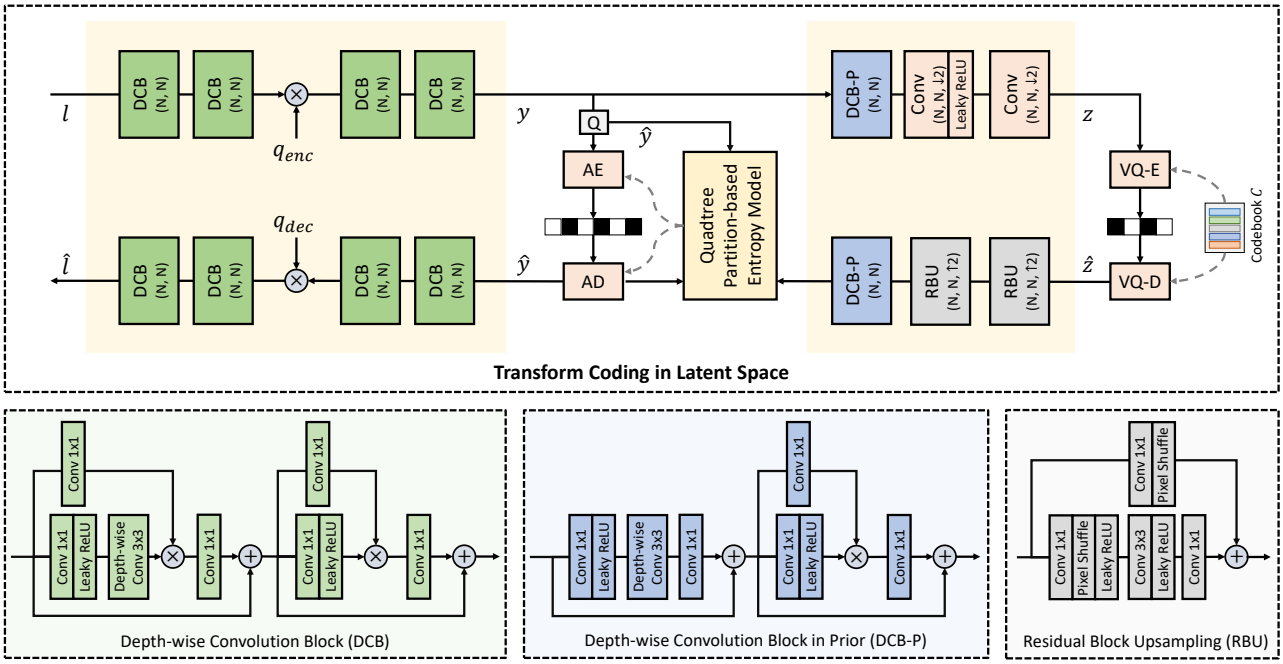

Figure 2. Structure of the transform coding module in the latent space.

texture realism, as also discussed in [4, 13]. We demonstrate this limitation with an example in Figure 1. Clearly, the image on the right is perceptually superior to the one in the middle, despite having worse PSNR, MS-SSIM, and LPIPS scores. In contrast, the image-level metric DISTS provides a more accurate assessment of image quality. For this reason, our primary focus in this paper is on DISTS, FID, and KID rather than PSNR, MS-SSIM, and LPIPS.

Measurement of FID and KID. For the facial image dataset CelebAHQ[11], FID and KID are directly calculated on all 30,000 images with a resolution of 512×512. For natural images, following established practices in generative image compression methods [17, 18], we measure them by splitting the image into 256×256 patches. Specifically, we split a $H \times W$ image into $\lfloor H/256 \rfloor \cdot \lfloor W/256 \rfloor$ patches, and then shift the extraction origin by 128 pixels in both dimensions to extract another $(\lfloor H/256 \rfloor - 1) \cdot (\lfloor W/256 \rfloor - 1)$ patches. This process yields 28,650 patches for the CLIC2020 test set [20] and 6,573 patches for the DIV2K validation set [1]. Following [17, 18], we omit FID and KID on Kodak [12] since only 192 patches are generated from the 24 images.

2.2. Quantitative Results

In this section, we present additional comparison results. In Figure 7, we compare GLC with other methods VVC [21], TCM [16], EVC [6], FCC [9], Text+Sketch [13], HiFiC [17] and MS-ILLM [18] on Kodak [12] and DIV2K validation set [1]. Figure 8 displays results on PSNR and MS-SSIM.

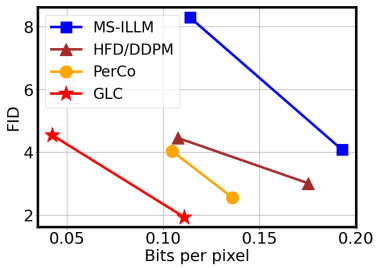

Figure 3. Comparison results on MS-COCO 30K.

Despite the limitations of these pixel-space metrics in evaluating perceptual quality, which has been discussed in Section 2.1, they are still included for completeness. Results for Text+Sketch [13] on Kodak are not shown in the figure due to its significant deviation from other curves, with PSNR=11.97dB and MS-SSIM=0.3127 at BPP=0.0289.

In addition, we compare our GLC with recent works HFD [8] and PerCo [3], along with MS-ILLM, on the MS-COCO 30K dataset [15]. Following the methodology of [8], we select the same images as them from the 2014 validation set to generate 256×256 patches. To match the quality range of their models, we further train a codec around 0.12 bpp for comparison (the correspondencing latent autoencoder has $f = \frac{1}{8}$ and $M = 256$). As shown in Fig. 3, our model exhibits significant performance improvement.

2.3. Visual Results

We provide visual comparisons with other methods on Kodak (Figure 9), CelebAHQ (Figure 10), CLIC2020 and

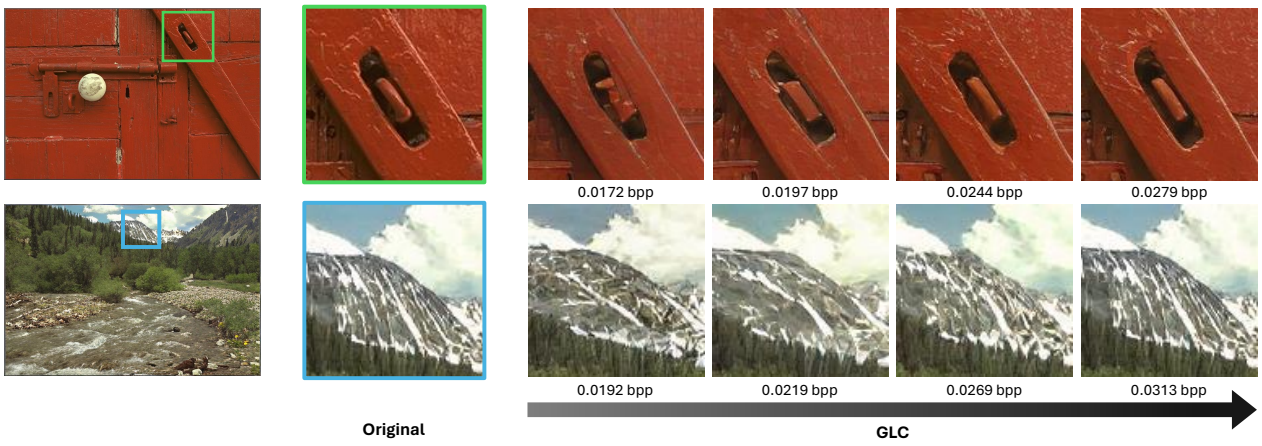

Figure 4. Examples of rate-variable compression of GLC using a single model.

Table 2. Complexity comparison for facial image on CelebAHQ with a resolution of 512 × 512.

Model	Latency (ms) Enc.	Dec.	Params	BD-DISTS
MS-ILLM	31.4	39.7	181 M	0.070
GLC	19.2	26.6	92 M	0

Table 3. Complexity comparison for natural image on Kodak with a resolution of 512 × 768.

Model	Latency (ms) Enc.	Dec.	Params	BD-DISTS
Text+Sketch	2.0×10^4	1.9×10^4	409 M	0.140
MS-ILLM	41.8	53.5	181 M	0.047
GLC	37.1	58.6	105 M	0

DIV2K (Figure 11 and 12). These comparisons reveal that GLC significantly outperforms other methods in both fidelity and realism. Additionally, we show the rate-variable characteristic of GLC in Figure 4. As the bitrate increases, GLC enhances semantic consistency and produces more intricate textures, which illustrates the impact of latent-space compression on visual quality. It should be noted that rate-variable compression is a core functionality for a practical image compression application.

2.4. Complexity

We compare the complexity of GLC with previous SOTA methods using a NVIDIA Tesla A100 GPU. The results of facial image compression on CelebAHQ are presented in Table 2, where GLC achieves a 0.070 lower BD-DISTS value and less latency compared to MS-ILLM. The results for natural image compression on Kodak are shown in Table 3, where GLC achieves a 0.047 lower BD-DISTS value and comparable latency compared to MS-ILLM, and achieves a 0.140 lower BD-DISTS value and much less latency compared to Text+Sketch. It is worth note that we do not consider the cost of the caption generation process in Text+Sketch.

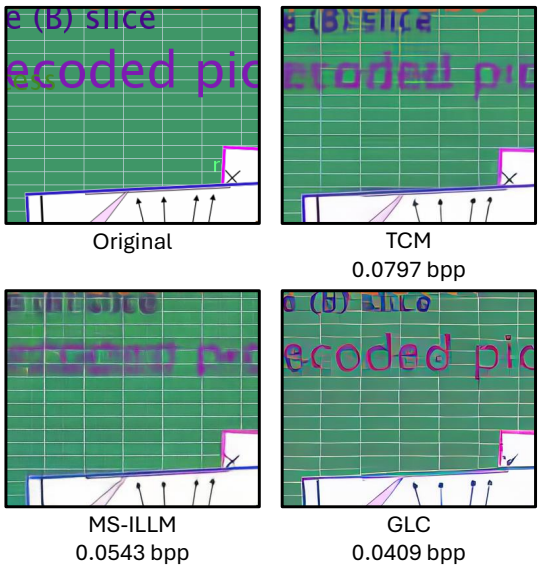

Figure 5. Generalization test on a screen image.

3. Discussion on Limitations

While the proposed GLC demonstrates superior performance in natural and facial images, its generalization ca-

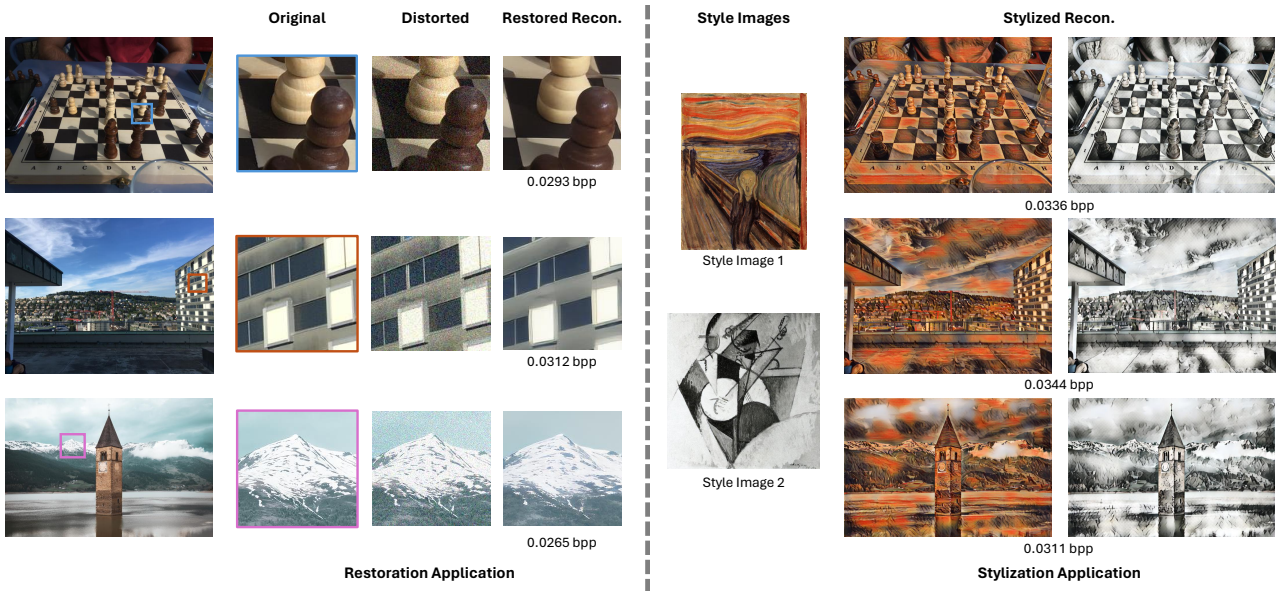

Figure 6. Examples of the restoration and stylization application implemented on GLC pipeline. The distortion is Gaussian noise with $\sigma = 20$. The first style image is sourced from the Wikiart dataset [19], and the second is *The Scream* by Edvard Munch, 1893.

pability is not always satisfactory. For instance, it may not achieve comparable quality for screen images, which is a common but significant challenge for image compression. As shown in Figure 5, GLC, while producing clearer results than TCM and MS-ILLM in text regions, still falls short in generating straight grid lines in the background. In the future, we hope this problem can be solved by enhancing the generalization capability of the generative latent autoencoders or employing a more suitable training strategy for GLC.

4. Applications

In this section, we demonstrate the details of the proposed restoration application and stylization application implemented on GLC pipeline.

Restoration Application. This application integrates the restoration task into a compression system, enabling users to compress a distorted image directly into codes and then decode it for a restored reconstruction. To accomplish it, we train a restoration encoder to map the distorted images x_d into clean latents l_c. The structure of this encoder is the same as the generative latent encoder used in the compression task. Visual results for our restoration application are provided in the middle of Figure 6, where the images are distorted by adding Gaussian noise with $\sigma = 20$.

Stylization Application. This application integrates the style transfer task into the compression system, allowing users to decode images with different styles. This is achieved by training a stylization decoder to replace the latent decoder, which is supervised by both content loss and style loss [10]. As depicted in the right of Figure 6, the proposed stylization application can decode codes into different styles.

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. 2

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 1

[3] 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. 2

[4] 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, 2

[5] 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

[6] 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. 2

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a

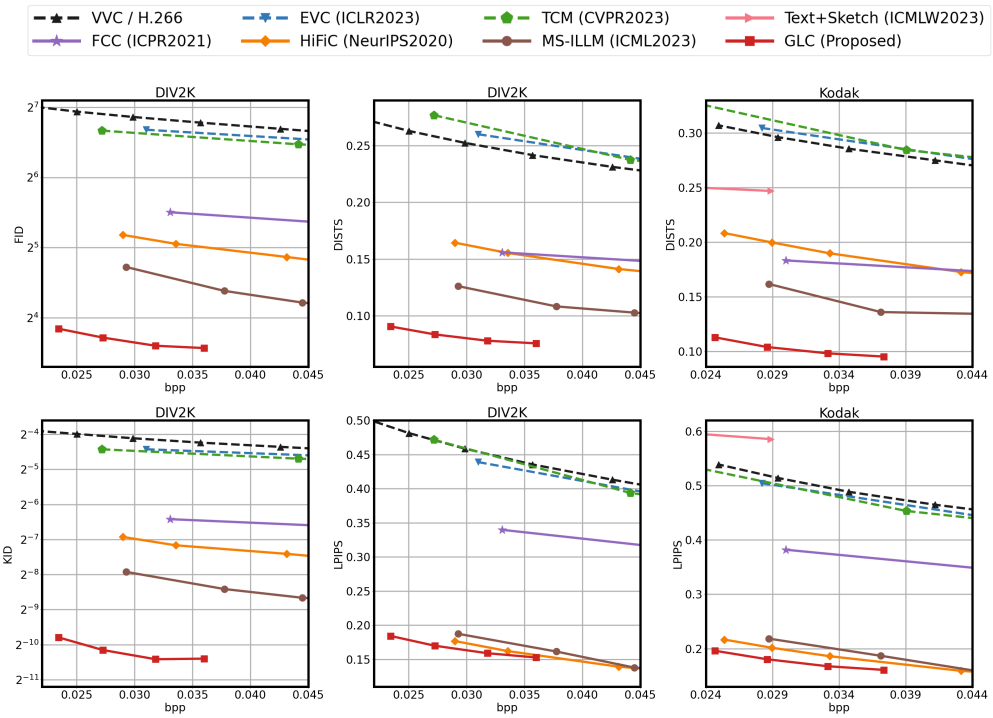

Figure 7. Comparison of methods on Kodak and DIV2K vilication set.

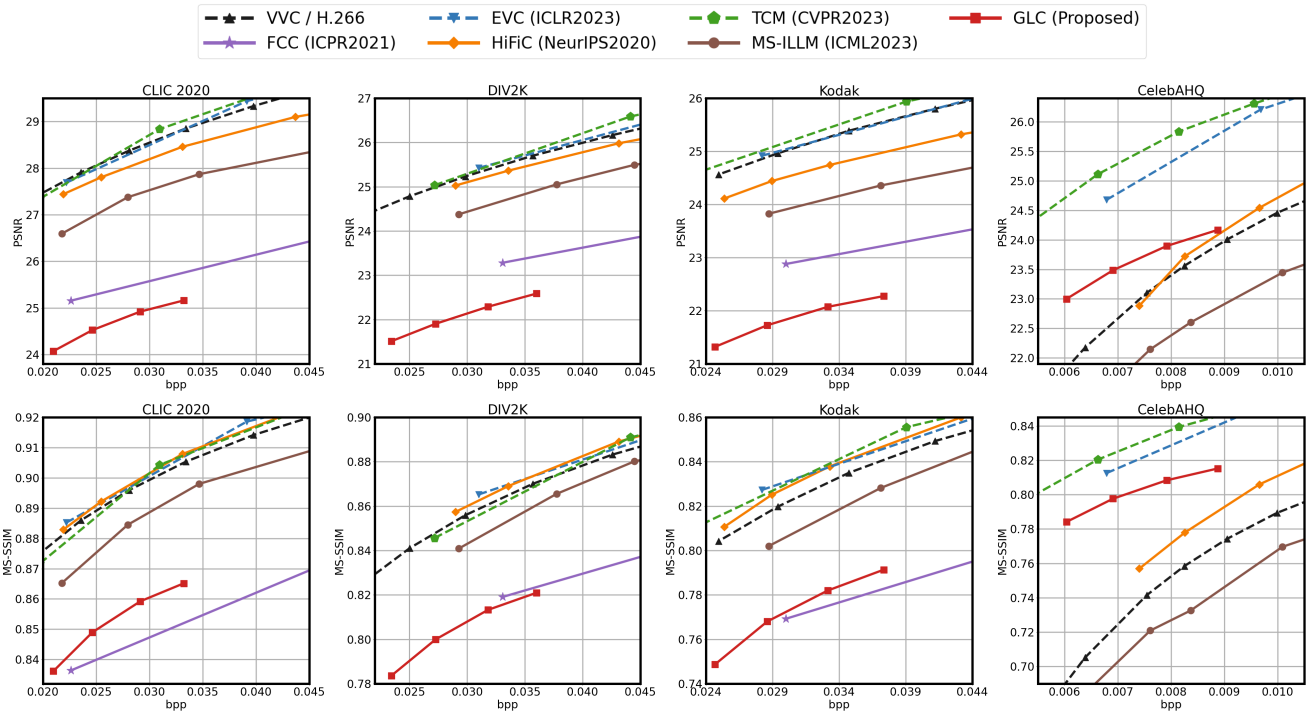

Figure 8. Comparison of methods measured by PSNR and MS-SSIM.

two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[8] Emiel Hoogeboom, Eirikur Agustsson, Fabian Mentzer,

Figure 9. Visual comparison on Kodak.

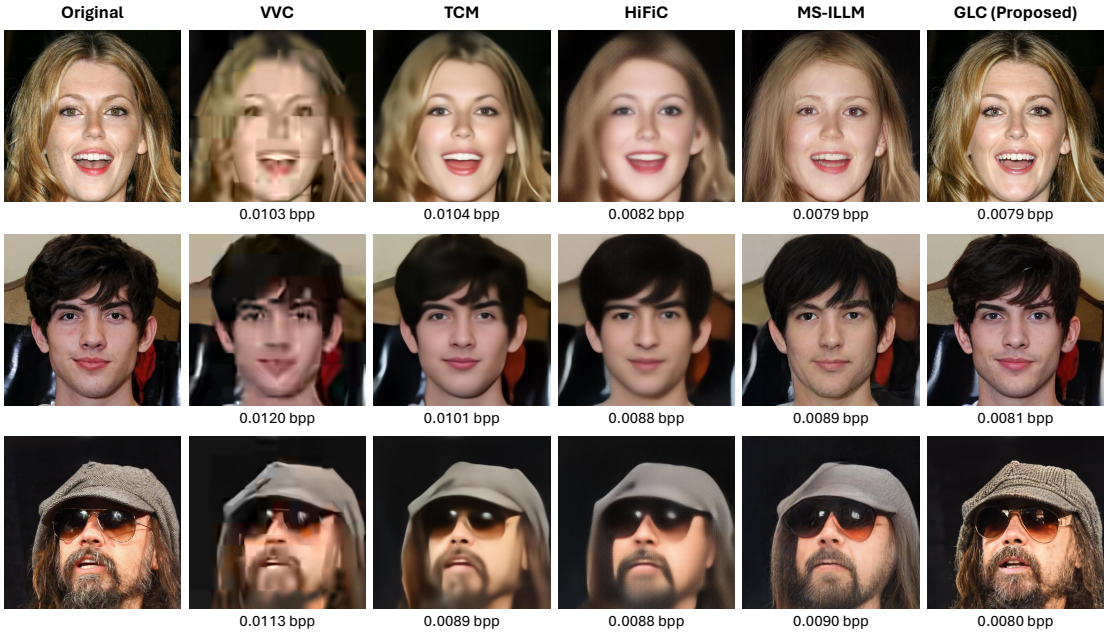

Figure 10. Visual comparison on CelebAHQ.

Luca Versari, George Toderici, and Lucas Theis. High-fidelity image compression with score-based generative models. *arXiv preprint arXiv:2305.18231*, 2023. 2

[9] 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. 2

[10] 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, 4

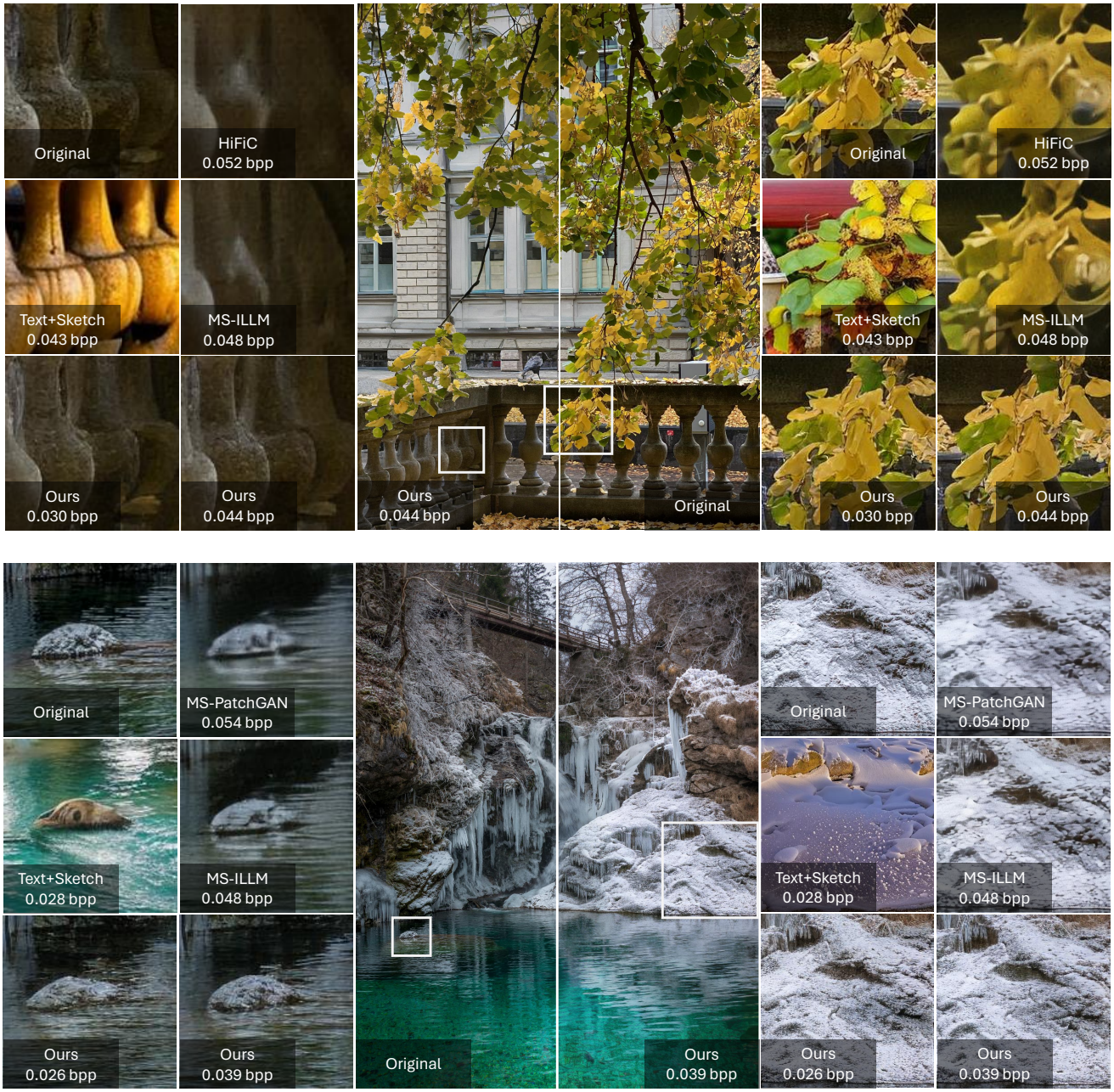

Figure 11. Visual comparison of high-resolution images in CLIC2020 and DIV2K.

[11] 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

[12] Kodak Lossless True Color Image Suite. http://r0k.us/graphics/kodak/. 2

[13] 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

[14] 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. 1

[15] 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,*

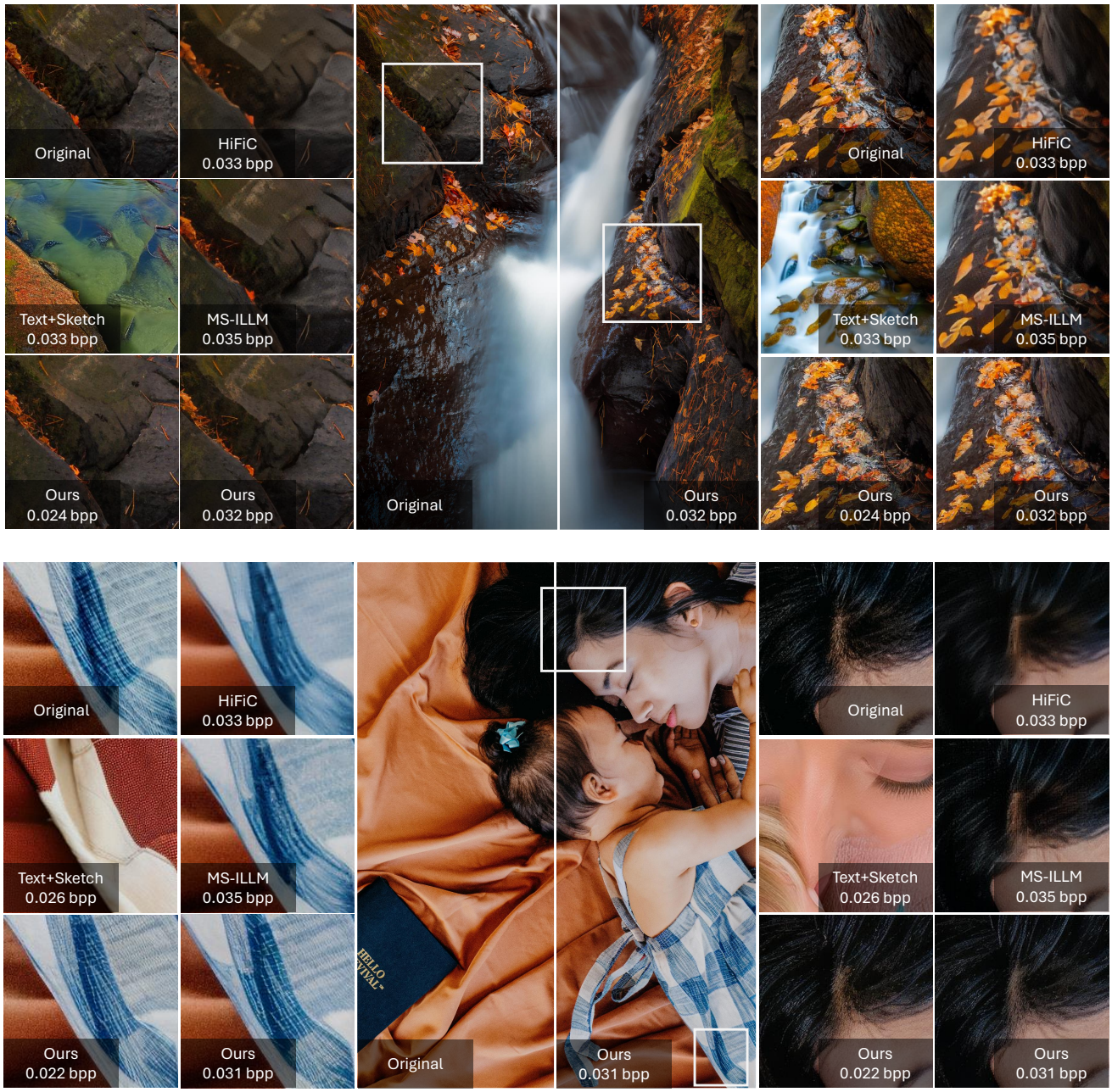

Figure 12. Visual comparison of high-resolution images in CLIC2020 and DIV2K.

Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. 2

[16] 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. 2

[17] 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

[18] 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. 2

[19] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015. 4

[20] 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. 1, 2

[21] VVC-21.2. https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tree/VTM-21.2. Accessed: 10/23/2023, 2023. 2

[22] 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. 1

[23] 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. 1