Neural Image Compression via Attentional Multi-scale Back Projection and Frequency Decomposition: *Supplementary Materials*

Ge Gao¹, Pei You¹, Rong Pan¹, Shunyuan Han¹, Yuanyuan Zhang¹, Yuchao Dai², Hojae Lee¹ ¹Samsung R&D Institute China Xi'an ²Northwestern Polytechnical University, Xi'an, China

¹{gel.gao, pei.you, rong.pan, shuny.han, yuan2.zhang, hojae72.lee}@samsung.com ²daiyuchao@nwpu.edu.cn

As denoted in the main paper, our proposed network attains substantial improvements over the current state-of-the-art methods and outperforms the advanced image compression codec VVC-intra (VTM 12.0) [2]. The supplementary materials provide additional experimental results and visual comparisons of reconstructed images, which are not included in the submitted paper due to the page limitation.

1. Performance Evaluation

1.1. Performance on the Tecnick Image Set

We further tested our model on the SAMPLING test set (100 RGB images, resolution: 1200 x 1200) of the Tecnick image set [1]. The subplot (a) in Fig. 1 shows that our model consistently outperforms previous neural compression methods and VVC-intra over the entire bpp range.



Figure 1. RD curve comparison. (a) The RD curve comparison regarding PSNR on the Tecnick image set. (b) The RD curve comparison regarding LPIPS on the Kodak image set. (c) The RD curve comparison regarding LPIPS on the 2020 CLIC Validation set. (d) The RD curve comparison regarding LPIPS on the Tecnick image set.

1.2. Evaluation regarding LPIPS

We quantitatively evaluated the quality of images reconstructed by our method using a commonly used perceptual metric - LPIPS [3], which measures the distance of distorted images in the feature space. As shown in subplots (b), (c) and (d) in Fig. 1, our model outperforms VVC-intra in terms of LPIPS for all bitrates measured.

2. Subjective Quality

Four sets of comparisons are presented, as shown in Fig. 2- 5. Note that both sides of the images have been cropped to fit the page width to facilitate visual examination. In comparison to VVC-intra (VTM 12.0) [2], our MSE-optimized model generates images with better perceptual quality.

In Fig. 2, for example, the bushes enclosed by the red and orange boxes are less distorted in the decoded image by our method. We could see from Fig. 3 that the white brick region enclosed by the red box contains some artifacts in the decoded image by VTM 12.0 [1], while these artifacts have been removed by our methods. Further, our method produces petals with richer details as denoted by the orange box. In Fig. 4, the regions enclosed by the red and the orange boxes contain more faithful textures and edges than that by VTM 12.0 [1]. The texture of the hairs and the eyelashes, as shown in Fig. 5, are better restored by our method.

References

- [1] Nicola Asuni and Andrea Giachetti. Testimages: a large-scale archive for testing visual devices and basic image processing algorithms. In *STAG*, pages 63–70, 2014. 1
- [2] Gary Sullivan and Jens-Rainer Ohm. Versatile video coding. JVET-T2002, 2020. 1, 2, 3, 4, 5, 6
- [3] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 2



Figure 2. Comparison of decoded image *kodim13.png* by our method, optimized for MSE and MS-SSIM, respectively, and VTM 12.0 [2]. Top-left, the original; top-right, our method (MS-SSIM-optimized; bpp, 0.112; PSNR, 28.299; MS-SSIM, 0.978); bottom-left, VTM 12.0 (bpp, 0.135; PSNR, 31.445; MS-SSIM, 0.967); bottom-right, our method (MSE-optimized; bpp, 0.146; PSNR, 32.427; MS-SSIM, 0.977)



Figure 3. Comparison of decoded image *kodim07.png* by our method, optimized for MSE and MS-SSIM, respectively, and VTM 12.0 [2]. Top-left, the original; top-right, our method (MS-SSIM-optimized; bpp, 0.389; PSNR, 24.018; MS-SSIM, 0.957); bottom-left, VTM 12.0 (bpp, 0.414; PSNR, 25.520; MS-SSIM, 0.917); bottom-right, our method (MSE-optimized; bpp, 0.405; PSNR, 24.945; MS-SSIM, 0.932)



Figure 4. Comparison of decoded image *kodim01.png* by our method, optimized for MSE and MS-SSIM, respectively, and VTM 12.0 [2]. Top-left, the original; top-right, our method (MS-SSIM-optimized; bpp, 0.199; PSNR, 24.841; MS-SSIM, 0.955); bottom-left, VTM 12.0 (bpp, 0.236; PSNR, 26.925; MS-SSIM, 0.936); bottom-right, our method (MSE-optimized; bpp, 0.283; PSNR, 27.598; MS-SSIM, 0.956)



Figure 5. Comparison of decoded image *kodim04.png* by our method, optimized for MSE and MS-SSIM, respectively, and VTM 12.0 [2]. Top-left, the original; top-right, our method (MS-SSIM-optimized; bpp, 0.138; PSNR, 28.162; MS-SSIM, 0.963); bottom-left, VTM 12.0 (bpp, 0.087; PSNR, 30.459; MS-SSIM, 0.920); bottom-right, our method (MSE-optimized; bpp, 0.108; PSNR, 31.303; MS-SSIM, 0.939)