MambaIC: State Space Models for High-Performance Learned Image Compression

Supplementary Material

In the appendix, we provide details about evaluation metrics (Appendix A), datasets (Appendix B) and experimental settings (Appendix C). We also carry out more experimental results (Appendix D) and visualizations (Appendix E) to showcase the effectiveness of the proposed method qualitatively and quantitatively.

A. Explanation of Evaluation Metrics

PSNR. Peak Signal-to-Noise Ratio (PSNR) is a widely used metric to measure the quality of reconstructed images compared to the original images. It quantifies how much the noise (*i.e.*, distortion) has affected the quality of the image. Higher PSNR values typically indicate better quality, with less distortion or degradation in the image. In the main results, we convert PSNR into a logarithmic decibel unit for a better comparison.

MS-SSIM. Multiscale Structural Similarity Index (MS-SSIM) is an extension of SSIM (Structural Similarity Index). Concretely, SSIM evaluates the perceived quality of an image based on three main factors: luminance, contrast, and structure. The combination of these three components gives a measure of image quality that aligns more closely with human perception. Furthermore, MS-SSIM improves upon the original metric by evaluating similarity at multiple scales (resolutions) to better simulate human perception. In practical calculations, MS-SSIM combines the SSIM values from different scales using a weighted average.

B. Details about Evaluation datasets

Kodak. kodak is made up of 24 high-quality color images, each of them with 768 × 512 pixels. These images contains a diverse set of scenes, including landscapes, portraits, indoor settings, and textures, making the dataset representative of real-world visual content.

Tecnick. Tecnick consists of 100 images with 1200×1200 pixels. It is significant in evaluating image compression performance on numerous images with medium resolution.

CLIC Professional Valid. CLIC Professional Valid is a collection of images with 2K resolution proposed by the Third Challenge on Learned Image Compression. It validates the effectiveness of learned image compression approaches on high-resolution scenarios.

C. More Explanation of Experimental Settings

Detailed structure of channel-spatial context model is shown in Table A1. In the main paper, structure of hyper encoder/decoder are a stack of convolution/deconvolution, VSS block and convolution/deconvolution. The convolution in spatial and channel entropy modeling Φ and Ψ_k holds `kernel=3, stride=2` by default. In training procedure, the images are randomly cropped to 256×256. We use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate is set to 1e-4 by default. During evaluation, the image is padded to fit for the compression and all evaluations are conducted on NVIDIA A100 under the same condition.

Table A1. Detailed architecture of channel-spatial context model.

Spatial context model Φ	**Channel context model Ψ_k**
in channel: M/K (spatial, $K = 5$)	in channel: k*M/K (k^{th}, channel $k = 1, \cdots, K$)
VSS block Conv 3×3, s1, 2*M/K	VSS block Conv 3×3, s1, 2*M/K
WLA module with channel-spatial aggregation	
fixed channel: 2*M/K+2*M/K+2*M (spatial+channel+hyper) spatial reshape	
partition window size w Local Attention reverse window size w	

D. More Experimental Results

Effectiveness of SSM block in different modules. We apply SSM block as foundation block in both nonlinear transform and context model. To figure out the utility of different foundation blocks in each modules, we additional conduct experiments and comprehensively compare the results with different variants of model that is equipped with nonlinear transform/context model of CNN/Transformer/SMM structure. Results in Table A2 reveals that the structure of main transform, *i.e.*, encoder/decoder, also influences the performance and further demonstrates the effectiveness of the proposed structure incorporating SSM blocks.

E. More Attention Map Comparison

Corresponding to visualization results in the main paper, we showcase more comparison of attention map as opposed to models w/o context entropy model and w/o window-based local attention in Figure A1 and Figure A2 to further verify the effectiveness of each proposed component in MambaIC.

Table A2. Different variants of nonlinear transform architecture.

Main Transform	Hyper Transform	Context Model	Decoding Latency (ms)	BD-Rate
CNN	CNN		35.53	-3.81%
	Transformer		-	-
	State Space Model		35.64	-7.15%
Transformer	CNN		-	-
	Transformer		48.74	-7.19%
	State Space Model		37.82	-9.30%
State Space Model	CNN		-	-
	Transformer		-	-
	State Space Model		39.42	-12.52%

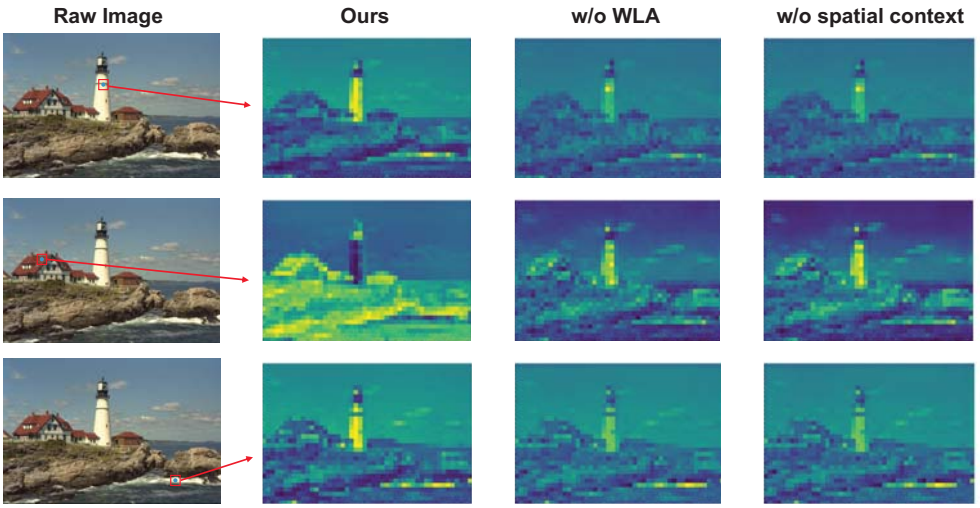

Figure A1. Attention maps of latent representations y of *kodim21.png* in Kodak.

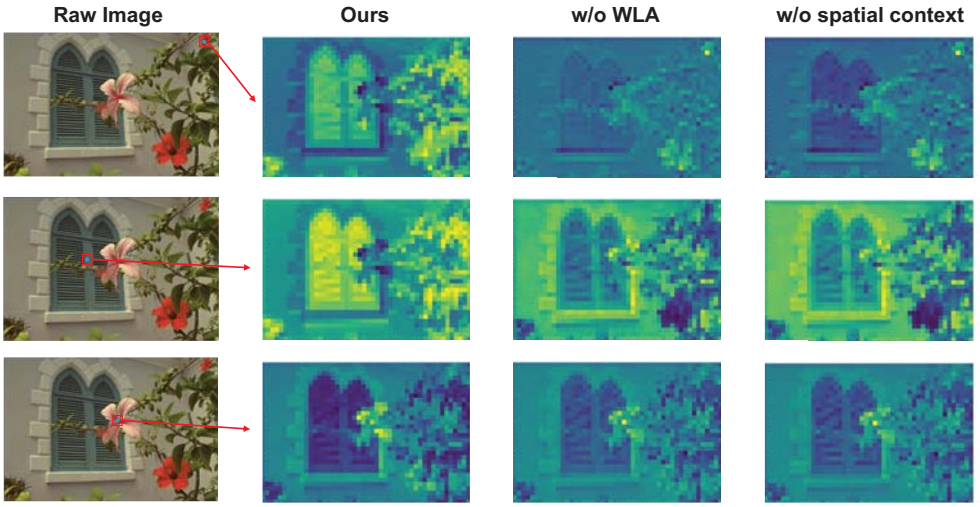

Figure A2. Attention maps of latent representations y of *kodim07.png* in Kodak.