MambaSIC: Mamba-based Stereo Image Compression with Bi-directional Multi-reference Entropy Model

Supplementary Material

More details about Section 4.3

Owing to space constraints, Section 4.3 presents a concise summary of the results. In this section, we conduct an in-depth analysis of each component.

Intra-module ablation. Variant V1 removes the cross-view control matrix $\boldsymbol{C}$ and modulation parameter $\boldsymbol{\alpha}$ in Stereo 2DSS. V2 builds on V1 by removing stereo gating, relying only on $\hat{\boldsymbol{f}}_{l\to l}$ and $\hat{\boldsymbol{f}}_{r\to r}$ without cross-view interaction—essentially reducing to the VSSL used in Vmamba [26]. V3 extends V2 by eliminating the context transform, where each view's features are split by channel and independently processed through convolutional layers and VSSL, without any cross-view interaction. Among them, V3 shows the largest performance drop, while V1 shows the least, confirming the importance of each component.

We also ablate local and global modeling by removing the convolutional branch and retaining only the Stereo VSSL. This results in BDBR increases of 8.57% and 3.48% on two benchmarks, verifying the benefit of combining local and non-local features.

Different Entropy Models. In V4, we replace our entropy model with that of single image compression [20], which only models multi intra-view priors for both anchor and non-anchor parts, without leveraging inter-view priors from the Stereo VSSB. Compared with our full model, V4 leads to bitrate increases of 11.67% and 13.01%, the most significant performance drop among all variants. These results underscore the importance of incorporating inter-view priors, which enable more accurate probability estimation and more efficient entropy coding. We also evaluate entropy models from state-of-the-art SIC methods [21, 27]. As shown in Table 5, the proposed entropy model achieves better rate-distortion performance than baselines (V5,V6 and V7) This suggests that our model provides more accurate probability estimations, which in turn minimizes the coding overhead.

Inter-view Fusion. To evaluate the effectiveness of the proposed Stereo VSSB, we consider two baselines for comparisons. We replace the Stereo VSSB with the mutual attention block in BiSIC [27] and Bi-CTM in BCSIC [21]. We apologize for the mistake in Table 5—values for V8 and V9 were inadvertently swapped. The correct results should indicate that V8 yields bit rate increases of 13.59% and 15.74% on the two datasets, while V9 results in increases of 8.81% and 9.26%, respectively. We will correct this in the final version. As shown in Table 5, our proposed model outperforms all baselines by a large margin.

Experimental Details

All training and testing settings strictly follow prior works [27, 40, 41, 47], to ensure fair comparisons. Specifically, each image in the InStereo2K dataset is pre-processed so that its dimensions are divisible by 64. For the Cityscapes dataset, rectification artifacts and the self-vehicle are removed by cropping 64 pixels from the top, 256 pixels from the bottom, and 128 pixels from each side of every image. During testing, we evaluate on images with resolutions of 1,024×832 from InStereo2K and 1,792×704 from Cityscapes.

For traditional codec baselines, BPG [5] is evaluated using the YUV 4:4:4 format to retain high visual quality. HEVC and VVC are implemented using the JVET standard. Stereo image pairs are first converted into YUV 4:4:4 videos via ffmpeg, where the left image is encoded as an I-frame and the right as a P-frame. It is worth noting that MV-HEVC only supports YUV 4:2:0, which leads to degraded PSNR performance at higher bitrates. Additionally, we reproduce BCSIC [21] and evaluate it using the same image settings as in [27, 40, 41, 47], instead of the original 512×512 resolution used in [21], to ensure comparability. The original setup in [21] yields significantly lower RD values, hence we report all results under a unified and fair evaluation protocol.

Additional Visualization Results

We visualize the qualitative results in Fig.6, Fig.7, Fig.8, Fig.9 and Fig.10, to demonstrate the effectiveness of the proposed method compared with baseline models, including VVC[6], BCSIC [21], LDMIC [47], SASIC [40], ECSIC [41], CAMSIC [48] and BiSIC [27]. Our proposed MambaSIC achieves higher PSNR at lower BPP for both the left and right views, outperforming the compared methods. Besides, the reconstruction details and texture of BiSIC are closer to the ground truth. Notably, thanks to our bidirectional design, the image qualities of the left and right views remain consistent, effectively mitigating the imbalance issue often observed in unidirectional approaches. In contrast, VVC adopts a predictive compression framework where one view is encoded independently, and the other is generated based on the disparity between the predicted and actual views. This unidirectional approach results in a PSNR gap between stereo views. ECSIC compresses the

right image using spatial context from the left image, yielding higher quality on the right view. SASIC uses the left image as a shift to assist the compression of the right image, which also results in a similar phenomenon. Compared with BiSIC, which also adopts a bidirectional structure, our method achieves a smaller PSNR discrepancy between views, indicating that the proposed Stereo VSSB is more effective than the mutual attention block in maintaining balanced reconstruction quality across views.

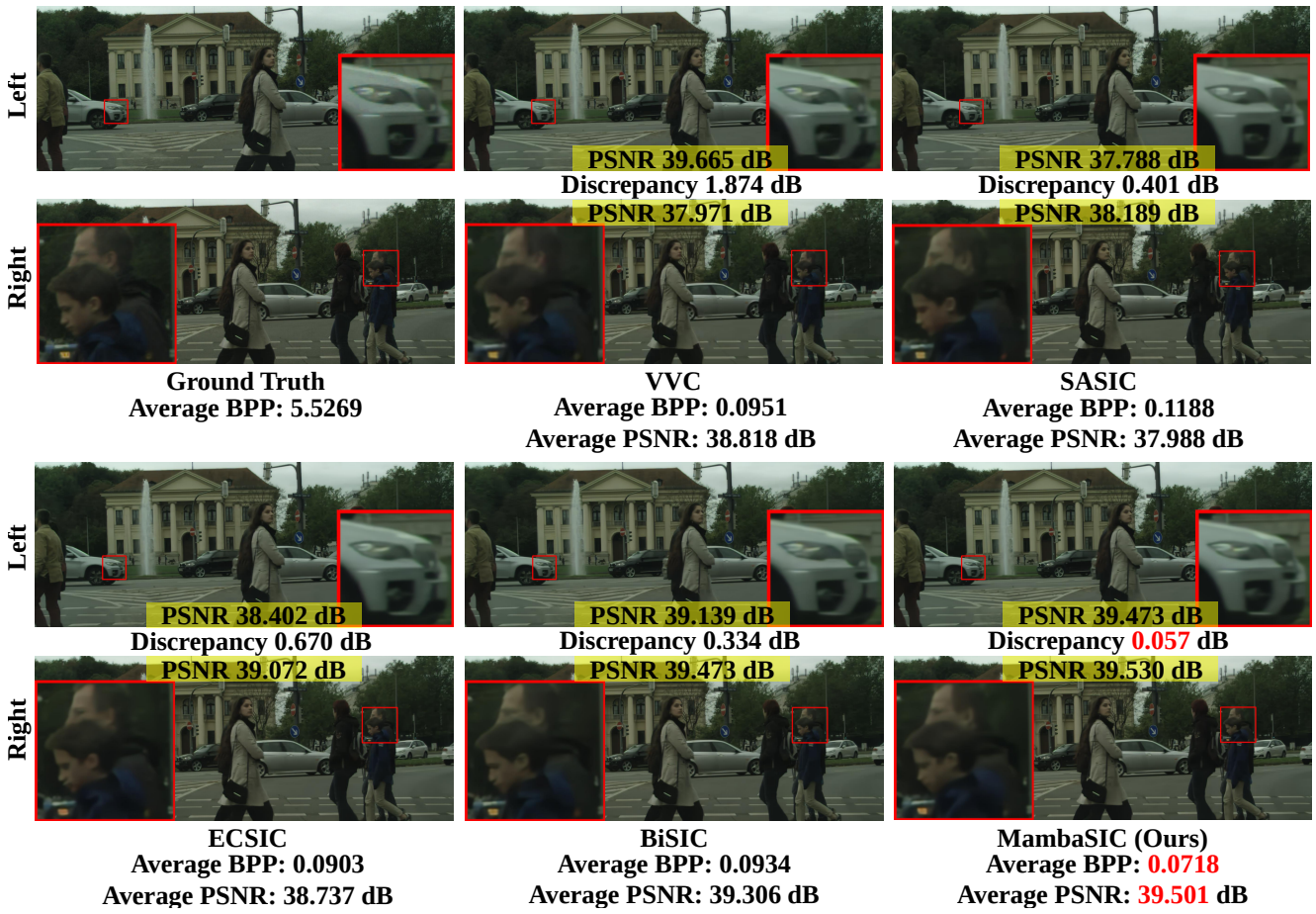

Figure 6. Qualitative comparison on reconstructed image across various codecs. Our MambaSIC achieves the lowest bit rate, the highest reconstruction quality, and the least PSNR discrepancy.

Figure 7. Qualitative comparison on reconstructed image across various codecs.

Figure 8. Qualitative comparison on reconstructed image across various codecs.

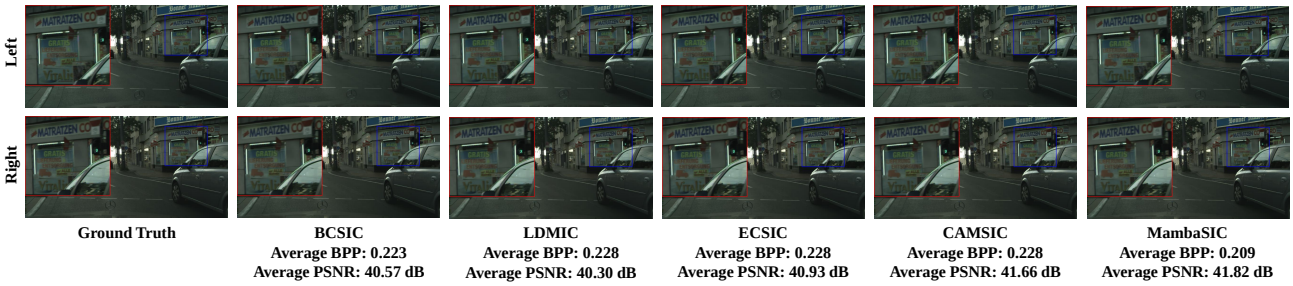

Figure 9. Qualitative comparison on reconstructed image across various codecs.

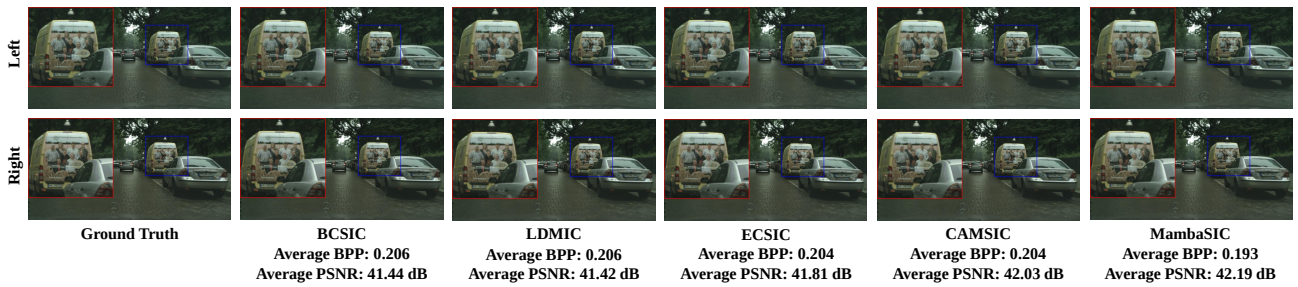

Figure 10. Qualitative comparison on reconstructed image across various codecs.