1. Supplementary Material

We provide additional results for our ablation study in Subsection 1.1, MS-SSIM [2] rate-distortion curves in Subsection 1.2, a detailed description of the layer structure of our model in Subsection 1.3, and finally additional qualitative examples in Subsection 1.4.

1.1. Additional Ablation Study Results

We provide quantitative measurements of the performance differences between different experiments from our ablation study in the main paper. We show the Bjøntegaard Delta bitrate (BD-Rate) [1] and BD-PSNR scores relative to the single image baseline in Table 1 for Cityscapes and Table 2 for InStereo2k. The description of the experiments can be found in Section 4.3 of the main paper. We also provide BD rate scores for higher quality (higher bitrates) and lower quality (lower bitrates). Our method shows consistent improvements of 37%[1] in the low bitrate range for both datasets. At high bitrates, the rate savings decrease to 19% for Cityscapes and 12.6% for InStereo2k.

1.2. MS-SSIM Rate-Distortion Curves

We provide MS-SSIM [2] rate-distortion curves in Fig. 1 for Cityscapes on the left and for InStereo2k on the right. Our method outperforms all other methods on Cityscapes and is only slightly worse than the significantly slower (due to its autoregressive component) LDMIC model on InStereo2k.

1.3. Submodules

Fig. 2 shows the layer structure of the encoder E (top left), hyperprior encoder h_E (top right), hyperprior decoder h_D (bottom right) and decoder D (bottom left). The encoder contains three downsampling steps and the hyperprior encoder contains two. The decoder and hyperprior decoder contain an equal amount of upsampling steps each. The initial three convolutional and PReLU layers in the encoder E have shared weights between left and right input stream. In these modules the left and right streams are processed in parallel and are only connected in the SCA layers.

1.4. Qualitative Results

We provide additional qualitative results for examples from the InStereo2k dataset in Fig. 3 and for Cityscapes in Fig. 4 and Fig. 5. The example images always show the right image of the stereo image pair. Our method does not show noise patterns typical for traditional methods but instead results in a smoother appearance.

[1]The maximum asymptotic theoretical bitrate saving is 50.0%, which is equivalent to compressing a stereo image pair at the bitrate of a single image. Due to occlusions and non-overlapping fields of view, the true optimum is even lower.

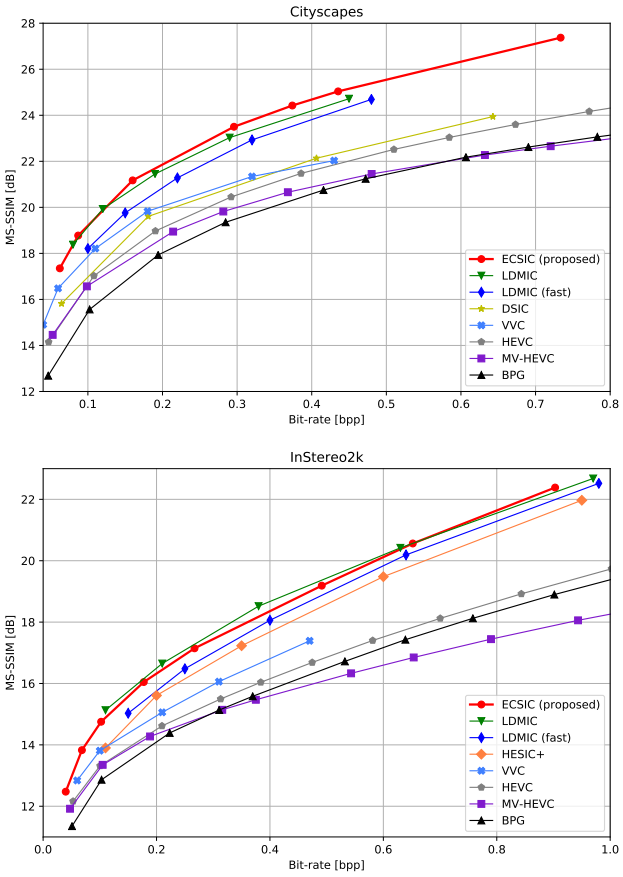

Figure 1. Rate-distortion curves for our method against other codecs for Cityscapes (left column) and InStereo2K (right column) measured by MS-SSIM.

Table 1. Relative quality difference (PSNR gain at the same bitrate; higher is better) and bitrate difference (bitrate gain for the same PSNR; lower is better) of the benchmarked methods on Cityscapes w.r.t. the baseline model. We also report the BD-Rate restricted to a low PSNR region ($34 - 38$dB) and high PSNR region ($44 - 46$dB).

Method	BD-PSNR [dB]↑	BD-Rate [%]↓	BD-Rate [%]↓ low PSNR	BD-Rate [%]↓ high PSNR
ECSIC (proposed)	**1.49**	**-30.18**	**-36.96**	**-19.02**
No context modules	1.02	-21.45	-26.95	-12.74
Only decoder SCA	0.54	-11.72	-15.57	-6.05
Only encoder SCA	0.02	-0.40	-0.36	-0.66
Baseline	0.0	-0.0	-0.0	-0.0

Table 2. Relative quality difference (PSNR gain at the same bitrate; higher is better) and bitrate difference (bitrate gain for the same PSNR; lower is better) of the benchmarked methods on InStereo2k w.r.t. the baseline model. We also report the BD-Rate restricted to a low PSNR region ($32 - 36$dB) and high PSNR region ($38 - 40$dB).

Method	BD-PSNR [dB]↑	BD-Rate [%]↓	BD-Rate [%]↓ low PSNR	BD-Rate [%]↓ high PSNR
ECSIC (proposed)	**0.77**	**-19.96**	**-37.04**	**-12.56**
No context modules	0.63	-18.20	-27.67	-9.61
Only decoder SCA	0.32	-9.36	-15.57	-5.56
Only encoder SCA	0.0	-0.70	-2.53	-0.81
Baseline	0.0	-0.0	-0.0	-0.0

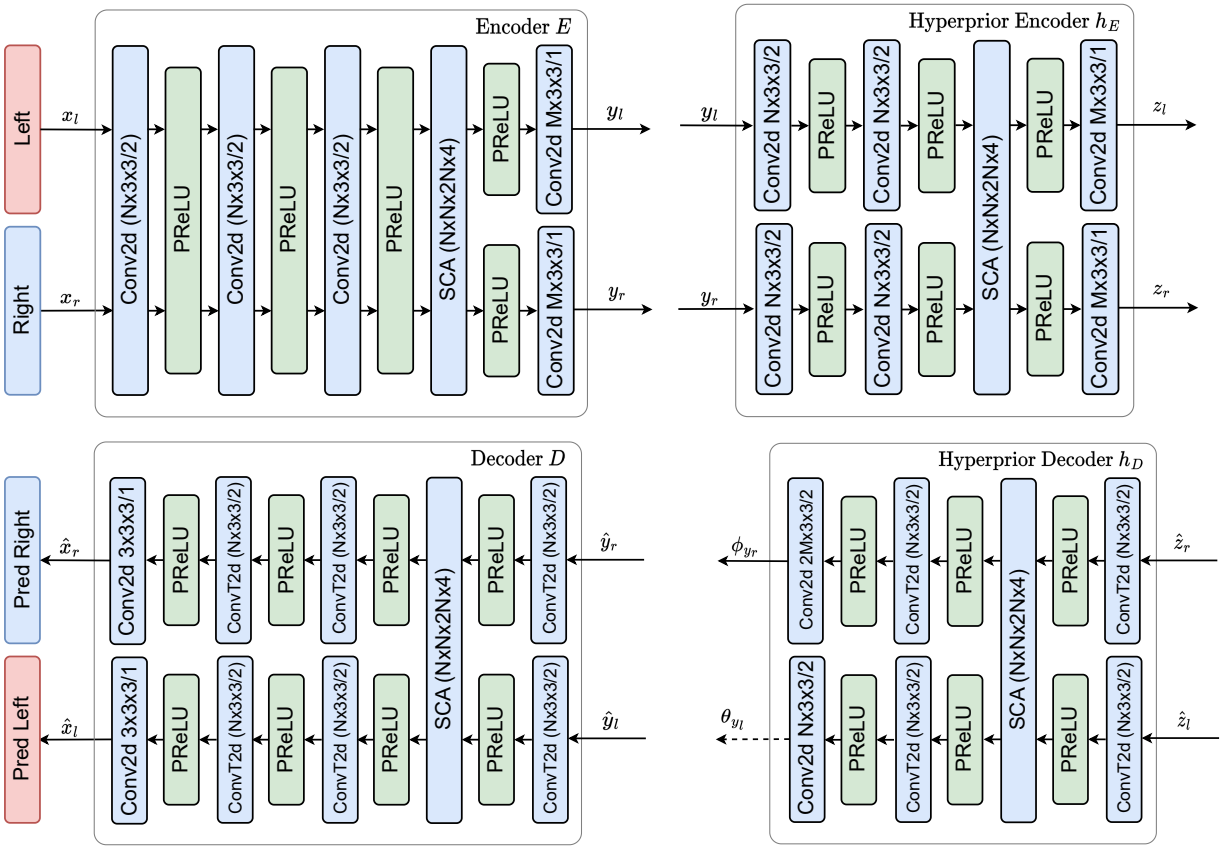

Figure 2. The left columns shows encoder E and decoder D. The right columns shows the encoder and decoder of the hyperior h_E and h_D. We set $N = 192$ and $M = 48$ for all our experiments. Conv2d denotes 2d convolutional layers and ConvT2d 2d transposed convolutional layers. The initial three convolutional and PReLU layers in the encoder have shared weights between left and right stream.

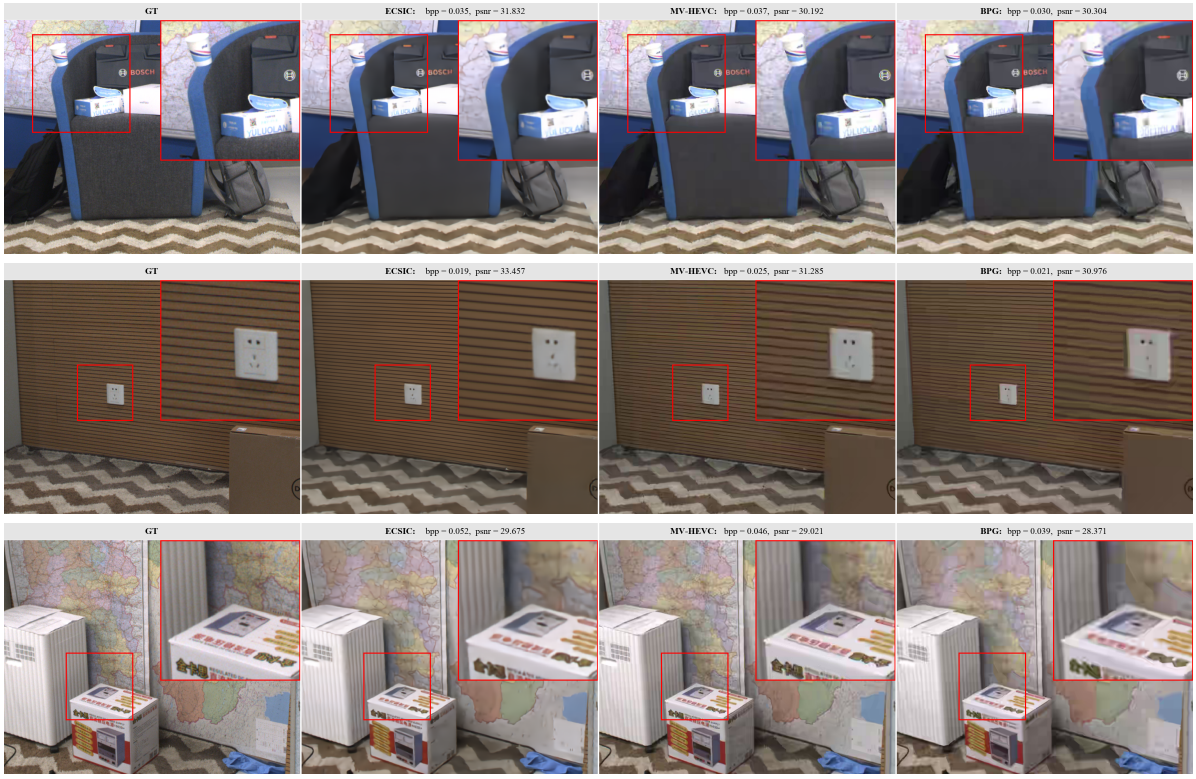

Figure 3. A qualitative comparison on images from the InStereo2K test set.

Figure 4. A qualitative comparison on images from the Cityscapes test set.

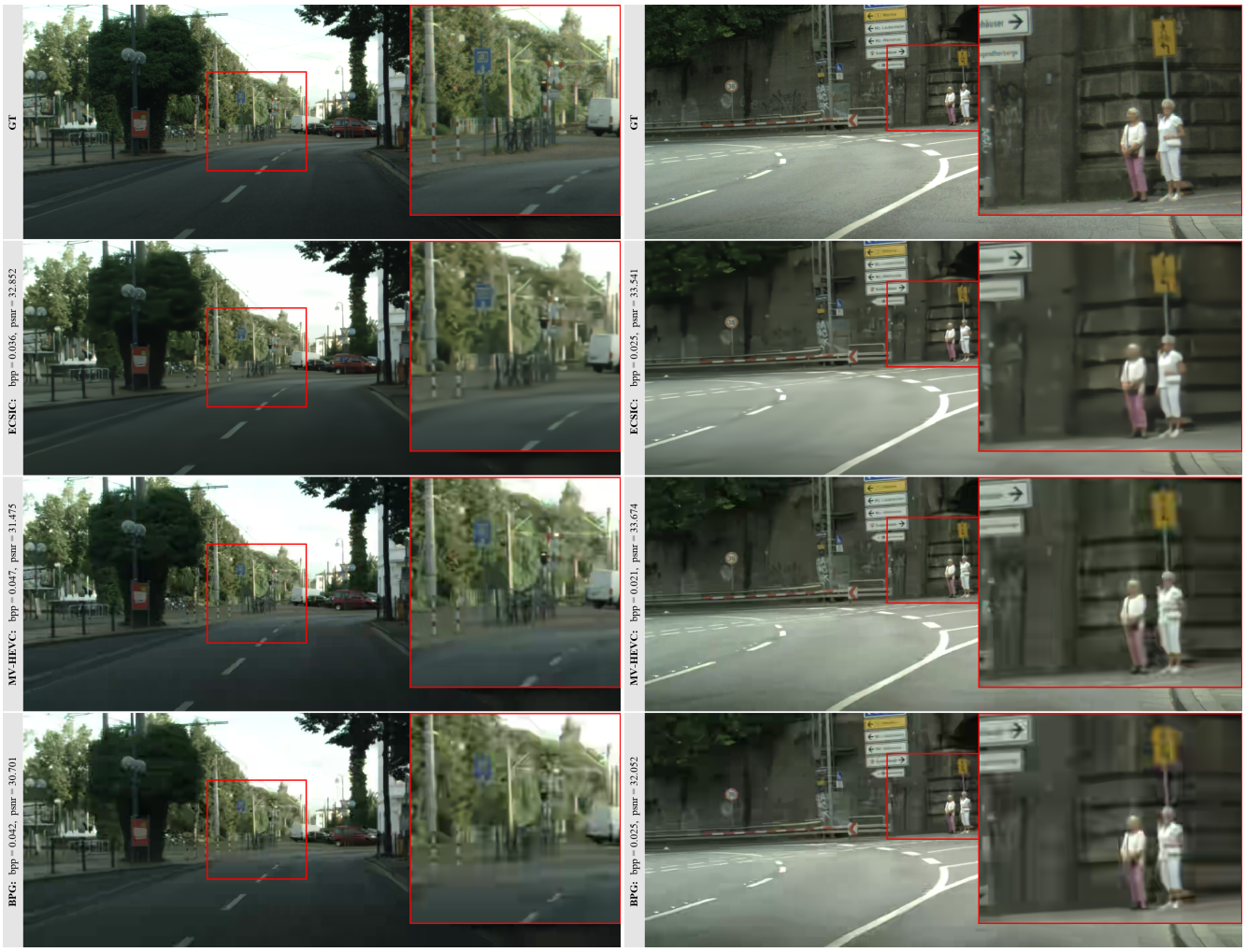

Figure 5. A qualitative comparison on images from the Cityscapes test set.

References

[1] Gisle Bjontegaard. Calculation of average PSNR differences between RD-curves. *VCEG-M33*, 2001. 1

[2] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003. 1