Variable-Rate Deep Image Compression through Spatially-Adaptive Feature Transform
Supplementary Material

Myungseo Song, Jinyoung Choi, Bohyung Han
ECE & ASRI, Seoul National University, Korea
{micmic123, jin0.choi, bhhan}@snu.ac.kr

A. Additional qualitative results

We present more qualitative results in the following three categories.

Text-preserving Figure S4 shows examples of text-preserving compression using a ROI mask.

Non-uniform quality maps Figure S5 presents compression results using non-uniform quality maps on the COCO [R6] validation set and the average bit allocation maps of quantized latent representation $\hat{y}$. All these example images were not included in our training set since we trained our model using the COCO train split.

Uniform quality maps Figure S6, S7, S8, S9 show qualitative comparison results of our model, Mean & Scale (M&S) Hyperprior model [R8], BPG (4:4:4) [R2] and JPEG (4:2:0) [R10] on several Kodak images. We adapt the compression rate of our model to that of M&S Hyperprior model by adjusting the value of the uniform quality map. Our model outperforms all other methods in terms of the visual quality and PSNR/MS-SSIM scores at similar bit rates.

B. Additional RD performance comparisons

To show the proposed model is particularly suitable for our task, we implement a variant of M&S Hyperprior model [R8] as a naive spatially-adpative variable-rate approach and present it in Figure S1. The naive model has same architecture as M&S Hyperprior model, but $(\mathbf{x}, \mathbf{m})$, $(\mathbf{y}, \mathbf{m})$ and $(\hat{\mathbf{y}}, \mathbf{w})$ are inputs to the encoder, hyper-encoder and decoder of it, respectively. We trained it with the same training setting as our model was trained with. We also report the official performances of M&S Hyperprior model, Minnen *et al*. [R9] which allows per-patch quality adaptation, and the performance of M&S Hyperprior model trained by us for 2M iterations.

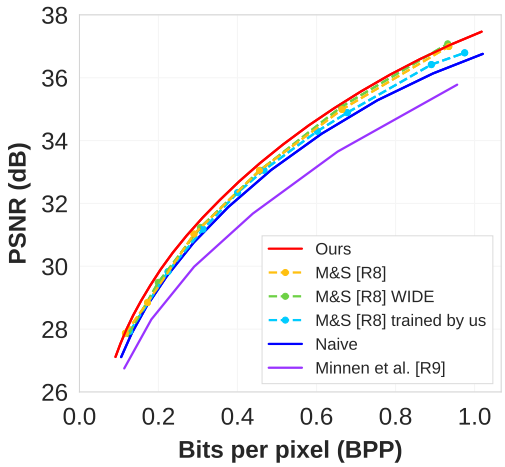

Figure S1. The rate-distortion performances on Kodak dataset. "M&S WIDE" is a wide-layer version of M&S Hyperprior model which has the similar number of parameters to that of our model. "Naive" is a naive implementation of spatially-adpative variable-rate model which is based on M&S Hyperprior model.

Compared to M&S Hyperprior model trained by us, the naive model shows performance degradation, which implies that the naive approach is not sufficient for the task of the image compression with quality map. Minnen *et al*. [R9] shows poor quality compared to other models. Meanwhile, we observe a slight drop in the performance of M&S Hyperprior model trained by us in comparison with the officially reported performance of M&S Hyperprior model. It may be because the original model was trained for about 6M iterations as the authors mentioned[1].

C. Model complexity

Table S1 compares the number of parameters and average encoding/decoding runtimes of our model and the base-

[1] https://groups.google.com/g/tensorflow-compression/c/LQtTAo6l26U/m/cD4ZzmJUAgAJ

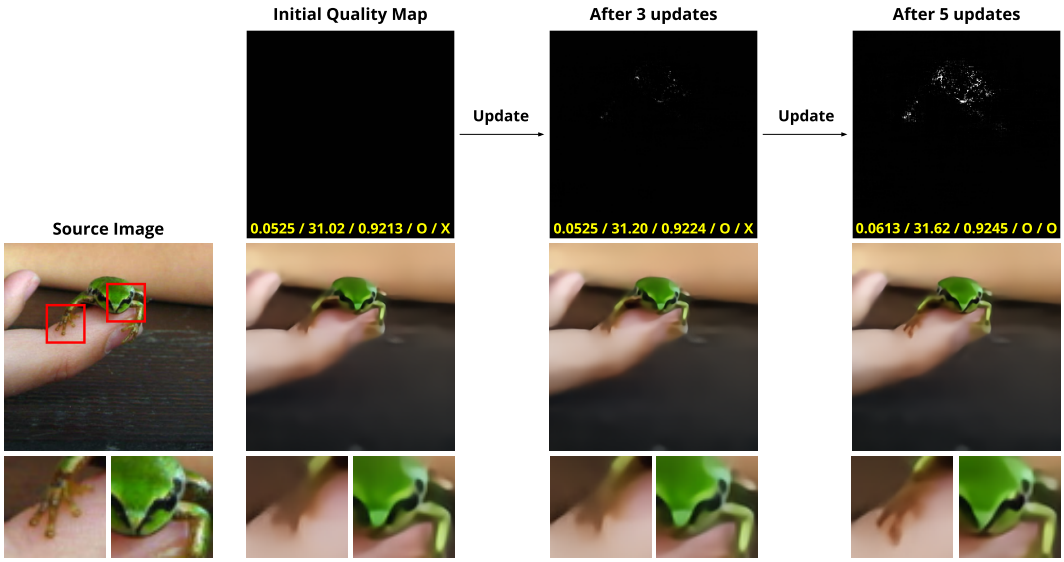

Figure S2. Classification-aware quality maps and recontructions through iterations when $\lambda = 0.01$. The annotation in each quality map denotes bpp/PSNR (dB)/MS-SSIM/Top-5/Top-1 classification result of the corresponding reconstructed image.

Table S1. Comparison of model complexity on Kodak dataset for our model and the baseline models using a GPU.

	# Parameters	Rate	Encoding (ms)	Decoding (ms)
M&S [R8]	11M	single	55	31
M&S [R8] WIDE	28M	single	143	45
M&S + Context [R8]	14M	single	5004 (82)	9876
Choi *et al*. [R4]	37M	variable	8004 (185)	12973
Ours W/O SC	27M	variable	250	37
Ours	28M	variable	254	36

line models. We additionally trained a wide-layer version of M&S Hyperprior model (304 channels) and denote it as M&S WIDE. We used a machine with a Titan Xp GPU and all models utilized the same entropy coder.

The M&S family requires multiple independent models to cover wide range of rate (*i.e*., 6 models of M&S in Figure S1), thus the actual number of the parameters is the multitude of the numbers in Table S1. M&S + Context [R8] and Choi *et al*. [R4] require much coding time than our model since they employ the autoregressive context model which have a serial coding process. The encoding of such serial context models can be made efficient by using masked convoltuion, *e.g*., for M&S + Context the encoding time decreased from 5004 ms to 82 ms. However, the decoding is inevitably slow and cannot utilize parallel processing. Our model outperforms M&S WIDE which has the similar number of parameters to ours as in Figure S1. The coding complexity of our approach increases marginally compared to the version without source conditioning (W/O SC) while the performance improves significantly. These results imply that simply increased parameters do not necessarily lead to performance gain while the proposed SFT with SC effectively improves the capacity of compression model.

D. Experiment details

D.1. Rate-distortion comparison

This section describes how we obtain the plots in Figure 6(a) in detail. For our model, we used a set of q-valued uniform quality maps ($q \in \{0, 0.05, 0.10, ..., 1.0\}$) and computed the average metrics over test images for each of 21 quality maps. Smaller spacing of q led to almost identical curves. For Choi *et al*. [R4], we extracted the RD curves from the original paper. For other methods, we used the results[2] provided by authors of [R1, R8].

D.2. Classification-aware compression

This section describes how we obtain the plots in Figure 6(b) in detail. We constructed a test set based on the ImageNet [R11] dataset by sampling 102 categories and choosing 5 images per a category randomly. We iteratively updated randomly initialized $\mathbf{m}$ by minimizing the loss for the test set:

$$\mathcal{L} = -\log P(\hat{\mathbf{y}}|\mathbf{m}) + \lambda \mathcal{L}_{\text{CE}}, \tag{1}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss. We took $\lambda \in \{0.0001, 0.001, 0.004, 0.01, 0.1, 1, 10, 100, 1000\}$ and used the results at 3 and 5 iterations for each plot. The accuracies converged to the known upper bound, the accuracies on the original images. We adopted the L-BFGS [R7] solver as an optimizer. During optimization, we used a pretrained VGG16 [R13] to compute $\mathcal{L}_{\text{CE}}$ loss while ResNet18 [R5] was used at test time to validate the generalization performance. For the Grad-CAM [R12] plots, we choose $\mathbf{m} =$

[2] https://github.com/tensorflow/compression/tree/master/results/image_compression

Table S2. Human evaluation results for 33 people when using semantic ROI masks as quality maps. Average response rates for 16 test cases are presented. Each number in parentheses indicate the quality value of ROI/non-ROI.

	Uniform (0.25/0.25)	ROI1 (0.65/0.15)	ROI2 (0.8↑/0.02)
Best	16.1%	30.5%	53.4%
Worst	72.7%	7.6%	19.7%

α**CAM** as the quality maps with $\alpha \in \{0, 0.1, 0.2, ..., 1.0\}$, where **CAM** denotes the map acquired by Grad-CAM. Figure S2 visualizes classification-aware image compression results together with how the classification-aware quality maps are acquired through iterations.

E. Human evaluation

We conducted a human evaluation to verify that different quality specifications in semantic and background regions can lead to perceptual improvement at same bitrates. We constructed 16 test cases from MSRA10K [R3] by randomly sampling original images and corresponding ROI masks. Each test case consists of the orinal image and three reconstructed images compressed with different qualtiy maps which give almost same bitrates. For each case, we asked 33 people to select the best and worst reconstructed images given the original image. The average response rates are presented in Table S2.

We observe that the higher the quality value is used in the semantic region, the more preferences occur as the best perceptual quality at the same bitrate. Similarly, for the selecitons of the worst images, the ROI-based quality maps lead to better perceptual quality than uniform quality maps. However, the votes for ROI2 as the worst images are more than those of ROI1, though the quality value in the semantic region of ROI2 is higher than that of ROI1. It implies that very poor qualtiy of the non-semantic region is not negligible for human perception.

F. Example quality maps for training

Figure S3 presents examples of quality maps we used for training. Specifically, we randomly generate the quality maps using one of four different ways for each instance in a mini-batch; (1) a uniform map (2) a semantic map of which each class label is converted to random value (3) a gradation image between two randomly selected values (4) a kernel density estimation map of Gaussian mixture with random mean, variance and number of mixtures.

G. Practical aspects of task-aware compression

One can raise some questions about feasibility and necessity of the task-aware compression; How would the task-aware compression be applied in a real-world application when a target label is not available? Why is the task-aware optimization not redundant when we already have the label? Wouldn't it be just cheaper to store the task outputs instead

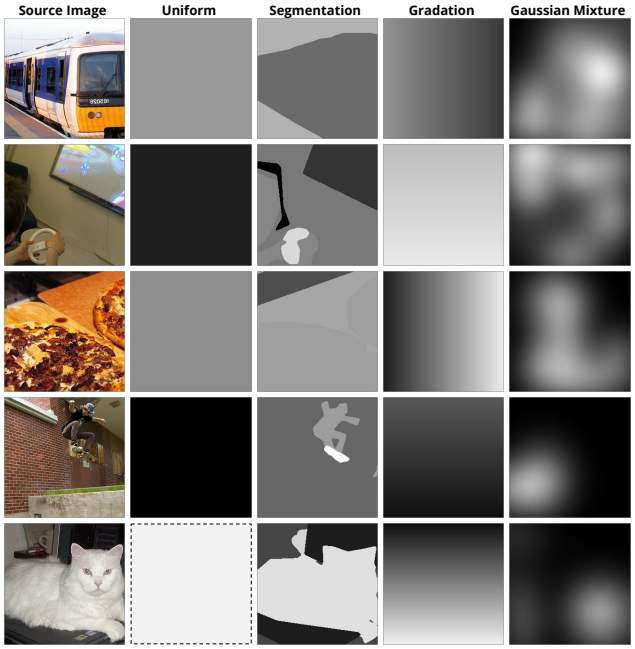

Figure S3. Examples of quality maps used for training. For each instance in a mini-batch, we randomly generated a quality map among the four types.

of optimization for the task? The goal of the task-aware compression is to reflect one's preference of spatial quality to a compressed image depending on target tasks. For example, when constructing street view images, one may want to decrease the quality of human faces while improving that of signs. In video conference applications, the quality enhancement only in human region may be required. In this respect, when the quality of particular semantic regions is important, the classification-aware compression would be useful as shown in the 5th column of Figure 2, S2. Even if obtaining a ground-truth label or calculating a task loss is unavailable, we can make an appropriate quality map using external task models, *e.g.*, ROI detection result, or Grad-CAM as shown in Figure 6(b), 7, S4 instead of optimizing the quality map. Note that for Grad-CAM in Figure 6(b), we used a class with the highest score predicted by the classifier for each test image instead of the ground-truth. Thus, the task-aware compression is still practically feasible without the task labels.

Meanwhile, the task-aware compression can be utilized when the task label itself is not a primary concern. For the example of video conference applications, we do not want to deliver the position of the human in the image or his/her personal information, but want to deliver the compressed images with high quality of the human region at the expense of the quality of background. Similarly, if we want to preserve an object region well, the classification-aware compression can be used regardless of the target label.

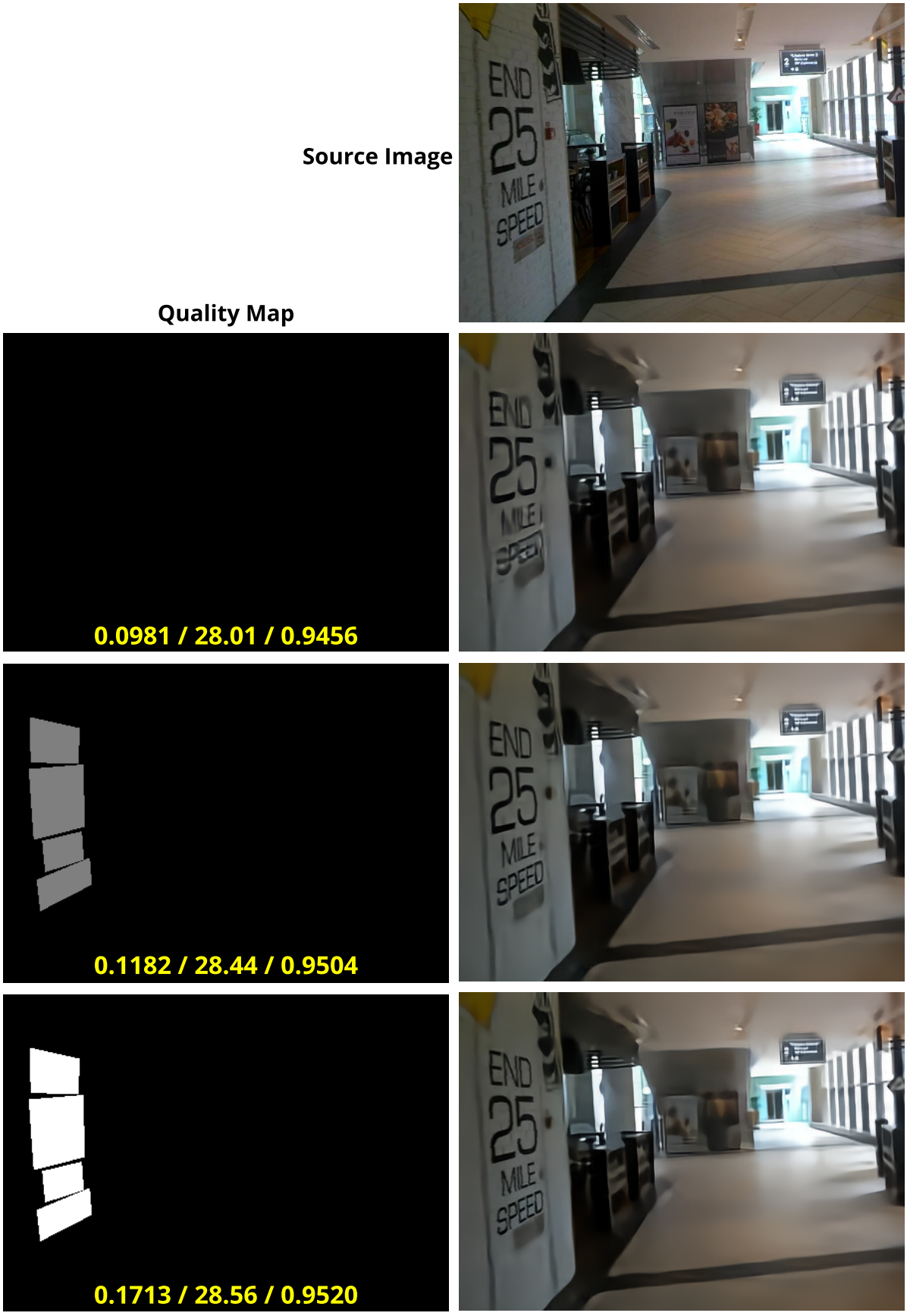

Figure S4. Reconstruction results using different quality maps (1st column) for the source image (1st row). The annotation in each quality map denotes bpp/PSNR (dB)/MS-SSIM of the corresponding reconstructed image. The ROI mask for the text was used as the quality map in 3rd and 4th row.

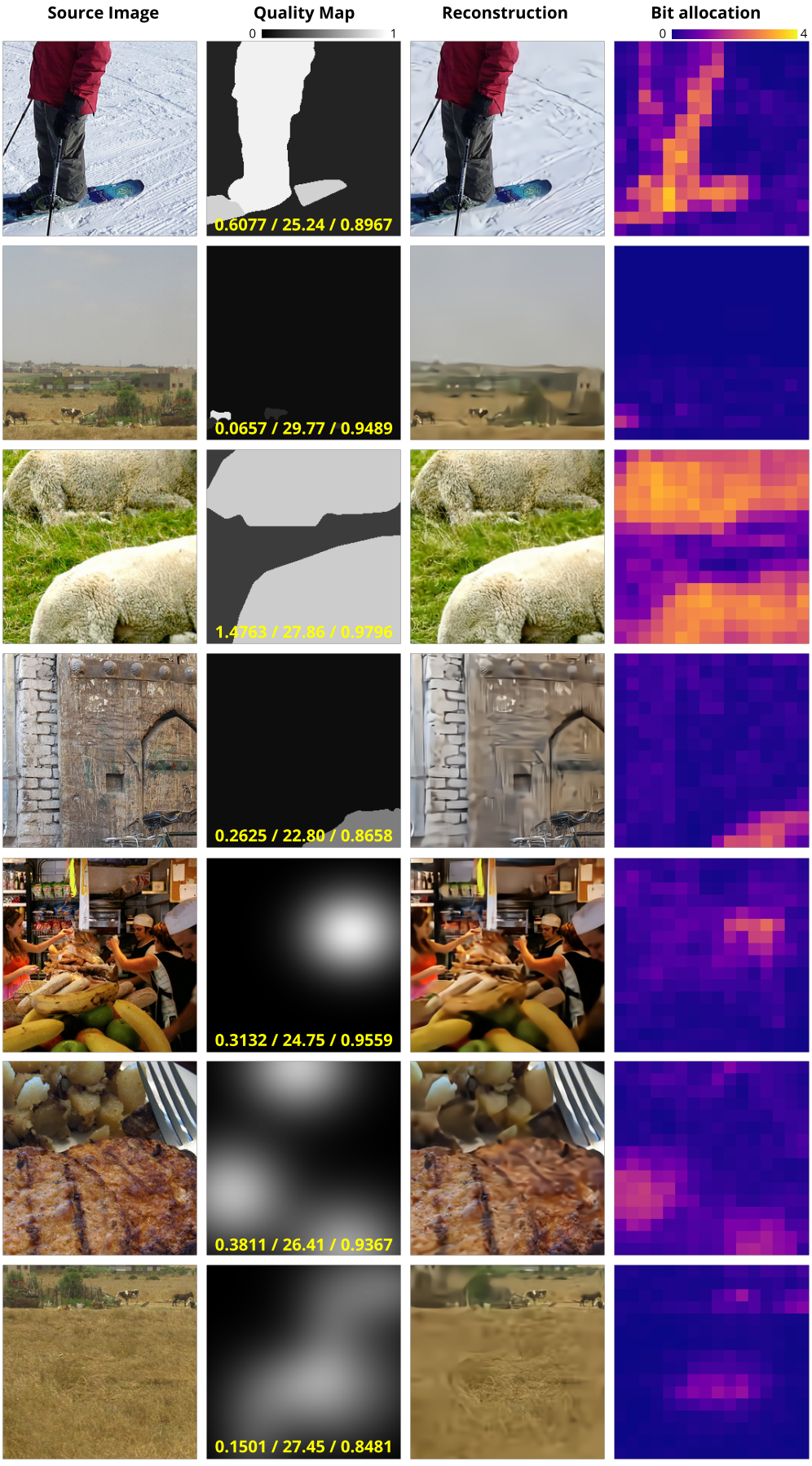

Figure S5. Examples of compression results using non-uniform quality maps.

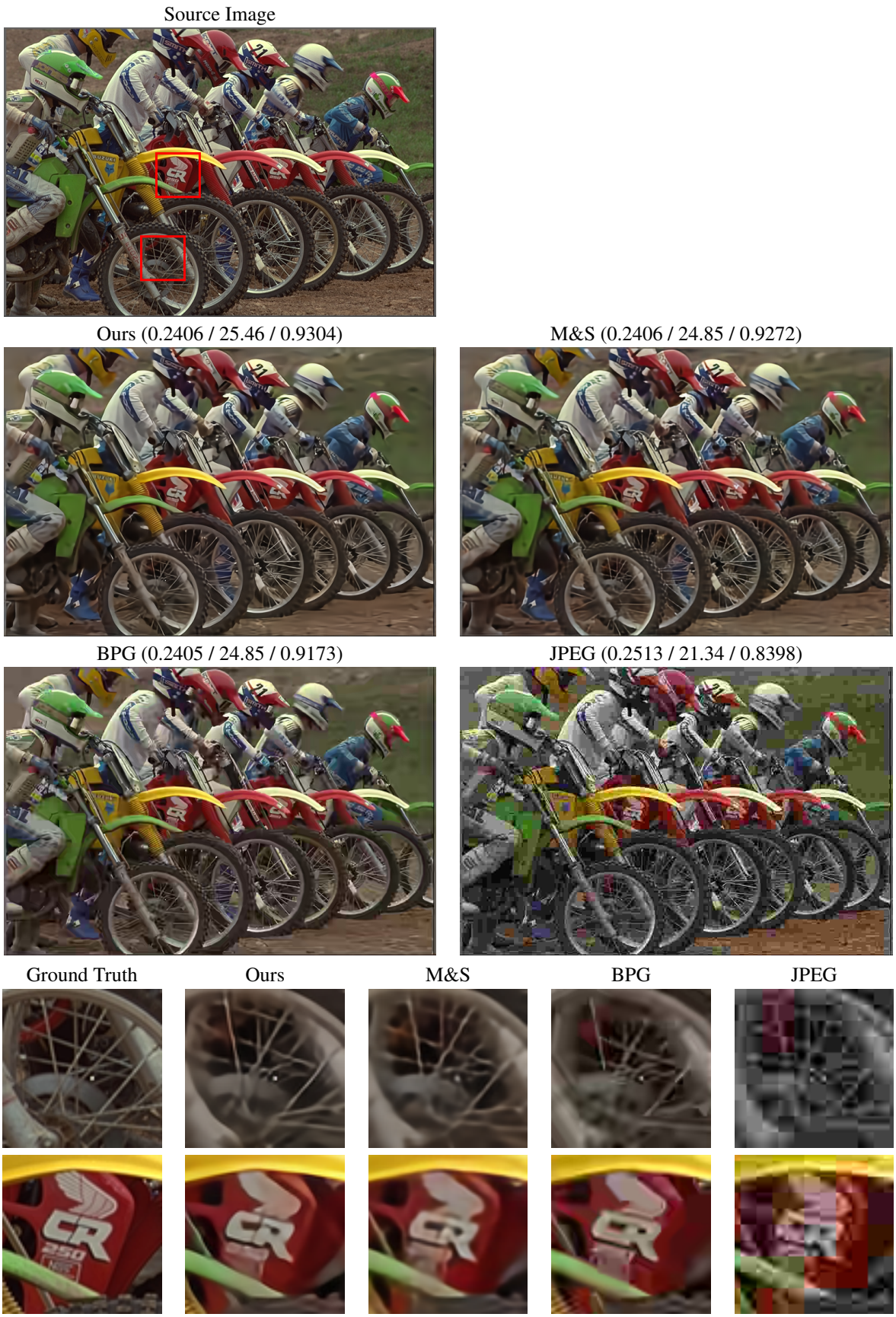

Figure S6. Compression results including the source image, our result, M&S Hyperprior model [R8], BPG (4:4:4) and JPEG (4:2:0) on Kodak 5 image. Each number in parentheses indicates bpp/PSNR (dB)/MS-SSIM of the reconstructed image. We adapt the compression rate of our model to that of M&S Hyperprior model by adjusting the value of the uniform quality map. Our model outperforms all other methods in terms of the visual quality and PSNR/MS-SSIM metrics at similar bit rates.

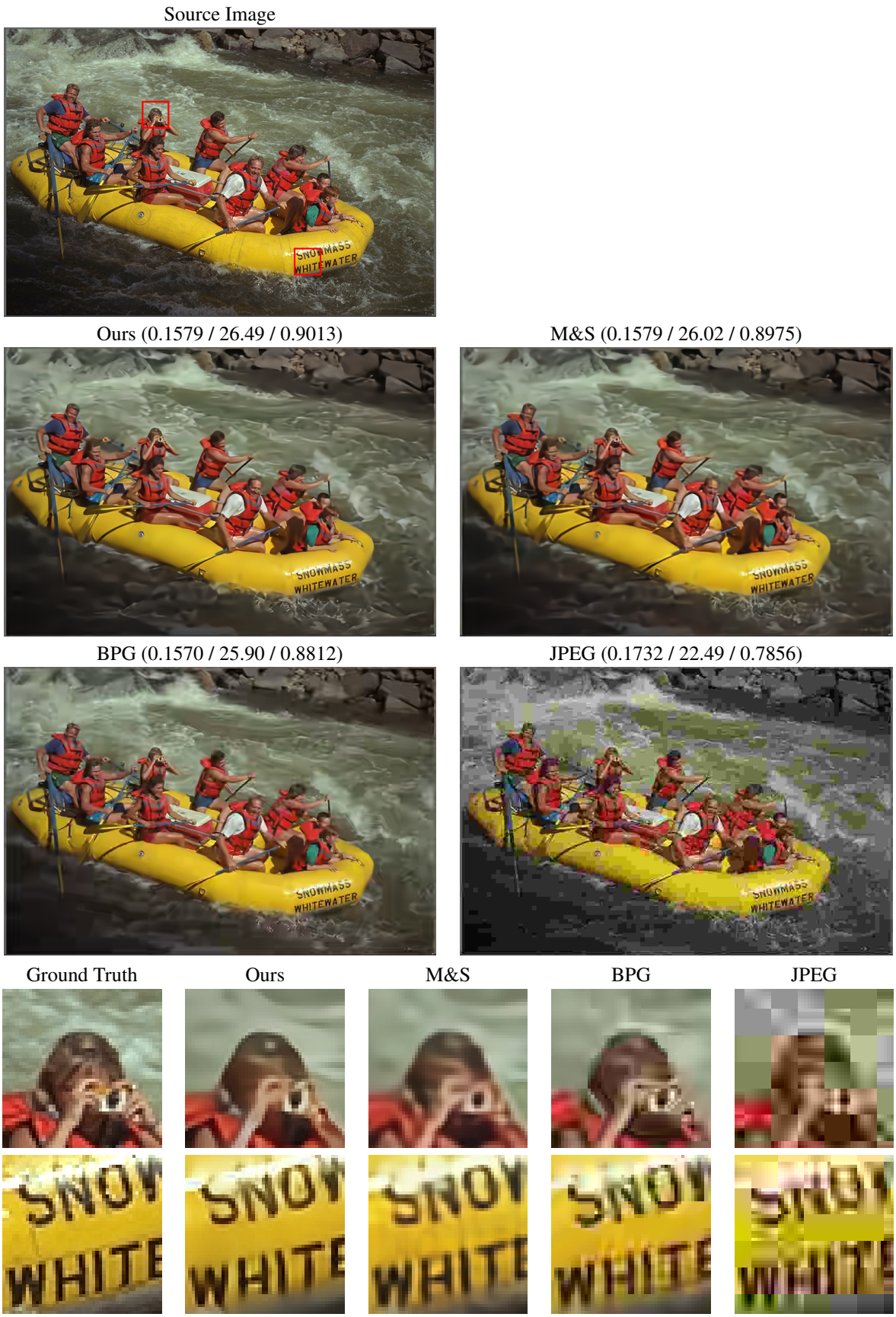

Figure S7. Compression results including the source image, our result, M&S Hyperprior model [R8], BPG (4:4:4) and JPEG (4:2:0) on Kodak 14 image. Each number in parentheses indicates bpp/PSNR (dB)/MS-SSIM of the reconstructed image. We adapt the compression rate of our model to that of M&S Hyperprior model by adjusting the value of the uniform quality map. Our model outperforms all other methods in terms of the visual quality and PSNR/MS-SSIM metrics at similar bit rates.

Figure S8. Compression results including the source image, our result, M&S Hyperprior model [R8], BPG (4:4:4) and JPEG (4:2:0) on Kodak 19 image. Each number in parentheses indicates bpp/PSNR (dB)/MS-SSIM of the reconstructed image. We adapt the compression rate of our model to that of M&S Hyperprior model by adjusting the value of the uniform quality map. Our model outperforms all other methods in terms of the visual quality and PSNR/MS-SSIM metrics at similar bit rates.

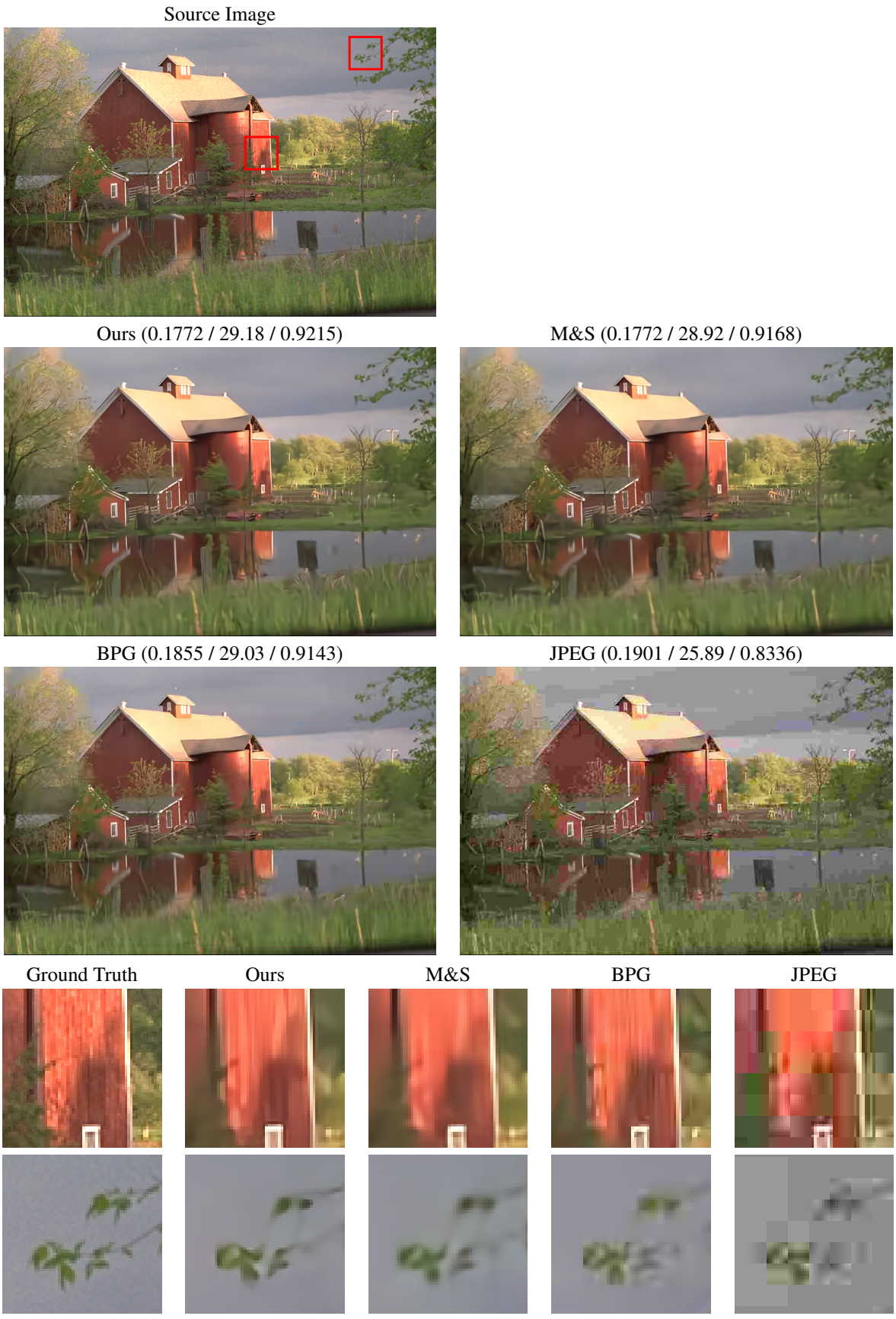

Figure S9. Compression results including the source image, our result, M&S Hyperprior model [R8], BPG (4:4:4) and JPEG (4:2:0) on Kodak 22 image. Each number in parentheses indicates bpp/PSNR (dB)/MS-SSIM of the reconstructed image. We adapt the compression rate of our model to that of M&S Hyperprior model by adjusting the value of the uniform quality map. Our model outperforms all other methods in terms of the visual quality and PSNR/MS-SSIM metrics at similar bit rates.

References

[R1] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *ICLR*, 2018. 2

[R2] Fabrice Bellard. BPG image format, 2014. 1

[R3] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *TPAMI*, 37(3):569–582, 2015. 3

[R4] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Variable rate deep image compression with a conditional autoencoder. In *ICCV*, pages 3146–3154, 2019. 2

[R5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[R6] 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 *ECCV*, pages 740–755. Springer, 2014. 1

[R7] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989. 2

[R8] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *NeurIPS*, 31:10771–10780, 2018. 1, 2, 6, 7, 8, 9

[R9] David Minnen, George Toderici, Michele Covell, Troy Chinen, Nick Johnston, Joel Shor, Sung Jin Hwang, Damien Vincent, and Saurabh Singh. Spatially adaptive image compression using a tiled deep network. In *ICIP*, pages 2796–2800. IEEE, 2017. 1

[R10] William B Pennebaker and Joan L Mitchell. *JPEG: Still image data compression standard*. Springer Science & Business Media, 1992. 1

[R11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 2

[R12] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 2

[R13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2