Raw Image Reconstruction with Learned Compact Metadata Supplementary Material

In the supplementary material, we provide more details about the implementation details, including the network design, and the details of the visualization in the main paper. Besides, we provide more ablation studies and qualitative comparison, including the reconstruction qualities under different bpp and the effectiveness of learned sampling masks, to demonstrate the effectiveness of the proposed method.

1. Implementation details

Hyper-parameter λ in Eq. 3. For the hyper-parameter λ which controls the tradeoff between the codelength and reconstruction quality defined in Eq. 3 in the main paper, we set it to 1 for uncompressed sRGB images and 0.05 for the compressed ones. We use a higher λ for the uncompressed sRGB images because it is relatively easier to do the reconstruction with high fidelity using a small codelength based on the uncompressed images.

Model architecture in Fig. 3. We further introduce details of the proposed architecture as a supplement to Fig. 3 and Fig. 4 in the main paper. For the residual connection with a different channel number in Fig. 3 of the main paper, we use an 1×1 convolutional kernel in the residual connection to make them compatible. We adopt the same attention block in [3]. In addition, $/2$ in h_a in Fig. 2 of the main paper represents the downsampling operation, *i.e.*, the resolution of auxiliary variable v is down-sampled by a factor of 4.

Quantization error map in Fig. 2. To illustrate the information loss caused by the non-linear transform and quantization step is non-uniform, we display the quantization error map in the main paper in Fig. 2. The quantization error map is estimated as follows: For a raw image $\mathbf{x}$, we first obtain its rendered sRGB image $\mathbf{y}$ without quantization. We calculate and save the mapping relationship $\mathbf{w} = \mathbf{x}/(\mathbf{y} + \epsilon)$ where $/$ is a pixel-wised division and ϵ is a small constant, *i.e.*, 1e-3. For an sRGB image after quantization $\hat{\mathbf{y}}$, the quantization error map $\mathbf{e}$ is calculated as $\mathbf{e} = |\mathbf{w} \cdot (\hat{\mathbf{y}} + \epsilon) - \mathbf{x}|$.

The estimation of bpp maps in Fig. 9. We estimate the likelihood of latent code y and z using Eq. 4 in the main paper. The number of bits can be estimated by $-\log_2 p(x)$ where $p(x)$ is the estimated likelihood. Specifically, since the spatial resolution of $\mathbf{v}$ is smaller than $\mathbf{z}$, we interpolate the estimated bpp maps of $\mathbf{v}$ to the same size as $\mathbf{z}$ and keeps the total bits of $\mathbf{v}$ unchanged. Finally, the visualization of the bits map in Fig. 9 includes bits both allocated by $\mathbf{z}$ and $\mathbf{v}$.

2. Additional results

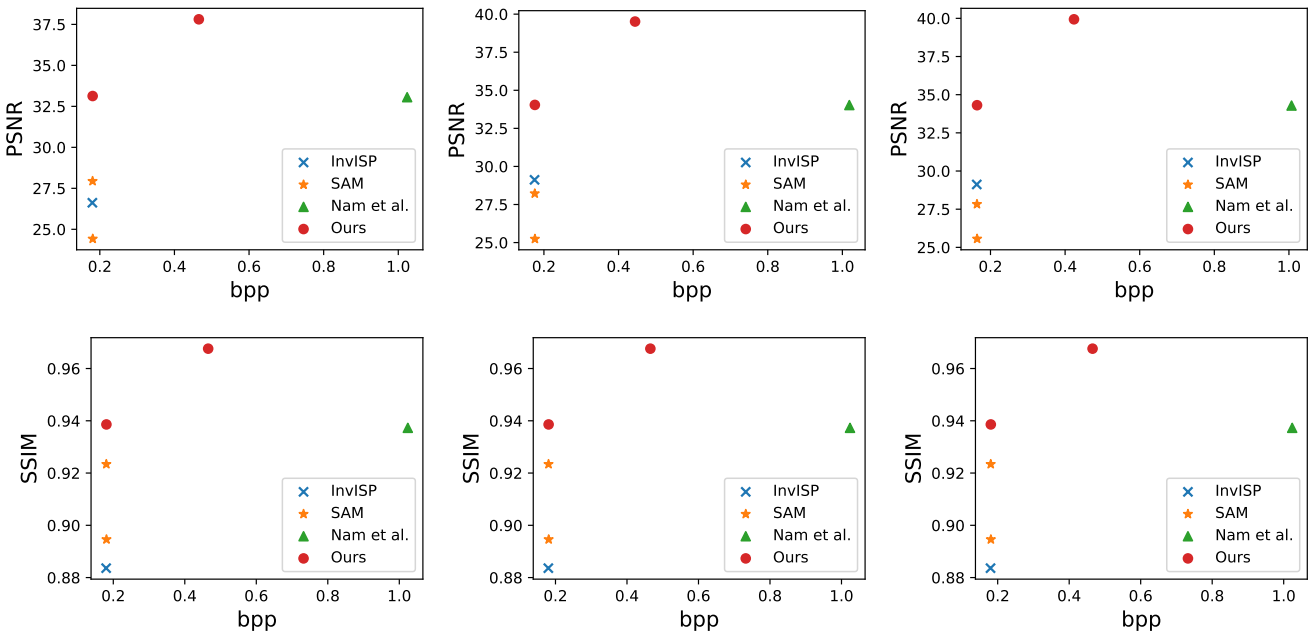
2.1. Highlights and shadows/Manipulation

For shadows, the proposed method achieves low reconstruction error and few bits are allocated to as in Fig. 2. Besides, the reconstructed RAW images achieve better exposure latitude, *i.e.*, fewer artifacts than JPEG or sometimes even the reference RAW image after contrast enhancement. More bits are adaptively allocated to highlights since they suffer more severe information loss due to dynamic range clipping and tone mapping as in Fig. 3.

2.2. Computational cost

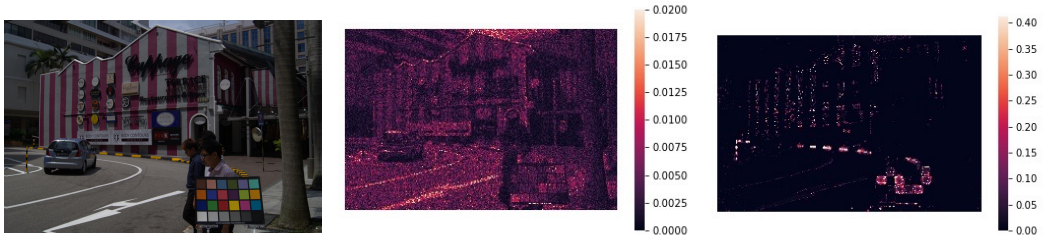
Our method can be trained and evaluated on a single RTX A5000 GPU. Specifically, we evaluate the computational cost of our proposed method and other SOTA methods. The results are shown in Table 1 measured by a commonly used library *thop* [1]. As we can see in the table, with the help of our proposed sRGB-guided context model, the context model becomes feasible in the raw image reconstruction task where the feature resolution is much higher. Compared with other learning based models, our method is lightweight with fewer parameters. In addition, we achieve comparable speed with other DNN-based networks, *e.g.*, InvISP [9] and SAM [6], and faster speed than test-time model SAM [6]. Besides, the speed bottleneck

[1] https://github.com/Lyken17/pytorch-OpCounter



(a) Samsung NX2000 (b) Olympus E-PL6 (c) Sony SLT-A57

Figure 1. The reconstruction qualities of models under different bpp. We evaluate the performance of models using compressed JPEG images with a quality factor of 10 from the NUS dataset [2]. The file size of JPEG images is also taken into consideration for the calculation of the bpp, *e.g.*, the bpp of InvISP is equal to the bpp of JPEG images since InvISP does not save additional metadata.



(a) JPEG image (bpp=1.02) (b) Error map (c) Bits allocation map (bpp=0.011)

(d) JPEG image after contrast enhancement (e) Zoom in of the patch from JPEG (f) The patch from reconstructed Raw (g) The patch from the lossless Raw

Figure 2. The bits allocation and visualization of the shadow areas. All the second-row images have undergone contrast enhancement.

of our proposed method becomes the arithmetic coding instead of the previous context model [3, 4]. Our proposed method can be further accelerated by running the arithmetic coding on GPUs [1, 7, 8], which is out of the scope of the paper.

(a) JPEG image (bpp=1.35) (b) Bits allocation map (bpp=0.0128) (c) Error map (d) Reconstructed raw image (PSNR 46.18 SSIM: 0.9950) (e) Reference raw image

Figure 3. The results for highlights (red bounding box area).

	Parameters	FLOPs	Compress & Decompress time
InvISP [9]	1.41MB	18.06T	5.13s
SAM [6] (bpp 9.56e-4)	N/A	–	16.40s
SAM [6] (bpp 9.52e-3)	N/A	–	103.92s
Nam *et al.* [5]	2.59MB	5.23T	1.50s
Ours (serial)	0.53MB	–	$>$ 10 hours
Ours	0.55MB	5.82T	1.60+6.72 (arithmetic coding on CPU)=8.32s

Table 1. The comparison of the computational cost of different methods, given a 2920×4386 input. *N/A* in SAM [6] means it is a parameter free method. *Ours (serial)* means we substitute the commonly used PixelCNN based context model [3, 4] for our proposed sRGB-guided context model. The arithmetic coding in our method can be further accelerated [1, 7, 8] by running it on GPU.

2.3. Reconstruction qualities with different bpp

We further explore the performance of the proposed model when we increase the bpp, *i.e.,* by increasing the weight of the reconstruction term. We re-train a model with $\lambda = 0.5$ instead of 0.05 in the main paper. The results evaluated on JPEG images with the quality factor of 10 are reported in Fig. 1. As shown in the figure, the performance of our method is greatly improved (around 6 in terms of PSNR) with the increase of the bpp while still remaining lower than Nam *et al.* [5]. Besides, we can significantly improve the reconstruction quality even if the file size of additionally saved metadata is almost negligible compared with the size of JPEG images. The visualizations of error maps are shown in Fig. 5 and Fig. 4. Some visualization results in the image space can be seen in Fig. 6. As we can see, our method achieves much better reconstruction quality with lower bpp than all other SOTA methods. In addition, due to the low quality of conditioned JPEG images, SAM [6] fails in some of the blocks due to the non-exitance of solutions for the systems of linear equations.

	Fidelity		**bpp**		
	PSNR	SSIM	Deterministic mask	Random sparse mask	The proposed learnable mask
Samsung	37.814	0.96757	3.102e-01	2.928e-01	**2.854e-01**
Olympus	39.517	0.97745	2.967e-01	2.764e-01	**2.694e-01**
Sony	39.936	0.97972	2.835e-01	2.691e-01	**2.606e-01**
Mean	39.089	0.97491	2.968e-01	2.794e-01	**2.718e-01**

Table 2. The ablation study on the proposed learnable order prediction model. We evaluate the performance of methods conditioned on JPEG images with quality factor of 10 on NUS dataset.

2.4. Ablation study on learnable order masks

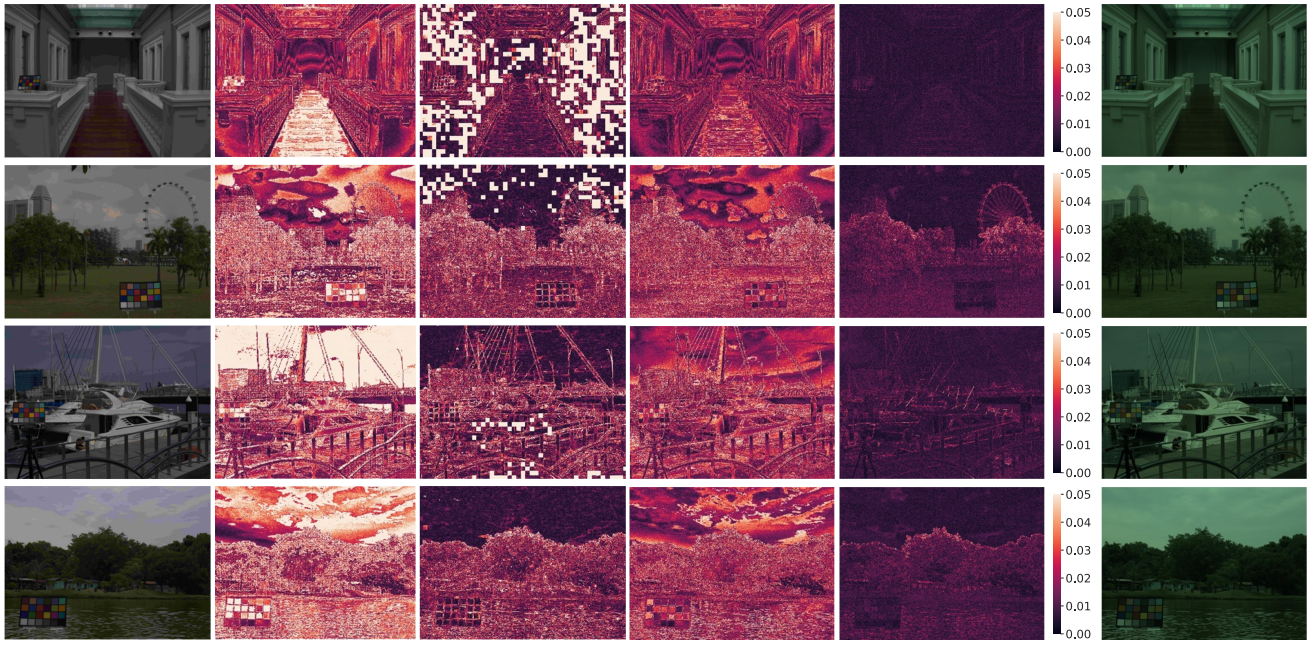
As illustrated in the main paper, our proposed sRGB-guided context model has two parts: the order prediction module and the iterative Gaussian entropy model. In this subsection, we explore the effectiveness of the design of the sampling masks during the compress/decompress processes. We evaluate the performance of the following strategies

- A deterministic mapping formulated by

$$M_{i,j}^{k} = \begin{cases} 1, & \arg\max_c(\mathbf{m}_{i,j}^{c}) = k \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

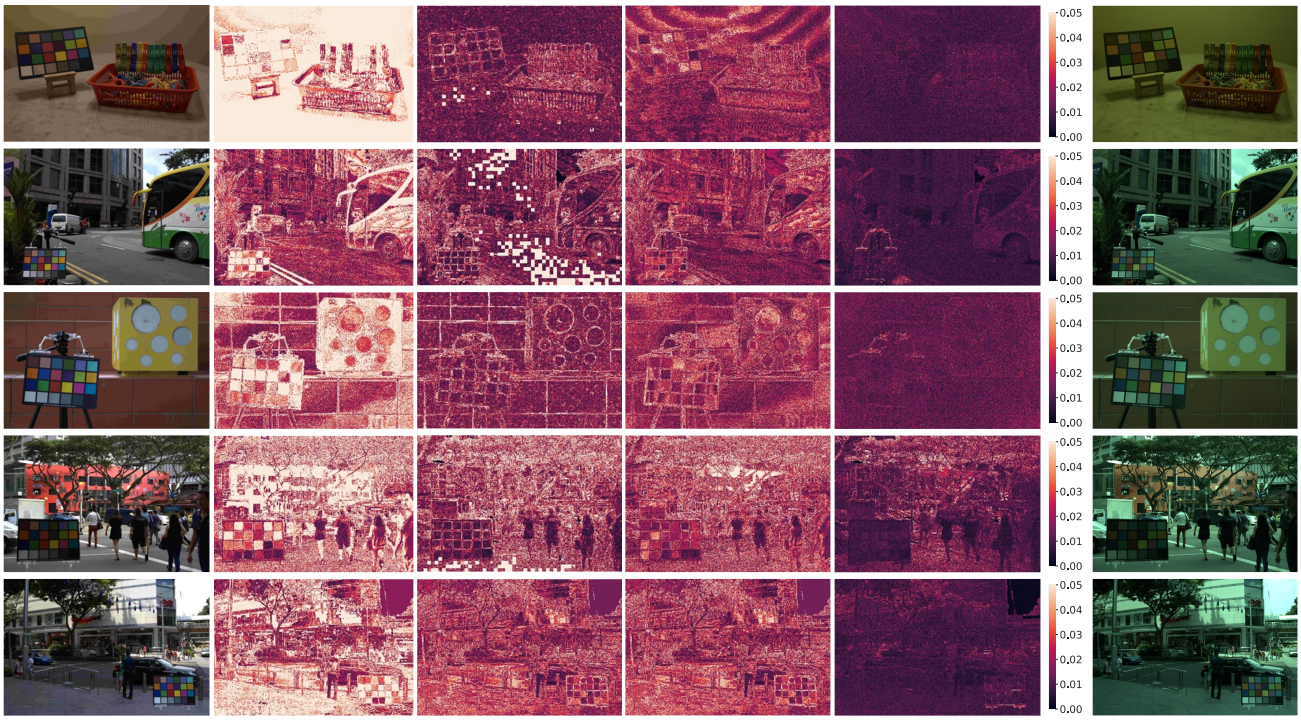
- Random sparse masks $\mathbf{M}^i$ that satisfy $\sum_{i=0}^{N} \mathbf{M}^i = \mathbf{1}$ where $\mathbf{1}$ is an all-one mask.
- The strategy adopted in the main paper. Specifically, we add a buffer to the model to save a pre-sampled random matrix $\mathbf{g}$ so that we can achieve the same sparse sampling masks during the compress/decompress process. Besides, the order masks are obtained with the guidance of sRGB images.

We evaluate the performance of the aforementioned strategies using the model illustrated in Sec. 2.3, *i.e.*, the model trained with $\lambda = 0.5$. The results are shown in Table 2. Since the proposed sRGB-guided context model is a lossless compression model, all the competitors have exactly the same reconstruction quality. In terms of the bpp, the random sparse mask achieves better performance compared with the deterministic mask, due to the sparsity of sampling masks, i.e., the similarity between adjacent pixels can be better utilized. Our proposed learnable mask achieves the best performance which demonstrates the effectiveness of the proposed pre-sampling strategy and the introduced sRGB guidance.



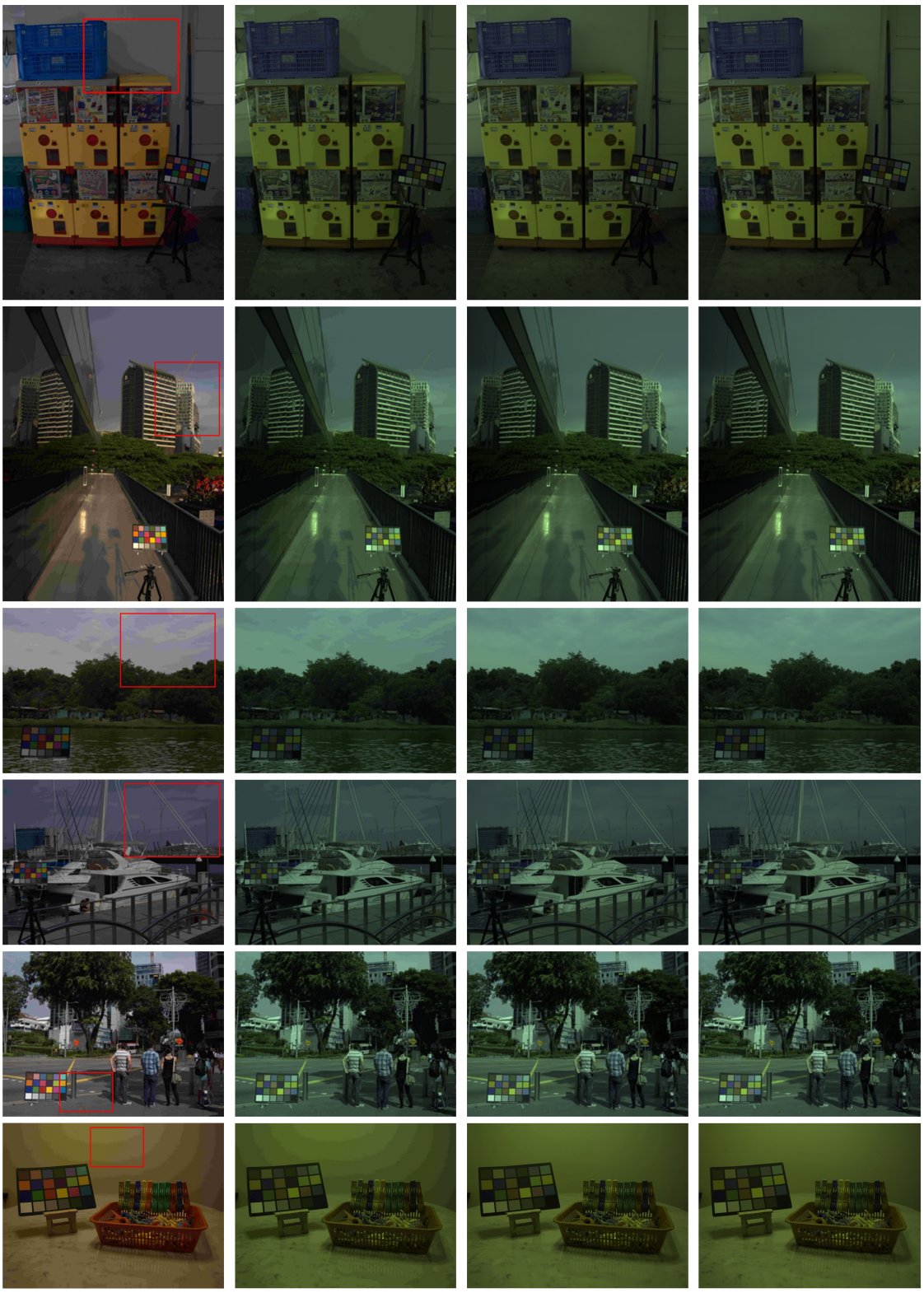
(a) Input (8 bit sRGB image) | (b) InvISP [9] (bpp: N/A) | (c) SAM [6] (bpp: 9.52e-3) | (d) Nam *et al.* [5] (bpp: 8.44e-1) | (e) Ours (bpp: 0.2978) | (f) Raw image (After gamma correction)

Figure 4. Qualitative comparison results on the Olympus subset of NUS dataset [2]. We visualize the maximum value of the error among three channels on the pixel level. For better visualization, we apply gamma correction to the raw image to increase the visibility.



(a) Input (8 bit sRGB image) | (b) InvISP [9] (bpp: N/A) | (c) SAM [6] (bpp: 9.52e-3) | (d) Nam *et al.* [5] (bpp: 8.44e-1) | (e) Ours (bpp: 0.3141) | (f) Raw image (After gamma correction)

Figure 5. Qualitative comparison results on the Samsung subset of NUS dataset [2]. We visualize the maximum value of the error among three channels on the pixel level. For better visualization, we apply gamma correction to the raw image to increase the visibility.



(a) Input 8 bit JPEG images (quality factor 10) | (b) Nam *et al.* [5] (bpp: 0.844) | (c) Ours (bpp: 0.269) | (d) Raw image (After gamma correction)

Figure 6. Qualitative comparison results on NUS dataset [2] in the image space. For better visualization, we apply gamma correction to raw/reconstructed raw images to increase the visibility.

References

[1] Liang Chen, Yong Fang, and Bormin Huang. Accelerating arithmetic coding on a graphic processing unit. In *High-Performance Computing in Remote Sensing*, volume 8183, pages 88–97. SPIE, 2011. 2, 3

[2] Dongliang Cheng, Dilip K Prasad, and Michael S Brown. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *JOSA A*, 31(5):1049–1058, 2014. 2, 4, 5, 6

[3] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7939–7948, 2020. 1, 2, 3

[4] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31, 2018. 2, 3

[5] Seonghyeon Nam, Abhijith Punnappurath, Marcus A Brubaker, and Michael S Brown. Learning srgb-to-raw-rgb de-rendering with content-aware metadata. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17704–17713, 2022. 3, 4, 5, 6

[6] Abhijith Punnappurath and Michael S Brown. Spatially aware metadata for raw reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 218–226, 2021. 1, 3, 4, 5

[7] Jiahan Su. A cuda implementation of arithmetic coding. https://github.com/jiahansu/GPUAR. 2, 3

[8] Jan Šupol and Bořivoj Melichar. Arithmetic coding in parallel. *International Journal of Foundations of Computer Science*, 16(06):1207–1217, 2005. 2, 3

[9] Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6287–6296, 2021. 1, 3, 4, 5