Raw Image Reconstruction with Learned Compact Metadata

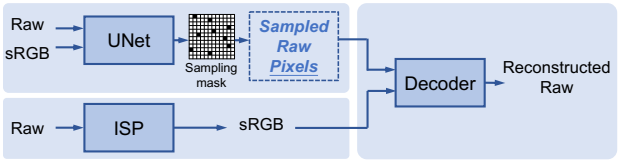
Yufei Wang[1], Yi Yu[1], Wenhan Yang[2], Lanqing Guo[1], Lap-Pui Chau[3], Alex C. Kot[1], Bihan Wen[1*]
[1]Nanyang Technological University [2]Peng Cheng Laboratory
[3]The Hong Kong Polytechnic University
{yufei001, yuyi0010, lanqing001, eackot, bihan.wen}@ntu.edu.sg
yangwh@pcl.ac.cn lap-pui.chau@polyu.edu.hk

Abstract

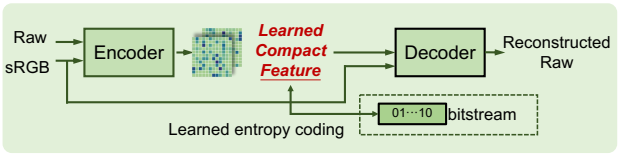
While raw images exhibit advantages over sRGB images (e.g., linearity and fine-grained quantization level), they are not widely used by common users due to the large storage requirements. Very recent works propose to compress raw images by designing the sampling masks in the raw image pixel space, leading to suboptimal image representations and redundant metadata. In this paper, we propose a novel framework to learn a compact representation in the latent space serving as the metadata in an end-to-end manner. Furthermore, we propose a novel sRGB-guided context model with the improved entropy estimation strategies, which leads to better reconstruction quality, smaller size of metadata, and faster speed. We illustrate how the proposed raw image compression scheme can adaptively allocate more bits to image regions that are important from a global perspective. The experimental results show that the proposed method can achieve superior raw image reconstruction results using a smaller size of the metadata on both uncompressed sRGB images and JPEG images. The code will be released at https://github.com/wyf0912/R2LCM.

1. Introduction

As an unprocessed and uncompressed data format directly obtained from camera sensors, raw images has unique advantages for computer vision tasks in practice. For example, it is easier to model the distribution of real image noise in raw space, which enables generalized deep real denoising networks [1, 40]; As pixel values in raw images have a linear relationship with scene radiance, they own benefits to recover shadows and highlights without bringing in the grainy noise usually associated with high ISO [12, 16, 33, 35], which greatly contributes to the low-light image enhancement. Besides, with richer colors, raw images offer more room for correction and artistic manipulation.



(a) Previous SOTA methods sample in the raw space [26, 30].

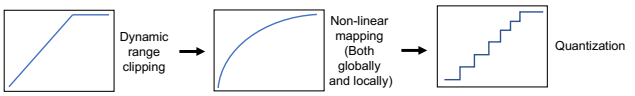


(b) The proposed method samples in the latent space.

Figure 1. The comparison between the previous SOTA methods (in blue) and our proposed method (in green). Different from the previous work where the sampling strategy is hand-crafted or learned by a pre-defined sampling loss, we learn the sampling and reconstruction process in a unified end-to-end manner. In addition, the sampling of previous works is in the raw pixel space, which in fact still includes a large amount of spatial redundancy and precision redundancy. Instead, we conduct sampling in the feature space, and more compact metadata is obtained for pixels in the feature space via the adaptive allocation. The saved metadata is annotated in dashed box.

Despite of these merits, raw images are not widely adopted by common users due to large file sizes. In addition, since raw images are unprocessed, additional post processing steps, *e.g.*, demosaicing and denoising, are always needed before displaying them. For fast image rendering in practice, a copy of JPEG image is usually saved along with its raw data [2]. To improve the storage efficiency, raw-image reconstruction problem attracts more and more attention, *i.e.*, how to minimize the amount of metadata required for de-rendering sRGB images back to raw space. Classic metadata-based raw image reconstruction methods model the workflow of image signal processing (ISP) pipeline and save the required parameters in ISP as metadata [27]. To further reduce the storage and computational complexity towards a lightweight and flexible reverse ISP reconstruction,

*Corresponding author.



(a) Simplified ISP adopted from [21]

(b) Processed Raw (c) Quantized sRGB (d) Quantization error map

Figure 2. An illustration of the information loss caused by the ISP. (a) A simplified ISP suffers from the information loss caused by nonlinear transformations. (b) Raw image after process to better display the details. (c) Quantized sRGB image after ISP which suffers information loss, *e.g.*, the red bounding box area. (d) The quantization error map. As we can see from the above figures, the information loss caused by the quantization is non-uniformly distributed in both over-exposed areas and normally-exposed areas.

very recent methods focus on sparse sampling of raw image pixels [26, 30]. Specifically, in [30], a uniform sampling strategy is proposed to combine with an interpolation algorithm that solves systems of linear equations. The work in [26] proposes a sampling network and approximates the reconstruction process by deep learning to further improve the sampling strategy.

Though lots of progress has been made, existing sparse sampling based raw image reconstruction methods still face limitations in terms of coding efficiency and image reconstruction quality. Specifically, the bit allocation should be adaptive and globally optimized for the image contents, given the non-linear transformation and quantization steps in ISP as shown in Fig. 2. For example, the smooth regions of an image can be well reconstructed with much sparser samples, comparing to the texture-rich regions which deserve denser sampling. In constrast, in existing practices, even for the state-of-the-art method [26] where the sampling is enforced to be locally non-uniform, it is still almost uniform from the global perspective, which causes metadata redundancy and limits the reconstruction performance. In addition, very recent works [26, 30] sample in a fixed sampling space, *i.e.*, raw image space, with a fixed bit depth of sampled pixels, leading to limited representation ability and precision redundancy.

To address the above issues, instead of adopting a predefined sampling strategy or sampling loss, *e.g.*, super-pixel loss [37], we propose a novel end-to-end learned raw image reconstruction framework based on encoded latent features. Specifically, the *latent features* are obtained by minimizing the reconstruction loss and its bitstream cost simultaneously. To further improve the rate-distortion performance, we propose an sRGB-guided context model based on a learnable order prediction network. Different from the commonly used auto-regressive models [9, 24] which encode/decode the latent features pixel-by-pixel in a sequential way, the proposed sRGB-guided context requires much fewer steps (reduce by more than 10^6-fold) with the aid of a learned order mask, which makes the computational cost feasible while maintaining comparable performance. Fig. 1 compares the proposed raw image reconstruction method with the previous strategies [9, 24].

Our contributions are summarized as follows,

1. We propose the first end-to-end deep encoding framework for raw image reconstruction, by fully optimizing the use of stored metadata.
2. A novel sRGB-guided context model is proposed by introducing two improved entropy estimation strategies, which leads to better reconstruction quality, smaller size of metadata, and faster speed.
3. We evaluate our method over popular raw image datasets. The experimental results demonstrate that we can achieve better reconstruction quality with less metadata required comparing with SOTA methods.

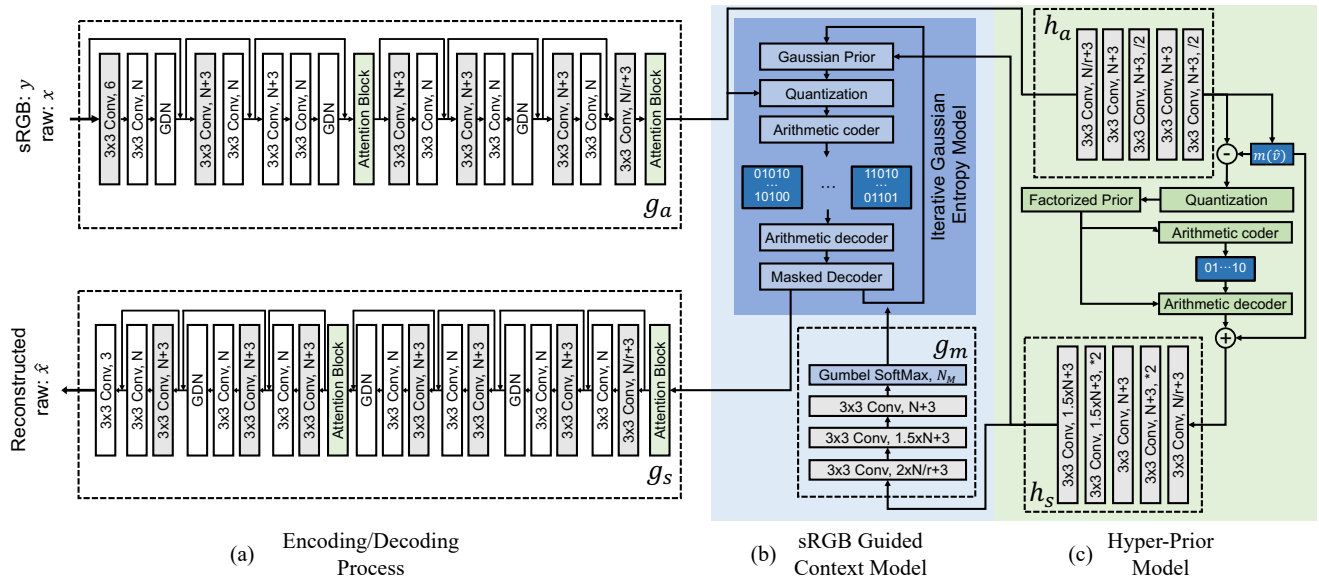
2. Related Work

2.1. Raw image reconstruction

The current raw image reconstruction works can be categorized into two categories: blind raw reconstruction and raw reconstruction with metadata.

Blind raw reconstruction. Blind raw reconstruction aims to reconstruct the raw image only based on the rendered sRGB image [32, 41]. Early works aim to recover the linearity of the image by radiometric calibration [10]. More complex models [7, 11, 18] are subsequently proposed to better describe the workflows of the ISP pipeline. With the development of the deep-learning, deep-learning based models are rapidly developing. For example, [22] directly learns a mapping from LDR (low dynamic range) to HDR (high dynamic range). [21] uses three specialized CNNs to reserve the proposed subdivided pipeline from HDR to LDR. Recently, [36] proposes to use an invertible network to learn the mapping between sRGB space and raw space and vice versa. Though great progress has been done, due to the information loss during the ISP pipeline, *e.g.*, quantization, the fidelity of the reconstructed ones is inevitably constrained.

Raw reconstruction with metadata. To further improve the fidelity of the raw image reconstruction, an alternative way is to save additional metadata to assist the reconstruction [26, 29, 39]. For instance, the work in [39] proposes to save a low resolution raw file to model the tone mapping curve. The works in [27, 28] propose to save the estimated parameters of the simplified ISP pipeline. The work in [30] proposes a spatial aware algorithm that estimates the parameters of interpolation during the test time based on saved uniformly sampled raw image pixels. A recent



(a) Encoding/Decoding Process (b) sRGB Guided Context Model (c) Hyper-Prior Model

Figure 3. The overall framework of our method includes three parts: (a) The encoding/decoding process maps the raw image into the latent space and vice versa. (b) The sRGB-guided context model, which is based on the iterative Gaussian entropy model, encodes/decodes the latent variable $\hat{\mathbf{z}}$ into/from bitstream. (c) The hyper-prior model encodes/decodes the auxiliary variable $\hat{\mathbf{v}}$ into/from the bitstream based on the saved channel-mean value $m(\hat{\mathbf{v}})$. The blue blocks with white text represent the saved metadata, and layers in gray represent that the sRGB image is additionally concatenated to their inputs.

work [26] improves the sampling strategy by sampling representative raw pixels based on the superpixel. Besides, a UNet is adopted [26] to further speed-up the inference process. However, the training of the sampling network is based a pre-defined loss which leads to suboptimal sampling strategy and affects the restoration performance. Different from previous works that usually save discrete pixels of raw images, we propose an end-to-end network that can learn to extract necessary metadata in the latent space.

2.2. Learned image compression

Recently, a great number of deep learning based image compression methods [4, 15, 20, 38] have been proposed and achieve promising results. End-to-end training is made possible thanks to the development of differential quantization and rate estimation [3, 4, 31]. Besides, the introduction of the contextual model [19, 24] greatly improves the compression rate of learned compression models and attracts more and more attention recently. Specifically, the works in [19, 24] propose to utilize an autoregressive model to utilize the information that already decompressed from the bitstream. However, due to the nature of the context model, both compress and decompress processes are extremely slow for the image with high resolution. To minimize the serial processing, [25] proposes a channel-conditioning and [13] proposes a checkerboard context model. Besides, though the learning-based image compression exhibits very promising results on the low bpp (bit per pixel) scenarios, the network architecture needs to be carefully designed for the settings that require high fidelity as shown in [14, 23].

3. Methodology

3.1. Motivation

Our goal is to reconstruct the raw image $\mathbf{x}$ which has a linear relationship with the scene radiance based on the sRGB image $\mathbf{y}$ after the ISP pipeline. Due to the operations like quantization and tone mapping, the process from the raw image to sRGB image is non-linear and the information loss is spatially non-uniform as shown in Fig. 2. Different from the previous works that uniformly/approximately uniformly save the sparse *raw-pixel* values with a *fixed* number of bits, we propose to learn the coding of information in the *latent space* with an adaptively allocated number of bits for each pixel in an end-to-end manner.

As shown in Fig. 3, our method aims to obtain a compact representation $\hat{\mathbf{z}}$ of the image conditioned on the corresponding sRGB image. The latent feature $\hat{\mathbf{z}}$ is expected to have necessary information to reconstruct the raw image with high fidelity and its code-length shall be as small as possible. To this end, we propose an sRGB-guided context model which can make better use of decoded information and greatly improve computational efficiency. Besides, an improved hyper-prior is proposed to further improve the coding efficiency and reconstruction quality.

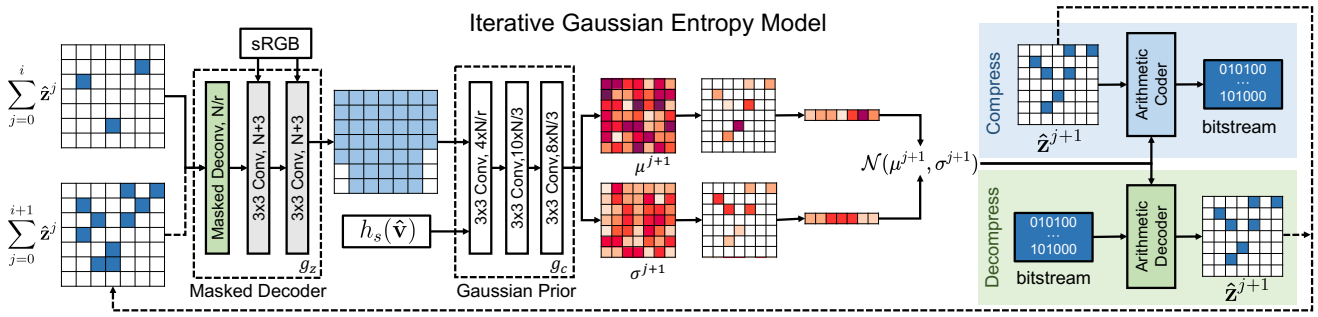


Figure 4. An illustration of a step of the proposed iterative Gaussian entropy model. We model the distribution of $\hat{\mathbf{z}}^{i+1}$ based on the existing information $\sum_{j=0}^{i} \hat{\mathbf{z}}^j$ and $h_s(\hat{\mathbf{v}})$ where $\hat{\mathbf{v}}$ is the auxiliary variable from the hyper-prior. Then arithmetic coding is used to compress/decompress the latent code $\hat{\mathbf{z}}^{i+1}$ losslessly. The dashed arrows represent the following step.

3.2. The overall of entropy-based coding

Specifically, our framework can be formulated by

$$\begin{aligned} \mathbf{z} &= g_a(\mathbf{x}, \mathbf{y}; \phi) \\ \hat{\mathbf{z}} &= Q(\mathbf{z}) \\ \hat{\mathbf{x}} &= g_s(\hat{\mathbf{z}}, \mathbf{y}; \theta), \end{aligned} \tag{1}$$

where $\mathbf{z}$ and $\hat{\mathbf{z}}$ are latent codes w/o and w/ quantization. g_a and g_s are the analysis and synthesis transforms. ϕ and θ represent the parameters of these two transforms respectively. Q is the quantization operation. We further introduce hypeprior [5] to model spatial dependencies in $\mathbf{z}$ as follows

$$\begin{aligned} \mathbf{v} &= h_a(\mathbf{z}, \mathbf{y}; \phi_h) \\ \hat{\mathbf{v}} &= Q(\mathbf{v}) \\ q_{\hat{\mathbf{z}}|\hat{\mathbf{v}},\mathbf{y}}(\hat{\mathbf{z}}|\hat{\mathbf{v}}, \mathbf{y}) &\leftarrow h_s(\hat{\mathbf{v}}, \mathbf{y}; \theta_h), \end{aligned} \tag{2}$$

where h_a and h_s represent the auxiliary analysis and synthesis transforms respectively, and ϕ_h and θ_h are the learned parameters of them. The optimization objective that simultaneously minimizes the raw image reconstruction loss and the codelength of latent codes is defined as follows

$$\begin{aligned} \mathcal{L} &= \underbrace{\mathcal{R}(\hat{\mathbf{z}}) + \mathcal{R}(\hat{\mathbf{v}})}_{\text{rate}} + \lambda \cdot \underbrace{\mathcal{D}(\hat{\mathbf{x}}, \mathbf{x})}_{\text{distortion}} \\ &= \mathbb{E}[-\log_2 q_{\hat{\mathbf{z}}|\hat{\mathbf{v}},\mathbf{y}}(\hat{\mathbf{z}}|\hat{\mathbf{v}}, \mathbf{y})] \\ &+ \mathbb{E}[-\log_2 q_{\hat{\mathbf{v}}|m(\mathbf{v})}(\hat{\mathbf{v}}|m(\mathbf{v}))] + \lambda \cdot \mathcal{D}(\hat{\mathbf{x}}, \mathbf{x}), \end{aligned} \tag{3}$$

where $m(\mathbf{v})$ represents the mean value of different channels, and $\mathcal{D}$ is the mean square error to measure the reconstruction loss. The details of the likelihood estimations of different latent codes will be introduced below.

3.3. The estimation of the likelihood

As revealed by the cross entropy $H(p, q) = H(p) + \mathcal{D}_{KL}(p||q)$ that measures the number of extra bits to code the desired distribution p using an estimated one q, the key of reducing the code length is to accurately model the distribution of latent codes. To this end, we propose to model the distribution of different latent variables with different strategies since they depend on different information.

Following a similar way of previous works [5,9], we use a non-parametric, fully factorized density model to encode the auxiliary latent codes $\mathbf{v}$. However, due to the limitation of the network design, *e.g.*, the domain of the hyper-prior model must be univariate and the network must be monotonic increasing [5], we find that there is still lots of redundant information in the auxiliary latent codes v as shown in Fig. 5. To further improve the coding efficiency, we propose to additionally save the mean value of each channel $m(\mathbf{v})$ to the metadata to reduce the redundancy. Specifically, we model the conditional distribution $q_{\hat{\mathbf{v}}|m(\mathbf{v})}$ as follows

$$\begin{aligned} &q_{\hat{\mathbf{v}}|m(\mathbf{v})}(\hat{\mathbf{v}}|m(\mathbf{v})) = \\ &\prod_i [(q_{v_i|m(\mathbf{v})}(\psi^{(i)}) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2}))(\hat{\mathbf{v}} - m(\mathbf{v}))_i], \end{aligned} \tag{4}$$

where ψ^i is the parameters of each univariate conditional distribution $p_{v_i|m(\mathbf{v})}$, and i is the position index.

For the encoding of $\hat{\mathbf{z}}$, previous works show great improvement by introducing the context model, *i.e.*, the already decompressed pixels can help to predict the pixels which are not decompressed yet to further improve the coding efficiency. However, due to its serialization property, the autoregressive model incurs a significant computational cost which is unacceptable in the raw image reconstruction since its high-resolution, *e.g.*, 4000×6000. To improve the computational efficiency while preserve the advantages of the autoregressive model, we propose a novel sRGB-guided context model. More specifically, our proposed context model includes two parts: a learnable order prediction network g_m as shown in Fig. 3, and an iterative Gaussian entropy model as shown in Fig. 4.

The learnable order prediction network. As the prerequisite of our proposed context model, the order masks of com-

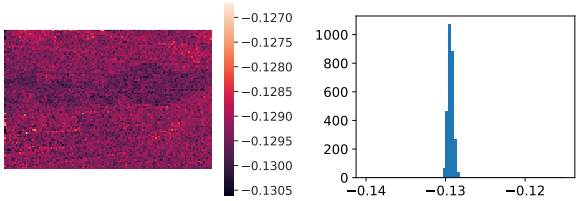


Figure 5. Visualization of a channel of the auxiliary latent features $\mathbf{v}$ and its value distribution.

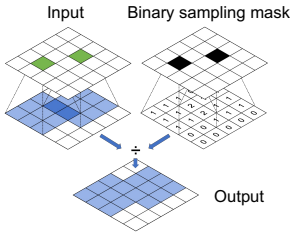


Figure 6. The proposed masked deconvolution layer.

pression/decompression play a significant role. To make sampling order masks learnable, we propose a training strategy that makes end-to-end training of the whole framework feasible. Specifically, we utilize Gumbel-softmax [17] to make the binary mask derivable for training as follows

$$M^k_{i,j} = \frac{\exp((\log(m^k_{i,j}) + g^k_{i,j})/\tau)}{\sum_{t=1}^{N} \exp((\log(m^t_{i,j}) + g^t_{i,j})/\tau)}, \tag{5}$$

where k is the index of sampling masks, N is the number of iterations, $\mathbf{g} \in \mathbb{R}^{N\times h\times w}$ is a random matrix i.i.d sampled from Gumbel(0,1) distribution, τ is a temperature hyper-parameter, and $\mathbf{m} \in \mathbb{R}^{N\times h\times w}$ denotes unnormalized log probabilities predicted by a subnetwork. For inference, to make sure that we have exactly the same random vector $\mathbf{g}$ during compress/decompress processes, we add a registered buffer to the model to save a pre-sampled $\mathbf{g}$. The pre-sampled $\mathbf{g}$ is then cropped to the same size as the $\mathbf{m}$ to generate a set of sparsely sampled $\mathbf{M}^k$.

In addition, we find that the vanilla convolutional layer cannot well utilize the information from the randomly sparsely sampled features (can refer to the sampling mask in Fig. 9). Therefore, we further propose a new masked deconvolution layer that can alleviate negative impacts from the randomness and sparsity as shown in Fig. 6. For an input feature $\hat{\mathbf{z}}$ and its corresponding mask $\mathbf{M}^c = \sum_{i=0}^{k} \mathbf{M}^k$ which records the positions of all already decoded ones, the output $\hat{\mathbf{z}}'$ is as follows

$$\hat{\mathbf{z}}' = \frac{\text{Deconv}(\mathbf{z})}{\max(1, \text{Deconv}_{\mathbf{1}}(\mathbf{M}^c))}, \tag{6}$$

where Deconv and $\text{Conv}_{\mathbf{1}}$ are deconvolution layers with stride of 1. Besides, $\text{Deconv}_{\mathbf{1}}$ is a fixed layer that the weights are all one and the bias is zero.

The iterative Gaussian entropy model. After obtaining the predicted order mask, we can iteratively compress/decompress the information as shown in Fig 4. Specifically, we use information from the auxiliary latent variable $\hat{\mathbf{v}}$ and already encoded/decoded partial of $\hat{\mathbf{z}}$ to predict the distribution of the to-be-processed part of $\hat{\mathbf{z}}$ as follows

$$\boldsymbol{\mu}^{i+1}, \boldsymbol{\sigma}^{i+1} = g_c(g_z((\sum_{k=0}^{i} \mathbf{M}^i) \odot \hat{\mathbf{z}}, y), h_s(\hat{\mathbf{v}})), \tag{7}$$

where $\odot$ is a pixel-wise multiplication, g_z is the masked decoder, and g_c is the Gaussian prior module to predict the distribution of $\hat{\mathbf{z}}^{i+1}$ that are not encoded. Then, the likelihood of $\hat{\mathbf{z}}$ is formulated as

$$\begin{aligned} q_{\hat{\mathbf{z}}|\hat{\mathbf{v}},\mathbf{y}}(\hat{z}_i|\hat{\mathbf{v}},\mathbf{y}) &= (\mathcal{N}(\mu_i^{k_i}, \sigma_i^{k_i}) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2}))(\hat{z}_i) \\ &= c_{\mu_i^{k_i}, \sigma_i^{k_i}}(\hat{z}_i + 0.5) - c_{\mu_i^{k_i}, \sigma_i^{k_i}}(\hat{z}_i - 0.5), \end{aligned} \tag{8}$$

where the subscript i is the index of the pixel position, k_i is the index of parameter groups defined in Eq. 7, and $c_{\mu_i^{k_i}, \sigma_i^{k_i}}(\cdot)$ is its corresponding cumulative function.

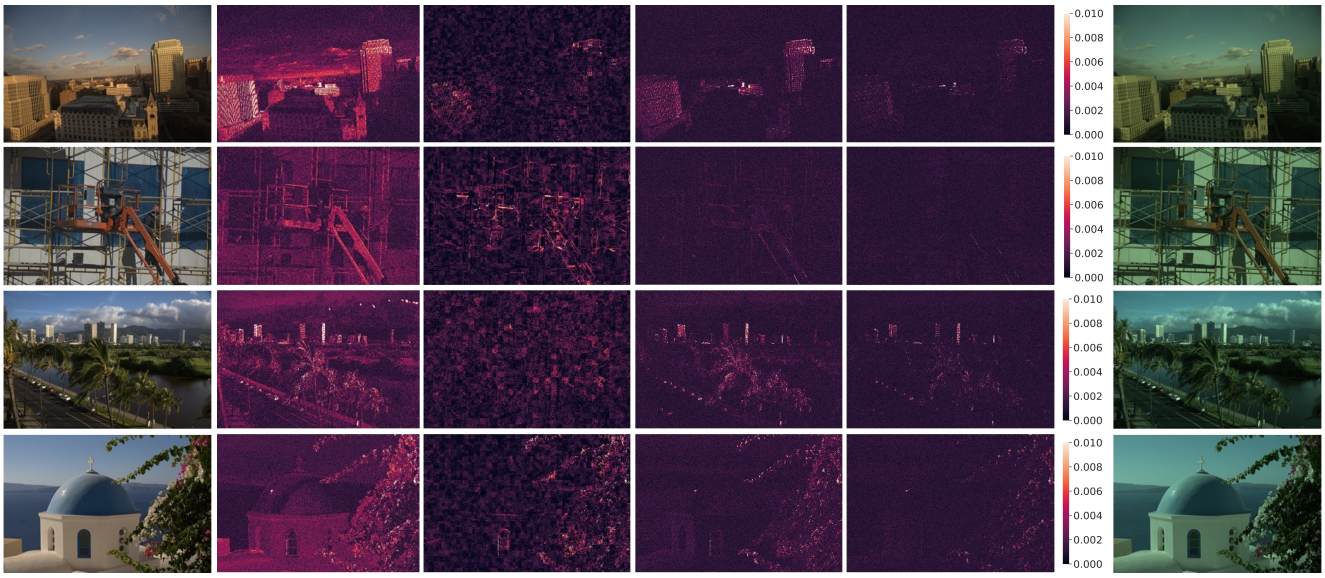
4. Experiments

4.1. Experimental settings

Datasets. We utilize two widely-used datasets, NUS dataset [8] and AdobeFiveK dataset [6], to evaluate the effectiveness of our proposed methods. These datasets are all natural images collected from different scenarios and devices. Following previous work [26, 30], we use the raw image after demosaic and render sRGB images using a software ISP. Specifically, AdobeFiveK dataset [6] includes 5000 photographs taken by different photographers and devices so that it covers a wide range of scenes and lighting conditions. We randomly split the whole dataset into training set and validation set which include 4900 and 100 images respectively. For the NUS dataset [8], we select the same subsets of devices with the previous work [26].

Baselines. We compared the proposed method with several SOTA methods, including InvISP [36], SAM [30], and Nam *et al.* [26]. Specifically, InvISP [36] is a SOTA raw image reconstruction model that utilizes a single invertible network to learn the mapping from sRGB image to raw image and vice verse. SAM [30] is a test-time adaptation model that saves the uniformly sampled raw pixels as metadata. Nam *et al.* [26] learn the sampling process and reconstruction process using two separate neural networks [37].

Implementation details. All the code will be released after acceptance. For training, we use a batch size of 1 and patch size of 1024 to reduce the I/O time. Adam is used as the



(a) Input (8 bit sRGB image) | (b) InvISP [36] (bpp: N/A) | (c) SAM [30] (bpp: 9.52e-3) | (d) Nam *et al.* [26] (bpp: 8.44e-1) | (e) Ours (bpp: 4.90e-4) | (f) Raw image (After gamma correction)

Figure 7. Qualitative comparison of the raw image reconstruction results. 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.

Method	bpp	PSNR	SSIM
InvISP [36]	N/A	52.69	0.99938
SAM [30]	9.566e-4	49.61	0.99874
SAM [30]	9.5219-3	54.76	0.99945
Nam *et al.* [26]	8.438e-1	56.72	0.99958
Ours (w/o metadata)	N/A	53.03	0.99926
Ours	**4.901e-4**	**58.14**	**0.99969**

Table 1. Quantitative evaluation on AdobeFiveK dataset.

optimizer with a learning rate of 1e-4. We train the models for 100 epochs for AdobeFiveK dataset and 200 epochs for NUS dataset. We reduce the learning rate by a factor of 0.1 if there is no improvement in terms of the loss after every 20 epochs. For the sRGB-guided context model, we set $N = 2$ for the model conditioned on uncompressed sRGB images and $N = 4$ for the compressed JPEG data.

4.2. Experimental results

For the evaluation metrics, we utilize PSNR and SSIM [34] which are widely used to evaluate the reconstruction quality with the reference image. We also utilize bpp (bit per pixel) to evaluate the coding efficiency of the model.

4.2.1 Results on uncompressed sRGB data.

Results on AdobeFiveK dataset. We report the quantitative evaluation results in Table 1. As we can see in the table, raw image reconstruction models with metadata can achieve better performance than SOTA raw image reconstruction model without metadata [36]. Besides, compared with previous metadata-based SOTA methods [26, 30], our method achieves better reconstruction quality with lower storage overhead. Besides, to exclude the effect of network structure, we retrain and evaluate the performance of the model without the help of metadata using the same architecture as ours. Specifically, we set the original input of the raw image to zero, and remove the quantization step and code length loss. We find that its reconstruction quality is much lower than the results obtained from the same network with meta-data, which demonstrates the effectiveness of the saved metadata. Visual comparisons can be found in Fig. 7.

Results on NUS dataset. We also evaluate the performance of models on NUS dataset following a similar evaluation paradigm with Nam *et al.* [26]. The results are reported in Table 2. As can be seen in the table, we achieve huge performance improvement compared with SOTA method Nam *et al.* [26] (even compared with the test-time optimization version). In addition, the number of bits we need to save as metadata is less than 0.1% of [26].

4.2.2 Results on compressed sRGB data.

We further consider a more challenging and realistic setting that we reconstruct the raw image based on compressed JPEG image. To evaluate the robustness of our method, we train a single model across different devices and JPEG quality factors. The results are reported in Table 3. As we can see, our method can adaptively allocate different bpp to JPEG images with different quality factors, *i.e.*, assigning higher bpp to the image with worse JPEG quality. Our

Method	bpp ↓	Samsung NX2000		Olympus E-PL6		Sony SLT-A57	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
RIR [27]	3.253e-2	45.66	0.9939	48.42	0.9924	51.26	0.9982
SAM [30]	7.500e-1	47.03	0.9962	49.35	0.9978	50.44	0.9982
Nam *et al.* [26] [1]	8.438e-1	48.08	0.9968	50.71	0.9975	50.49	0.9973
[26] w/ fine-tuning	8.438e-1	49.57	0.9975	51.54	0.9980	53.11	0.9985
Ours	**2.887e-4**	**57.84**±0.89	**0.9997**±0.00	**59.08**±0.95	**0.9998**±0.00	**58.76**±0.95	**0.9997**±0.00

[1] The reason that the bpp of SAM [30] and Nam *et al.* [26] is slightly different is that Nam *et al.* [26] need extra bits to save the locations of sampled raw pixels as shown in Fig. 8.

[2] For the experiments in the gray area, we use a five-fold cross validation, and report the mean and std in the table.

Table 2. The quantitative results of NUS dataset [8] conditioned on sRGB images.

method achieves the best reconstruction quality with the least metadata compared with previous SOTA methods.

4.3. Ablation study

The comparison of bits allocation. One of the main advantages of the proposed method is that we could learn the bits allocation in an end-to-end manner. As shown in Fig. 2, the information loss is non-uniform so a good bits allocation algorithm is the core of raw image reconstruction algorithms. To this end, we visualize the bits allocation of both current SOTA methods and the proposed method in Fig. 8. In [30], the metadata are uniformly sampled in the raw pixel space, which leads to redundancy. Although Nam *et al.* [26] propose a superpixel based sampling network, the training of reconstruction and sampling networks are separated into two phases. In addition, even if its sampling is locally non-uniform, it still remains uniform globally, which limits the coding efficiency and reconstruction quality. As we can see in the figure, our method can adaptively allocate different bits to different areas. Specifically, for the flat area, *e.g.* flat area in the blue bounding box, our methods utilize few bits to encode. While for the area with more complicated context, *e.g.* boundary area in the blue bounding box, our method allocates relatively more bits. Besides, even for areas where we allocate bits, the need of bits is much lower than methods that sample in the raw pixel space.

The hand-crafted metadata of $\mathbf{v}$. To verify the effectiveness of our proposed modeling of the hyper-prior variable $\hat{\mathbf{v}}$. We compare the models trained w/ and w/o the hand-crafted metadata $m(\mathbf{v})$. We keep other settings the same as in Table 3 and evaluate the models on a fold of NUS dataset. The results are reported in Table 4. As we can see in the table, there are improvements in terms of both the bpp and the reconstruction quality that benefited from the more accurate value of $\mathbf{v}$ and alleviated redundancy.

The sRGB-guided context model. To better understand how the proposed context model works, we visualize each step of encoding/decoding $\mathbf{z}$ in Fig. 9 (a). As we can see, the order of the compress/decompress process is highly-related to the context of the image, which demonstrates that our proposed context model can well utilize the information

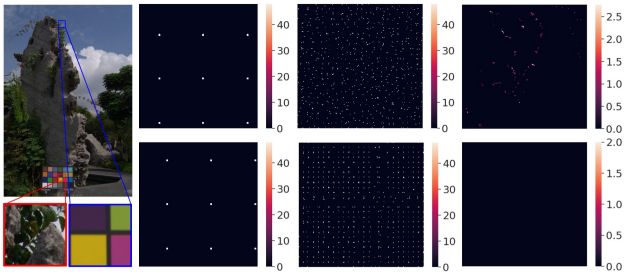


Figure 8. The comparison of the bits allocation. (a) The input 8-bit sRGB image. (b)-(d) The bits allocation maps of the red bounding box area (the first row) and the blue area (the second row). For better visualization, we enlarge the size of sampled pixels in (b). It is worth noting that for (b)-(c), each sampled raw pixel needs 48 bits to save. *Best zoom in for more details.*

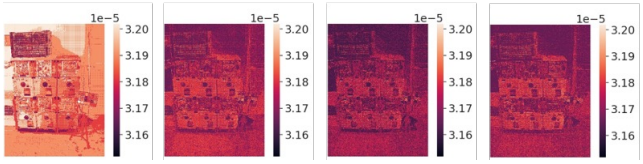


Figure 9. The visualization of each step of the sRGB-guided context model (a) The sampling masks from $\mathbf{M}^0$ to $\mathbf{M}^3$ respectively. The sampling rate and the bytes of the encoded string are displayed on the right side of the mask. (b) The bpp maps of the latent variable z based on the already decoded information. Specifically, the i_{th} (range from 0 to 3) bpp map is estimated using the information of $\sum_{k=-1}^{i-1} \mathbf{M}^k \odot \hat{\mathbf{z}}$, where $\mathbf{M}^{-1}$ is an all zero mask.

from the sRGB image. The sparse sampling mask can help a model better predict the distribution of the adjacent pixels that have not been compressed/decompressed. In addition, our method gradually increases the number of sampled pixels in the latent space as more and more pixels are available to help to predict the distribution of the unseen ones, which leads to better coding efficiency. As shown in Fig. 9 (b), we can get more accurate estimation of the likelihood (*i.e.*, smaller bpp) of $\mathbf{z}$ with the help of already decoded $\hat{\mathbf{z}}$. We also quantitatively evaluate the effectiveness of our proposed sRGB-guided context model as shown in Fig. 10. We

Quality	Method	bpp ↓	Samsung NX2000		Olympus E-PL6		Sony SLT-A57	
			PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
10	InvISP	N/A	26.62	0.8836	29.12	0.8980	29.12	0.9002
	SAM	9.556e-4	24.42	0.8946	25.24	0.9094	25.56	0.9110
	SAM	9.522e-3	27.94	0.9234	28.22	0.9376	27.83	0.9374
	Nam *et al.*	8.438e-1	33.06	0.9373	34.03	0.9477	34.29	0.9506
	Ours	**7.736e-4**	**33.13**	**0.9386**	**34.04**	**0.9482**	**34.31**	**0.9515**
30	InvISP	N/A	28.71	9.9316	31.76	0.9421	30.89	0.9459
	SAM	9.556e-4	28.88	0.9344	30.21	0.9465	29.65	0.9458
	SAM	9.522e-3	34.24	0.9553	35.87	0.9648	36.12	0.9677
	Nam *et al.*	8.438e-1	37.21	0.9630	38.70	0.9723	39.06	0.9750
	Ours	**3.613e-4**	**37.40**	**0.9640**	**38.81**	**0.9729**	**39.18**	**0.9757**
50	InvISP	N/A	30.02	0.9416	32.91	0.9529	32.97	0.9579
	SAM	9.556e-4	30.78	0.9448	32.41	0.9559	32.05	0.9567
	SAM	9.522e-3	36.32	0.9629	37.77	0.9705	38.24	0.9739
	Nam *et al.*	8.438e-1	38.34	0.9686	40.07	0.9767	40.04	0.9797
	Ours	**3.368e-4**	**38.67**	**0.9699**	**40.33**	**0.9776**	**40.73**	**0.9806**
70	InvISP	N/A	30.86	0.9458	32.91	0.9553	32.97	0.9592
	SAM	9.556e-4	32.08	0.9529	34.14	0.9620	33.90	0.9637
	SAM	9.522e-3	37.42	0.9684	38.96	0.9745	39.38	0.9780
	Nam *et al.*	8.438e-1	39.13	0.9724	41.01	0.9769	41.42	0.9825
	Ours	**3.210e-4**	**39.59**	**0.9742**	**41.36**	**0.9807**	**41.75**	**0.9836**
90	InvISP	N/A	31.55	0.9476	33.74	0.9598	33.68	0.9643
	SAM	9.556e-4	34.37	0.9663	36.60	0.9712	36.78	0.9747
	SAM	9.522e-3	39.17	0.9787	40.79	0.9812	41.20	0.9843
	Nam *et al.*	8.438e-1	40.32	0.9782	42.33	0.9838	42.82	0.9864
	Ours	**2.944e-4**	**41.19**	**0.9821**	**42.98**	**0.9856**	**43.43**	**0.9882**

Table 3. The quantitative results of NUS dataset [8] conditioned on the compressed JPEG image with different quality factors.

	Ours w/o $m(\mathbf{v})$			Ours		
	bpp ↓	PSNR	SSIM	bpp ↓	PSNR	SSIM
Samsung	2.92e-4	57.50	0.9997	2.88e-4	57.79	0.9997
Olympus	2.93e-4	58.93	0.9997	2.90e-4	59.35	0.9997
Sony	2.90e-4	59.05	0.9997	2.89e-4	59.24	0.9997
Mean	2.92e-4	58.66	0.9996	**2.89e-4**	**58.79**	**0.9997**

Table 4. The ablation study on the proposed modeling of $\hat{\mathbf{v}}$.

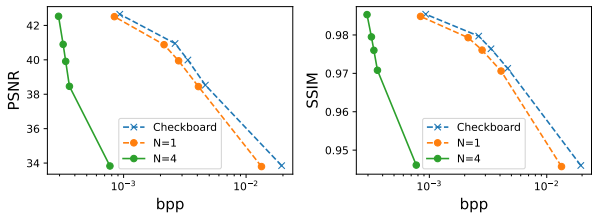


Figure 10. A comparison of models trained with different steps of the proposed sRGB-guided context model and He *et al.* [13]. The models are evaluated using JPEG images with varying quality factors (10, 30, 50, 70, 90). For all models, PSNR and SSIM decrease monotonously, with the lowering of the conditioned JPEG quality.

compare our proposed sRGB-guided context model with He *et al.* [13] which proposes an improved context model to obtain faster speed. For a fair comparison, we directly replace our proposed context model (Fig. 3 (b)) with the checkboard one in He *et al.* [13] and keep other settings the same. As we can see, all models achieve very similar reconstruction quality and our method achieves much lower bpp.

5. Conclusion

In this paper, we present a novel approach to reconstructing raw images using compact metadata. We introduce end-to-end coding techniques that encode the metadata in the latent space using an adaptive bits allocation strategy, resulting in improved reconstruction quality and higher coding efficiency. Additionally, our proposed sRGB-guided context model leads to better reconstruction quality, smaller metadata size, and faster processing speed. We evaluate our method on widely-used datasets and the results demonstrate that our method significantly improves performance over prior methods, *i.e.,* we achieve better reconstruction quality and smaller size of metadata.

Acknowledgement. This work was done at Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University. This research is supported in part by the NTU-PKU Joint Research Institute (a collaboration between the Nanyang Technological University and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation), the Basic and Frontier Research Project of PCL, the Major Key Project of PCL, and the MOE AcRF Tier 1 (RG61/22) and Start-Up Grant.

References

[1] Abdelrahman Abdelhamed, Marcus A Brubaker, and Michael S Brown. Noise flow: Noise modeling with conditional normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3165–3173, 2019. 1

[2] Adobe. Digital negative (dng) specification. https://www.kronometric.org/phot/processing/DNG/dng_spec_1.4.0.0.pdf. 1

[3] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc V Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. *Advances in neural information processing systems*, 30, 2017. 3

[4] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016. 3

[5] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018. 4

[6] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 5

[7] Ayan Chakrabarti, Ying Xiong, Baochen Sun, Trevor Darrell, Daniel Scharstein, Todd Zickler, and Kate Saenko. Modeling radiometric uncertainty for vision with tone-mapped color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2185–2198, 2014. 2

[8] 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. 5, 7, 8

[9] 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. 2, 4

[10] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. 2008. 2

[11] Han Gong, Graham D Finlayson, Maryam M Darrodi, and Robert B Fisher. Rank-based radiometric calibration. In *Color and Imaging Conference*, volume 2018, pages 59–66. Society for Imaging Science and Technology, 2018. 2

[12] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. *arXiv preprint arXiv:2212.04711*, 2022. 1

[13] Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14771–14780, 2021. 3, 8

[14] Leonhard Helminger, Abdelaziz Djelouah, Markus Gross, and Christopher Schroers. Lossy image compression with normalizing flows. *arXiv preprint arXiv:2008.10486*, 2020. 3

[15] Yueyu Hu, Wenhan Yang, Zhan Ma, and Jiaying Liu. Learning end-to-end lossy image compression: A benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[16] Haofeng Huang, Wenhan Yang, Yueyu Hu, Jiaying Liu, and Ling-Yu Duan. Towards low light enhancement with raw images. *IEEE Transactions on Image Processing*, 31:1391–1405, 2022. 1

[17] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 5

[18] Seon Joo Kim, Hai Ting Lin, Zheng Lu, Sabine Süsstrunk, Stephen Lin, and Michael S Brown. A new in-camera imaging model for color computer vision and its application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2289–2302, 2012. 2

[19] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. *arXiv preprint arXiv:1809.10452*, 2018. 3

[20] Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, and David Zhang. Learning convolutional networks for content-weighted image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3214–3223, 2018. 3

[21] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1651–1660, 2020. 2

[22] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, volume 37, pages 37–49. Wiley Online Library, 2018. 2

[23] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33:11913–11924, 2020. 3

[24] 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

[25] David Minnen and Saurabh Singh. Channel-wise autoregressive entropy models for learned image compression. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3339–3343. IEEE, 2020. 3

[26] 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 Pat-*

tern Recognition, pages 17704–17713, 2022. 1, 2, 3, 5, 6, 7

[27] Rang MH Nguyen and Michael S Brown. Raw image reconstruction using a self-contained srgb-jpeg image with only 64 kb overhead. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1655–1663, 2016. 1, 2, 7

[28] Rang MH Nguyen and Michael S Brown. Raw image reconstruction using a self-contained srgb–jpeg image with small memory overhead. *International journal of computer vision*, 126(6):637–650, 2018. 2

[29] Abhijith Punnappurath and Michael S Brown. Learning raw image reconstruction-aware deep image compressors. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):1013–1019, 2019. 2

[30] 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, 2, 5, 6, 7

[31] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, 2017. 3

[32] Lin Wang and Kuk-Jin Yoon. Deep learning for hdr imaging: State-of-the-art and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[33] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2604–2612, 2022. 1

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[35] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2758–2767, 2020. 1

[36] 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. 2, 5, 6

[37] Fengting Yang, Qian Sun, Hailin Jin, and Zihan Zhou. Superpixel segmentation with fully convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13964–13973, 2020. 2, 5

[38] Yi Yu, Yufei Wang, Wenhan Yang, Shijian Lu, Yap-peng Tan, and Alex C Kot. Backdoor attacks against deep image compression via adaptive frequency trigger. *arXiv preprint arXiv:2302.14677*, 2023. 3

[39] Lu Yuan and Jian Sun. High quality image reconstruction from raw and jpeg image pair. In *2011 International Conference on Computer Vision*, pages 2158–2165. IEEE, 2011. 2

[40] Yi Zhang, Hongwei Qin, Xiaogang Wang, and Hongsheng Li. Rethinking noise synthesis and modeling in raw denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4593–4601, 2021. 1

[41] Zhuoran Zheng, Wenqi Ren, Xiaochun Cao, Tao Wang, and Xiuyi Jia. Ultra-high-definition image hdr reconstruction via collaborative bilateral learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4449–4458, 2021. 2