Raw Image Reconstruction with Learned Compact Metadata

Yufei Wang¹, Yi Yu¹, Wenhan Yang², Lanqing Guo¹, Lap-Pui Chau³, Alex C. Kot¹, Bihan Wen¹*
¹Nanyang Technological University  ²Peng Cheng Laboratory  ³The 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 representation 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 have 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.

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,

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.
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 contrast, 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 10⁶-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
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 on 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 $x$ which has a linear relationship with the scene radiance based on the sRGB image $y$ 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 $\hat{z}$ of the image conditioned on the corresponding sRGB image. The latent feature $\hat{z}$ 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.

![Diagram](image-url)

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 $\hat{z}$ into/from bitstream. (c) The hyper-prior model encodes/decodes the auxiliary variable $\hat{v}$ into/from the bitstream based on the saved channel-mean value $m(V)$. 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.
3.2. The overall of entropy-based coding

Specifically, our framework can be formulated by

\[
\begin{align*}
    z &= g_a(x, y; \phi) \\
    \hat{z} &= Q(z) \\
    \hat{x} &= g_s(\hat{z}, y; \theta),
\end{align*}
\]

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

\[
\begin{align*}
    v &= h_a(z, y; \phi_h) \\
    \hat{v} &= Q(v)
\end{align*}
\]

for the encoding of \(z\), 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, \(e.g., 4000 \times 6000\). To improve the computational efficiency while preserving 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-

\[
\begin{align*}
    \mathcal{L} &= \mathcal{R}(\hat{z}) + \mathcal{R}(\hat{v}) + \lambda \cdot \mathcal{D}(\hat{x}, x) \\
    &= E[- \log g_{\hat{x}|\hat{v}, y}(\hat{z}|\hat{v}, y)] \\
    &\quad + E[- \log g_{\hat{v}|m(v)}(\hat{v}|m(v))] + \lambda \cdot \mathcal{D}(\hat{x}, x),
\end{align*}
\]
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

\[
M_{i,j}^k = \frac{\exp((\log(m_{i,j}^k) + g_{i,j}^k)/\tau)}{\sum_{i=1}^N \exp((\log(m_{i,j}^k) + g_{i,j}^k)/\tau)},
\]

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

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 \( \hat{z} \) and its corresponding mask \( M^c = \sum_{k=0}^{k} M^k \) which records the positions of all already decoded ones, the output \( \hat{z}' \) is as follows

\[
\hat{z}' = \frac{\text{Deconv}(z)}{\max(1, \text{Deconv}_1(M^c))},
\]

where Deconv and Conv1 are deconvolution layers with stride of 1. Besides, Deconv1 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 \( \hat{v} \) and already encoded/decoded partial of \( \hat{z} \) to predict the distribution of the to-be-processed part of \( \hat{z} \) as follows

\[
\mu^{i+1}, \sigma^{i+1} = g_c(g_z((\sum_{k=0}^{i} M^k) \odot \hat{z}, y), h_s(\hat{v})),
\]

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

\[
q_{\hat{z} | \hat{v}, y}(\hat{z}_i | \hat{v}, y) = (N(\mu_{k_i}^{c}, \sigma_{k_i}^{c}) \ast \mathcal{U}(-\frac{1}{2}, \frac{1}{2}))(\hat{z}_i)
\]

\[
= c_{\mu_{k_i}^{c}, \sigma_{k_i}^{c}}(\hat{z}_i + 0.5) - c_{\mu_{k_i}^{c}, \sigma_{k_i}^{c}}(\hat{z}_i - 0.5),
\]

where the subscript \( i \) is the index of the pixel position, \( k_i \) is the index of parameter groups defined in Eq. 7, and \( c_{\mu_{k_i}^{c}, \sigma_{k_i}^{c}}(\cdot) \) 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 versa. 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
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 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 \( \bm{n}(\mathbf{v}) \). 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 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.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>bpp ↓</th>
<th>Samsung NX2000</th>
<th>Olympus E-PL6</th>
<th>Sony SLT-A57</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR ↑</td>
<td>SSIM ↑</td>
<td>PSNR ↑</td>
<td>SSIM ↑</td>
</tr>
<tr>
<td>RIR [27]</td>
<td>3.253e-2</td>
<td>45.66</td>
<td>0.9939</td>
<td>48.42</td>
</tr>
<tr>
<td>SAM [30]</td>
<td>7.500e-1</td>
<td>47.03</td>
<td>0.9962</td>
<td>49.35</td>
</tr>
<tr>
<td>Nam et al. [26]</td>
<td>8.438e-1</td>
<td>48.08</td>
<td>0.9968</td>
<td>50.71</td>
</tr>
<tr>
<td>[26] w/ fine-tuning</td>
<td>8.438e-1</td>
<td>49.57</td>
<td>0.9975</td>
<td>51.54</td>
</tr>
<tr>
<td>Ours</td>
<td>2.887e-4</td>
<td>57.84±0.89</td>
<td>0.9997±0.00</td>
<td>59.08±0.95</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Quality</th>
<th>Method</th>
<th>bpp ↓</th>
<th>Samsung NX2000</th>
<th>Olympus E-PL6</th>
<th>Sony SLT-A57</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PSNR ↓</td>
<td>SSIM ↑</td>
<td>PSNR ↓</td>
<td>SSIM ↑</td>
</tr>
<tr>
<td>10</td>
<td>InvISP</td>
<td>N/A</td>
<td>26.62</td>
<td>0.8836</td>
<td>29.12</td>
</tr>
<tr>
<td></td>
<td>SAM</td>
<td>9.556e-4</td>
<td>24.42</td>
<td>0.8946</td>
<td>25.24</td>
</tr>
<tr>
<td></td>
<td>SAM</td>
<td>9.522e-3</td>
<td>27.94</td>
<td>0.9234</td>
<td>28.22</td>
</tr>
<tr>
<td></td>
<td>Nam et al.</td>
<td>8.438e-1</td>
<td>33.06</td>
<td>0.9737</td>
<td>34.03</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>7.736e-4</td>
<td>33.13</td>
<td>0.9836</td>
<td>34.04</td>
</tr>
<tr>
<td>30</td>
<td>InvISP</td>
<td>N/A</td>
<td>28.71</td>
<td>0.9931</td>
<td>31.76</td>
</tr>
<tr>
<td></td>
<td>SAM</td>
<td>9.556e-4</td>
<td>28.88</td>
<td>0.9344</td>
<td>30.21</td>
</tr>
<tr>
<td></td>
<td>SAM</td>
<td>9.522e-3</td>
<td>34.24</td>
<td>0.9553</td>
<td>35.87</td>
</tr>
<tr>
<td></td>
<td>Nam et al.</td>
<td>8.438e-1</td>
<td>37.21</td>
<td>0.9234</td>
<td>38.22</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>3.613e-4</td>
<td>37.40</td>
<td>0.9640</td>
<td>38.81</td>
</tr>
<tr>
<td>50</td>
<td>InvISP</td>
<td>N/A</td>
<td>30.02</td>
<td>0.9416</td>
<td>32.91</td>
</tr>
<tr>
<td></td>
<td>SAM</td>
<td>9.556e-4</td>
<td>30.78</td>
<td>0.9448</td>
<td>32.41</td>
</tr>
<tr>
<td></td>
<td>SAM</td>
<td>9.522e-3</td>
<td>36.32</td>
<td>0.9629</td>
<td>37.77</td>
</tr>
<tr>
<td></td>
<td>Nam et al.</td>
<td>8.438e-1</td>
<td>38.34</td>
<td>0.9686</td>
<td>40.07</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>3.368e-4</td>
<td>38.67</td>
<td>0.9699</td>
<td>40.33</td>
</tr>
<tr>
<td>70</td>
<td>InvISP</td>
<td>N/A</td>
<td>31.55</td>
<td>0.9476</td>
<td>33.74</td>
</tr>
<tr>
<td></td>
<td>SAM</td>
<td>9.556e-4</td>
<td>32.08</td>
<td>0.9529</td>
<td>34.14</td>
</tr>
<tr>
<td></td>
<td>SAM</td>
<td>9.522e-3</td>
<td>37.42</td>
<td>0.9684</td>
<td>38.96</td>
</tr>
<tr>
<td></td>
<td>Nam et al.</td>
<td>8.438e-1</td>
<td>39.13</td>
<td>0.9724</td>
<td>41.01</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>3.210e-4</td>
<td>39.59</td>
<td>0.9742</td>
<td>41.36</td>
</tr>
<tr>
<td>90</td>
<td>InvISP</td>
<td>N/A</td>
<td>33.86</td>
<td>0.9476</td>
<td>35.74</td>
</tr>
<tr>
<td></td>
<td>SAM</td>
<td>9.556e-4</td>
<td>34.37</td>
<td>0.9663</td>
<td>36.60</td>
</tr>
<tr>
<td></td>
<td>SAM</td>
<td>9.522e-3</td>
<td>39.17</td>
<td>0.9787</td>
<td>40.79</td>
</tr>
<tr>
<td></td>
<td>Nam et al.</td>
<td>8.438e-1</td>
<td>40.32</td>
<td>0.9782</td>
<td>42.33</td>
</tr>
</tbody>
</table>
|         | Ours   | 2.944e-4 | 41.19          | 0.9821        | 42.98       | 0.9856      | 43.43       | 0.9882     

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

<table>
<thead>
<tr>
<th>Ours w/o ( m(v) )</th>
<th>bpp ↓</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samsung</td>
<td>2.92e-4</td>
<td>57.50</td>
<td>0.9997</td>
</tr>
<tr>
<td>Olympus</td>
<td>2.93e-4</td>
<td>58.93</td>
<td>0.9997</td>
</tr>
<tr>
<td>Sony</td>
<td>2.90e-4</td>
<td>59.05</td>
<td>0.9997</td>
</tr>
<tr>
<td>Mean</td>
<td>2.92e-4</td>
<td>58.66</td>
<td>0.9997</td>
</tr>
</tbody>
</table>

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.

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


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


[40] Yi Zhang, Hongwei Qin, Xiaogang Wang, and Hongsheng Li. Rethinking noise synthesis and modeling in raw denois-