Compress Clean Signal from Noisy Raw Image: A Self-Supervised Approach

Zhihao Li [*1] **Yufei Wang** [*1] **Alex Kot** [1] **Bihan Wen** [1]

Abstract

Raw images offer unique advantages in many low-level visual tasks due to their unprocessed nature. However, this unprocessed state accentuates noise, making raw images challenging to compress effectively. Current compression methods often overlook the ubiquitous noise in raw space, leading to increased bitrates and reduced quality. In this paper, we propose a novel raw image compression scheme that selectively compresses the noise-free component of the input, while discarding its real noise using a self-supervised approach. By excluding noise from the bitstream, both the coding efficiency and reconstruction quality are significantly enhanced. We curate a full-day dataset of raw images with calibrated noise parameters and reference images to evaluate the performance of models under a wide range of input signal-noise ratios. Experimental results demonstrate that our method surpasses existing compression techniques, achieving a more advantageous rate-distortion balance with improvements ranging from +2 to +10dB and yielding a bit saving of 2 to 50 times. The code is available at https://lizhihao6.github.io/Cleans.

1. Introduction

In the camera imaging system, photons converge on the sensor chip to produce the raw image before an image signal processor (ISP) transforms it into an RGB image. Raw images are unprocessed and have a higher dynamic range, leading to various applications. For instance, their original noise distribution simplifies tasks such as image denoising (Abdelhamed et al., 2020; Feng et al., 2022; Wei et al., 2020) and low-light image enhancement (Ershov et al., 2022; Li et al., 2022b; Wang et al., 2022). Photographers and filmmakers favor raw images for extensive post-production flexibility.

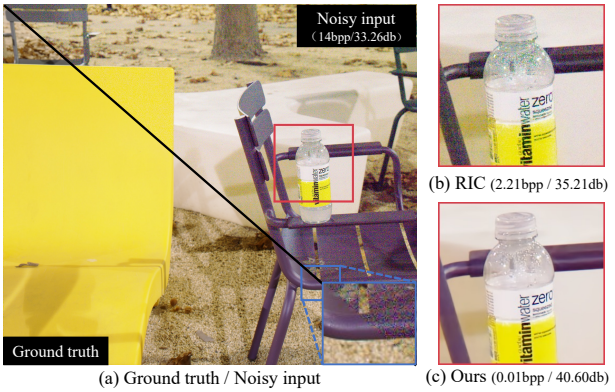


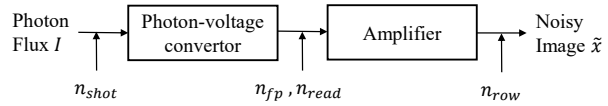
Figure 1. Comparison between the latest state-of-the-art (SOTA) raw image compression method, RIC (Li et al., 2022a), and our approach. Our method markedly surpasses RIC in both compression ratio and reconstruction quality, and achieves much higher PSNR comparing to the noisy input before compression. It is worth noting that our method is self-supervised, *i.e.*, we do not need clean images during training. For enhanced detail visibility, all raw images are converted to RGB. *(Zoom in for better view.)*

Additionally, the broader dynamic range of raw images benefits both low-level High Dynamic Range (HDR) imaging (Hasinoff et al., 2016) and high-level computer vision tasks (Li et al., 2022a).

Nevertheless, the substantial size of raw images poses significant challenges for storage and transfer, limiting their widespread use. The commonly used standard, Digital Negative (DNG), relies on the outdated JPEG-98, which no longer meets modern compression needs (Li et al., 2022a). Current image compression algorithms like JPEG, BPG, and PNG are primarily designed for RGB, monochrome, or YCbCr color spaces and require substantial modifications for raw images. Consequently, recent works (Wang et al., 2023a; Li et al., 2022a) have shifted towards learning-based methods for raw image compression, eliminating the need for manually designed features and providing a more efficient solution. Metadata-based methods, such as those described in (Wang et al., 2023a), utilize corresponding JPEG-compressed RGB images as a prior to reconstruct lossy raw images from additional bitstreams. Using JPEG images for raw compression is suboptimal, as it requires additional storage for RGB data converted from the raw files. Conversely, RIC (Li et al., 2022a) employs a learned image

[*]Equal contribution [1]Department of EEE, Nanyang Technology University, Singapore. Correspondence to: Bihan Wen <bihan.wen@ntu.edu.sg>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.



(a) Noise is inevitable in the imaging process

(b) Noisy input (c) Ground truth (d) Error map

Figure 2. An illustration of the prevalent noise in raw images: (a) The noise model for raw images demonstrates that noise is unavoidable in raw captures, even under sufficient illumination. (b) An example of a noisy input image shot at ISO 100 in well-lit indoor conditions. (c) A clean image, composed by merging 25 shots. (d) The error map highlights that noise remains conspicuous even in daylight conditions. All raw images have been processed to enhance visualization. *(Zoom in for better view.)*

compressor proposed in (Lu et al., 2022), which encodes raw images directly into bitstreams, bypassing the need for ISP conversion.

While substantial progress has been achieved, previous methods often neglect the prevalent noise characteristic of raw images, as illustrated in Fig.2. This noise is notably more pronounced in the raw domain compared to RGB, due to the lack of noise reduction and smoothing processes in the ISP pipeline. This not only increases the bitstream size but also potentially impairs the efficacy of deep learning-based compression techniques, as depicted in Fig.1. Cheng et al. (2022) have adapted the RGB image compression algorithm to be noise-sensitive, combining denoising and compression to address bit misallocation. However, their approach relies heavily on noise-free paired data for model optimization. For raw images, which exhibit significant variations in noise distribution and color space across different camera models, implementing this technique would necessitate the compilation of extensive, camera-specific datasets, which is largely impractical.

To overcome the outlined challenges, we propose a novel self-supervised framework for joint denoising and compression of raw images. Specifically, the underlying distribution of noise in the raw image is predicted, adhering to a physical-based prior. Simultaneously, the compression branch aims to reconstruct the clean image using the predicted noise model under a constrained bitstream. Given that real noise can hardly survive in low-dimensional subspace with a limited bitstream, the noise and clean signal can be well disentangled without the need for clean images. Compared to previous raw image compression methods, our approach significantly reduces bitstream sizes, by effectively decoupling noise from noisy images. Our contributions are summarized as follows:

- We propose the first self-supervised approach for raw image compression that incorporates joint denoising using a physical-based noise model.
- We propose a large-scale, full-day dataset for raw image compression, featuring accurately calibrated camera noise model parameters and noise-clean image pairs for evaluation. This establishes a solid benchmark for evaluating raw image compression methods.
- Our method significantly outperforms the existing raw image compression methods and those two-stage methods (both self-supervised and supervised) across a wide range of signal-to-noise ratios (SNRs) and cameras without any additional inference overhead.

2. Related work

2.1. Raw image denoising

Denoising w/ real paired data. Raw image denoising is crucial for enhancing image quality in both professional photography (Hasinoff et al., 2016) and scientific research (Levin et al., 2020; Joens et al., 2013). Traditional ISPs typically use methods like BM3D (Dabov et al., 2007) for noise reduction based on self-similarity. Recently, pioneering works like SID (Chen et al., 2018) and SIDD (Abdelhamed et al., 2018) have demonstrated that data-driven, deep-learning methods can effectively replace and even surpass traditional modules. Additionally, the ELD (Wei et al., 2020) dataset pushes raw image denoising capabilities into extreme low-light conditions. Beyond these datasets, recent developments in sophisticated restoration networks (Wang et al., 2023b; Chen et al., 2022; Abdelhamed et al., 2020) further advance the performance of raw image denoising.

Denoising w/ camera noise model. While noisy-clean paired datasets have led to significant advancements, collecting large-scale paired datasets including diverse noise patterns is laborious. As a result, noise model-based methods have emerged (Wei et al., 2020; Feng et al., 2022), utilizing synthetic noise from clean raw images. The effectiveness of these methods largely depends on the realism of the synthetic noise model. For instance, ELD (Wei et al., 2020) introduces an amended row-based noise model and replaces the Gaussian distribution with the Tukey Lambda distribution to better represent the long-tail noise pattern in extreme low-light conditions. PMN (Feng et al., 2022) further investigates the influence of Fixed Pattern Noise in dark frames, enhancing noise adherence to the Poisson-Gaussian (P-G) distribution. While calibration-based methods reduce the need for extensive data collection, they still require camera-specific clean raw images and are limited in dynamic scenes due to the long exposure times needed for clean image capture. To address these limitations, Neighbor2Neighbor (Huang et al., 2021) only uses spatially ir-

relevant noise characteristics, training denoising models on noisy images only. Yet, it does not incorporate prior of the precise noise model. In our framework, by integrating a noise model with a compressor, we effectively separate noise from noisy images, thereby obviating the need for clean images.

2.2. Raw image compression

Cameras typically employ Bayer color filter array (CFA) patterns to capture color information in scenes, with most CFAs consisting of RGGB four-channel arrangements. However, prevalent compression algorithms like JPEG (Wallace, 1991), HEVC (Sullivan et al., 2012), and VVC (Bross et al., 2021) are designed for three-channel RGB or single-channel grayscale images and thus don't directly support raw image compression. To leverage existing compression solutions, traditional raw image compression methods typically divide the raw image into one to three-channel sub-images for separate compression using current encoders (Lee & Ortega, 2001). Recently, deep-learning-based compression methods were proposed (Lu et al., 2022; Ballé et al., 2018) and even surpassed the advanced traditional codec VVC in the RGB domain. This trend has led to studies exploring deep-learning decoders specifically for raw images. For instance, Wang et al. (2023a; 2023c) utilize a context model to encode additional metadata, reconstructing raw images from JPEG-compressed RGB images. Apart from metadata-based approaches, RIC (Li et al., 2022a) directly employs a VAE for raw image compression, significantly outperforming traditional raw compression algorithms. Nonetheless, both conventional and deep-learning-based approaches generally overlooked the prevalent noise in raw images, which can increase bitstreams and degrade reconstruction quality. We find that with a known noise model, noise can be efficiently removed by a deep-learning-based VAE encoder.

3. Preliminaries

In this section, we elaborate the noise modeling we adopted to regularize the noise disentangled from raw images and the rate-distortion theory in learned raw image compression.

3.1. Noise model

In the camera imaging process, the captured raw image, denoted as $\tilde{x}$, is composed of the clean raw image x and the additive real noise n, which is defined as below:

$$\tilde{x} = x + n. \tag{1}$$

Inspired by the existing noise model (Wei et al., 2020; Feng et al., 2022), we decompose n into different types of noise components, which is represented as

$$n = n_{shot} + n_{read} + n_{row} + n_{fp}, \tag{2}$$

where n_{shot}, n_{read}, n_{row}, and n_{fp} stand for shot noise, read noise, row noise, and fixed pattern noise, respectively. A detailed description of each noise component can be found in the supplementary material.

Given that n_{shot} can be approximated from a Poisson distribution $\mathcal{P}(\frac{x}{k}) \cdot k - x$ to a Gaussian distribution $\mathcal{N}(0, x \cdot k)$, it can be combined with n_{read} to form a heteroscedastic Gaussian noise model as

$$n_{hg} \sim \mathcal{N}(0, \sigma_{hg}^2),\ \sigma_{hg}^2 = \sigma_{read}^2 + x \cdot k, \tag{3}$$

where k is the system gain and σ_{read} is the read noise standard deviation.

Consequently, the overall noise model can be simplified as

$$n = n_{hg} + n_{row} + n_{fp}. \tag{4}$$

By utilizing (4) as the prior of noising modeling, one can potentially decouple clean image from noise corruption, thereby obtaining more accurate estimation of the clean image x.

3.2. Lossy image compression

Lossy image compression for real data distribution, as grounded in rate-distortion theory (Shannon, 1948), aims to optimize the bit-rate $\mathcal{R}(D)$ through the rate-distortion function:

$$\mathcal{R}(D) = \min I(\hat{X}; X) \quad \text{s.t.} \quad \mathbb{E}[\Delta(\hat{X}, X)] \leq D, \tag{5}$$

where Δ denotes the distortion metric, I is the mutual information, and $\hat{x} \sim q_{\hat{X}|X}$ is the reconstructed image from the compressor. Achieving this for a given D involves optimizing a Lagrangian relaxation with corresponding λ_D:

$$\min[I(\hat{X}; X) + \lambda_D \cdot \mathbb{E}[\Delta(\hat{X}, X)]]. \tag{6}$$

Given the challenges in modeling p_{data}, Ballé et al. (2016) introduced a Variational Autoencoder (VAE) approach. This framework employs an encoder $y = g_a(x; \phi)$, mapping images to the latent space $q_{Y|X}$, and a decoder, reconstructing from quantized features $Q(y)$ as $\hat{x} = g_s(Q(y); \Omega)$. They approximate the upper bound of $I(\hat{X}; X)$ using:

$$I(\hat{X}; X) \leq I(Y; X) \leq \mathbb{E}[D_{\text{KL}}(q_{Y|X} \parallel p_Y)]. \tag{7}$$

Therefore, the total loss for the learned lossy image compression model is formulated as:

$$\mathcal{L}(\lambda_D) = \mathbb{E}[D_{\text{KL}}(q_{Y|X} \parallel p_Y)] + \lambda_D \cdot \mathbb{E}[\Delta(\hat{X}, X)]. \tag{8}$$

To further optimize the rate-distortion trade-off, considering the spatial redundancy in Y, we adopt the following overall framework for both our method and baselines where a

hyperprior p_V is employed in (Ballé et al., 2018):

$$\begin{aligned}\mathcal{L}(\lambda_D) &= \mathbb{E}[D_{\mathrm{KL}}(q_{Y|X,V} \parallel q_{Y|V})] \\ &\quad + \mathbb{E}[D_{\mathrm{KL}}(q_{V|X} \parallel p_V)] + \lambda_D \cdot \mathbb{E}[\Delta(\hat{X}, X)] \\ &= \underbrace{\mathcal{R}(Y) + \mathcal{R}(V)}_{\text{rate}} + \lambda_D \cdot \underbrace{\mathbb{E}[\Delta(\hat{X}, X)]}_{\text{distortion}}.\end{aligned} \tag{9}$$

4. Methodology

4.1. Motivation

Most of the existing raw image compression methods (Wang et al., 2023a; Li et al., 2022a) neglect the noise elements, by directly optimizing Eq.(5) with noisy raw images $\tilde{X}$. They inherently limit their optimization to:

$$\mathcal{R}(D) = \min I(\hat{X}; \tilde{X}) \quad \text{s.t.} \quad \mathbb{E}[\Delta(\hat{X}, \tilde{X})] \leq D. \tag{10}$$

Cheng et al.(2022) proposed a simultaneous denoising and compression strategy, optimizing:

$$U(D) = \min I(\hat{X}; \tilde{X}) \quad \text{s.t.} \quad \mathbb{E}[\Delta(\hat{X}, X)] \leq D. \tag{11}$$

This leads to the relationship:

$$U(0) = I(X; \tilde{X}) < I(\tilde{X}; \tilde{X}) = R(0), \tag{12}$$

indicating that joint denoising can enhance compression efficiency (the proof of Eq. (12) could be found in the supplementary). However, this method still relies on clean images X for training, which are challenging and expensive to obtain. To address this, as shown in Fig. 3, we propose a novel self-supervised framework that can implicitly optimize $I(X; \tilde{X})$ by regularizing the disentangled noise using a physical-based noise model. Specifically, the optimization objective is defined as follows:

$$\begin{aligned}&U(D, P) = \min I(\hat{X}; \tilde{X}) \\ &\text{s.t.} \quad \mathbb{E}[\Delta(\hat{X}, \tilde{X} - \hat{N})] \leq D,\ \mathbb{E}[d(\hat{N}, N)] \leq P,\end{aligned} \tag{13}$$

where d measures the divergence between predicted noise $\hat{N}$ and actual noise pattern N modeled in Sec. 3.1 and the implementation of d will be introudced in Sec. 4.2.

4.2. Regularization on the extracted noise

To optimize Eq. (13), we extend the Lagrangian relaxation approach in Eq. (6) as follows:

$$\begin{aligned}\min[&I(\hat{X}; \tilde{X}) + \lambda_D \cdot \mathbb{E}[\Delta(\hat{X}, \tilde{X} - \hat{N})] \\ &+ \lambda_P \cdot \ \mathbb{E}[d(\hat{N}, N)]].\end{aligned} \tag{14}$$

Estimating the divergence $d(\hat{N}, N)$ between $q_{\hat{N}}$ extracted from the noise extractor $F_n(\cdot; \Omega)$ and the physical-based noise modeling p_N in the original noise space is challenging due to the complex distribution of various noise components. To address this, inspired by normalizing flow (Rezende & Mohamed, 2015), we apply a series of bijective transformations to standardize the noise into a standard Gaussian distribution

$$z = f_{hg} \circ f_{row} \circ f_{fp}(n), \tag{15}$$

where f_{hg}, f_{row}, and f_{fp} are transformations for heteroscedastic Gaussian noise, row noise, and fixed pattern noise, respectively, $n \sim p_N$, and $\circ$ denotes function composition. We regularize the distribution of the predicted noises by first transforming into the space of Z as:

$$\hat{z} = f_{hg} \circ f_{row} \circ f_{fp} \circ F_n(\tilde{x}; \Omega), \tag{16}$$

where the details of each transformation is as follows:

- **Fixed pattern noise reduction f_{fp}.** Fixed pattern noise is challenging to extract via neural networks since it varies between different ISO levels. Therefore, we subtract it before applying the noise extractor F_n:

$$\hat{n}_{hg} + \hat{n}_{row} = f_{fp} \circ F_n(\tilde{x}; \Omega) = F_n(\tilde{x} - n_{fp}; \Omega). \tag{17}$$

- **Row noise reduction f_{row}.** Row noise, a zero-mean Gaussian noise with row-specific variances, is mitigated by subtracting each row's mean, as follows:

$$\begin{aligned}\hat{z}_r &= \mathrm{mean}_{row}(\hat{n}_{hg} + \hat{n}_{row})/\sigma_{row}, \\ \hat{n}_{hg} &= f_{row}(\hat{n}_{hg} + \hat{n}_{row}) \\ &= \hat{n}_{hg} + \hat{n}_{row} - \mathrm{mean}_{row}(\hat{n}_{hg} + \hat{n}_{row}),\end{aligned} \tag{18}$$

  where σ_{row} is each row's standard deviation, modeled by row noise model.

- **Heteroscedastic Gaussian noise reduction f_{hg}.** For the reconstructed image $\hat{x}$, we estimate the heteroscedastic Gaussian noise variance from Eq. (3) as $\hat{\sigma}^2_{hg} = \sigma^2_{read} + \hat{x} \cdot k$. This allows us to standardize the heteroscedastic Gaussian noise:

$$\hat{z} = f_{hg}(\hat{n}_{hg}) = \hat{n}_{hg}/\hat{\sigma}_{hg}. \tag{19}$$

Noise regularization loss. After transforming the predcited noise $\hat{z} = F_n(\tilde{x}, \Omega)$ to a latent space that the distribution is known, we can regualize the predcited noise distribution by minimizing the negative log-likelihood (NLL) loss as

$$\mathcal{L}_{\hat{Z}} = \mathbb{E}_{q_{\hat{z}|\tilde{x}}}[NLL_{p_{\mathcal{Z}}}(\hat{z})]. \tag{20}$$

While this formulation accounts for the overall distribution of $\hat{z}$, it does not consider the independent and identically distributed (i.i.d.) noise across pixels. To address this, we

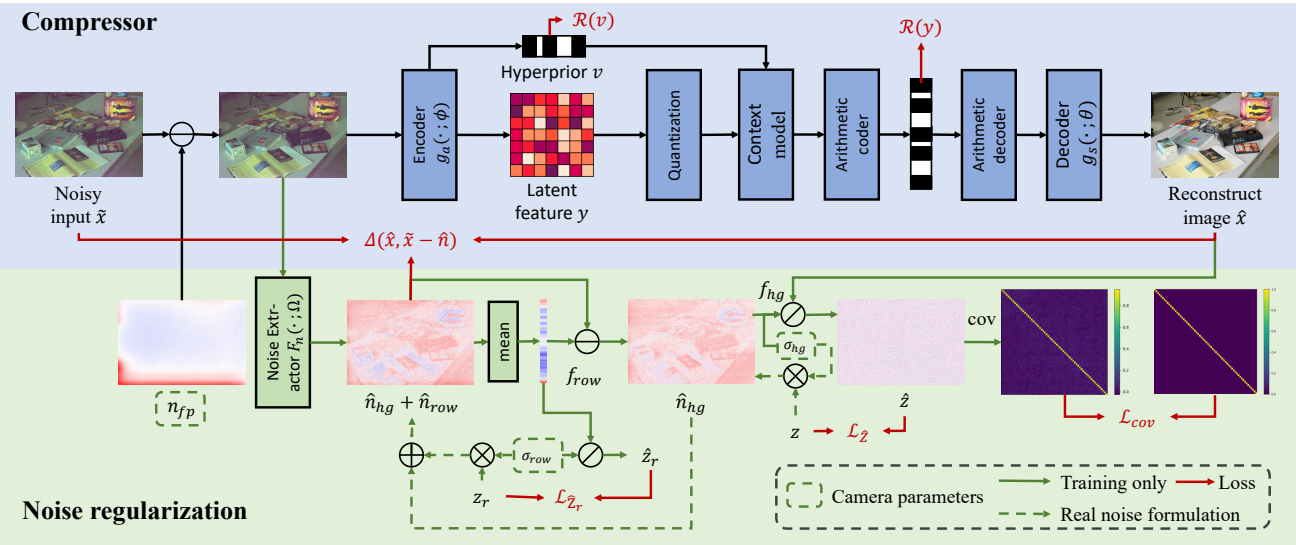


Figure 3. The proposed framework for self-supervised raw image denoising and compression without reliance on paired clean images. Distinct from typical learning-based compressors, our approach first subtracts fixed pattern noise n_{fp} from the noise input $\tilde{x}$ in the compressor encoder. Then, it compresses the predicted clean signal $\hat{x}$, constrained by $\tilde{x} - F_n(\tilde{x}; \Omega)$, using a compressor with an integrated hyperprior module. To regularize the predicted noise $\hat{n} = F_n(\tilde{x}; \Omega)$, we use a bijective mapping based on the physical-based noise model to map the complicated noise distribution to a latent space where the distribution is known. Besides, a covariance loss is used to enhance the spatial independence of the disentangled noise $\hat{n}$.

introduce a covariance loss $\mathcal{L}_{cov}$ to promote spatial independence the predicted noise as follows:

$$\mathcal{L}_{cov} = \mathbb{E}_{q_{\hat{z}|\tilde{x}}}[I - \hat{z}^T \hat{z}] \tag{21}$$

Besides, the subtracted each row's mean needs to obey the row-specific variances σ_{row} introudced in Eq. (3), which is regularized by the NLL loss below:

$$\mathcal{L}_{\hat{Z}_r} = \mathbb{E}_{q_{\hat{z}|\tilde{x}}}[NLL_{p_{\mathcal{Z}_r}}(\hat{z}_r)]. \tag{22}$$

Subsequently, the noise regularization loss $d(\hat{N}, N)$ is formulated as follows:

$$d(\hat{N}, N) = \mathcal{L}_{\hat{Z}} + \mathcal{L}_{\hat{Z}_r} + \lambda_{cov} \cdot \mathcal{L}_{cov}. \tag{23}$$

4.3. Overall training loss

As for the rate constraint $I(\hat{X}; \tilde{X})$ in Eq. (14), we adopt the common approximaiton of $\mathcal{R}(y) + \mathcal{R}(v)$ illustrated in Sec. 3.2. Specifically, the rates $\mathcal{R}(y)$ and $\mathcal{R}(v)$ are defined as follows:

$$\mathcal{R}(y) = \mathbb{E}[-\log_2 q_{\hat{y}|\hat{v}}(\hat{y}|\hat{v})], \mathcal{R}(v) = \mathbb{E}[-\log_2 q_{\hat{v}}(\hat{v})], \tag{24}$$

where $\hat{y}$ and $\hat{v}$ represent the quantized y and v, respectively.

To be consistent with established learned image compression methods (Ballé et al., 2018; Li et al., 2022a), the MSE loss L_{MSE} is selected as the distortion metric $\Delta(\hat{X}, X)$. Fianlly, the overall loss that simultaneous minimize the rate of the latent codes $\hat{y}, \hat{v}$, the reconstruction loss of the input image, and the noise regularization loss is summmarized as follows:

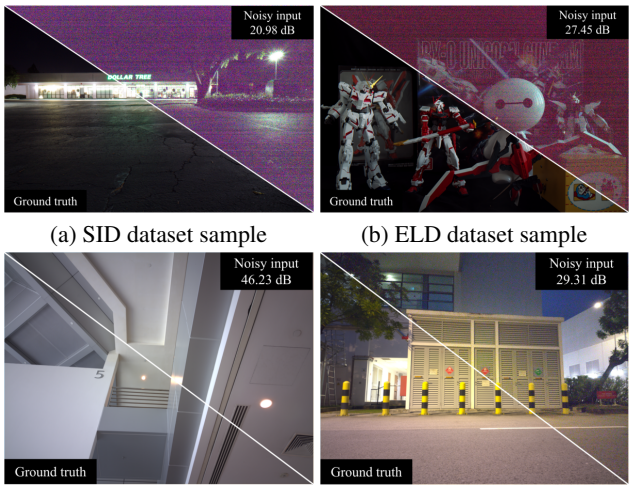
$$\mathcal{L} = \mathbb{E}[\underbrace{\mathcal{R}(\hat{y}) + \mathcal{R}(\hat{v})}_{\text{rate}} + \lambda_D \cdot \underbrace{L_{MSE}(\hat{x}, \tilde{x} - \hat{n})}_{\text{distortion}}] + \lambda_P \cdot \underbrace{d(\hat{N}, N)}_{\text{noise divergence}}. \tag{25}$$

Table 1. Comparison of various raw image denoising datasets. "Illum." denotes illumination, and "No." represents the number of training noise images available for each camera type.

Dataset	Illum. (lux)	Scenarios Indoor	Scenarios Outdoor	No.
SID (Chen et al., 2018)	< 5	✓	✓	280
ELD (Wei et al., 2020)	< 0.3	✓		120
FDRIC (Ours)	0.1-1e5	✓	✓	549

5. Full-day raw image compression dataset

Existing raw image denoising datasets mainly focus on low-light or nighttime scenes, *e.g.*, SID (Chen et al., 2018) captured under around 5 lux conditions and ELD (Wei et al., 2020) is even below 0.3 lux as shown in Fig. 4a. However, the demand for image compression is not only at night but throughout the whole day. Besides, as highlighted in Sec.1 and Fig.2, noise is also prevalent in daylight raw images, which possess considerably higher SNR. Due to the significant gap among input SNR, compressors trained solely on



(a) SID dataset sample (b) ELD dataset sample

(c) High SNR sample in FDRIC (d) Low SNR sample in FDRIC

Figure 4. Comparison of images across different datasets showcasing variations in noise levels. Subfigures (a) and (b) depict images from the SID and ELD datasets, respectively, while subfigures (c) and (d) illustrate high and low SNR images from our FDRIC dataset.

low SNR data are less effective in higher SNR scenarios. Therefore, the existing datasets cannot well meet the needs of training and performance evaluation.

To address this limitation, we develop a full-day raw image compression (FDRIC) dataset that covers a wide range of SNR, ensuring comprehensive training and evaluation. We collected our dataset using the Redmi Note12 Turbo smartphone, with an OV64B sensor of a resolution of 4624×3472. Our dataset includes 549 noisy images for training and 32 noise-clean image pairs for evaluation. To collect paired pairs for testing, we first captured short-exposure images with auto-adjusted ISO and exposure time, followed by 25 long-exposure shots at ISO-100. We ensured the same product of ISO and exposure time for both noisy and clean images to achieve consistent exposure levels. Our dataset contains indoor and outdoor scenes with illumination ranging from 0.1 to more than $1e5$ lux. Details on calibrating the camera noise model are provided in the supplementary.

6. Experiment

In this section, we begin by detailing the experimental setting. Next, we compare our methods to existing ones. Lastly, we perform comprehensive ablation studies for a thorough analysis of our approach.

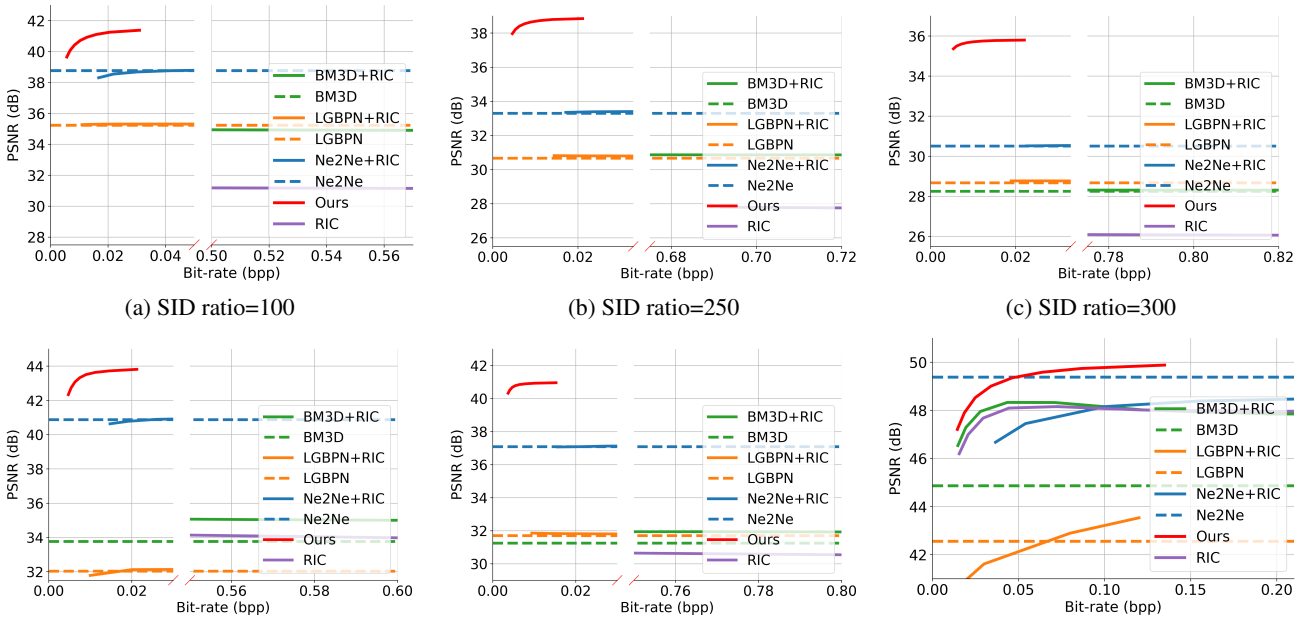
6.1. Experimental setting

Implementation details. For fair comparison, we use the same compressor as in RIC (2022a). Unlike RIC's need for eight different models to handle various rate-distribution trade-offs, our approach utilizes the quantization-error-aware variable rate framework (2023), enabling a wide range of continuous variable rates with a single model. For noise extraction F_n, we adopt the same U-Net architecture used in ELD (2020). We follow the train-test set split for the SID (2018) SonyA7S2 subset as outlined in PMN (2022). For the ELD SonyA7S2 subset, we directly evaluate it using a pretrained model, adhering to the same settings specified in PMN. For our FDRIC dataset, we crop these images into 512×512 patches for both training and testing. Adam optimizer with an initial learning rate of 1e-4 and a batch size of 6 is used, spanning 200,000 iterations with a learning rate decay to 1e-5 after 150,000 iterations. These hyperparameters remain consistent across all datasets, and we apply grad norm clipping for training stability as in RIC. For the hyperparameters λ_D, λ_P, and λ_{cov} in Eq. (23) and Eq. (25), the range of λ_D is set between 0.0018 and 0.18. The minimum value of λ_P is 0.05, and the maximum at 5, increasing at the same rate as λ_D. λ_{cov} is consistently maintained at 0.2. All the models are trained within CompressAI PyTorch framework (2020) using a single NVIDIA RTX 3090 GPU.

Compared methods. To validate the effectiveness of our framework, we compared our method with both one-stage compressors and two-stage denoiser+compressor approaches. As for the one-stage compressor baseline, we trained RIC using a noise-noise paired loss function, as defined in Eq. (10). In the two-stage setup, our method was compared against the traditional non-learned BM3D denoiser (2007), as well as the latest state-of-the-art (SOTA) learned self-supervised image denoisers: Neighbor2Neighbor (2021) and LGBPN (2023d). For these, the denoised outputs were compressed using the same compressor as RIC, noted as BM3D+RIC, Ne2Ne+RIC, and LGBPN+RIC, respectively. All self-supervised denoisers and compressors are trained on the SID and FDRIC datasets using the official codes.

6.2. Results

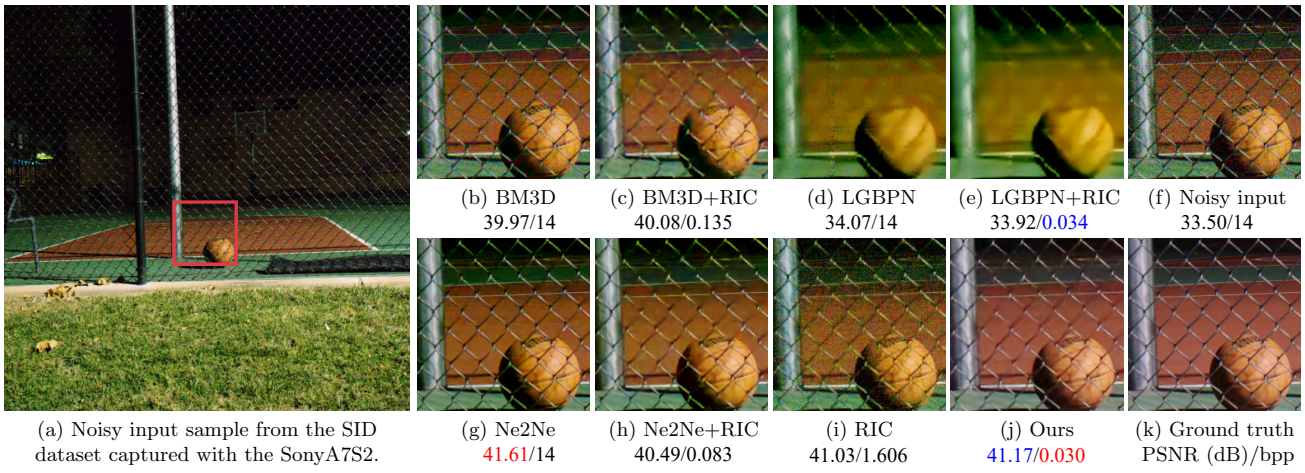
In Fig. 5, we present the Rate-Distortion (RD) curves across various SNR levels for the SID, ELD, and FDRIC datasets. At first glance, it is evident that our methods (indicated by the red RD curves) significantly surpass both the original RIC baselines and two-stage methods across all datasets and SNR levels. Compared to the original RIC baseline, which overlooks raw image noise, our method achieves a tenfold reduction in bitrates and more than 8dB gain in PSNR on the SID dataset. Additionally, our approach outperforms two-stage methods that combine state-of-the-art denoisers with the RIC compressor. The denoiser only results for the two-stage models are also illustrated in Fig. 5 as dotted lines, *i.e.*, the performances of denoisers. Our method exceeds both the traditional BM3D and the learned Neighbor2Neighbor denoisers, highlighting its potential as a superior raw image



(a) SID ratio=100 (b) SID ratio=250 (c) SID ratio=300

(d) ELD ratio=100 (e) ELD ratio=200 (f) FDRIC

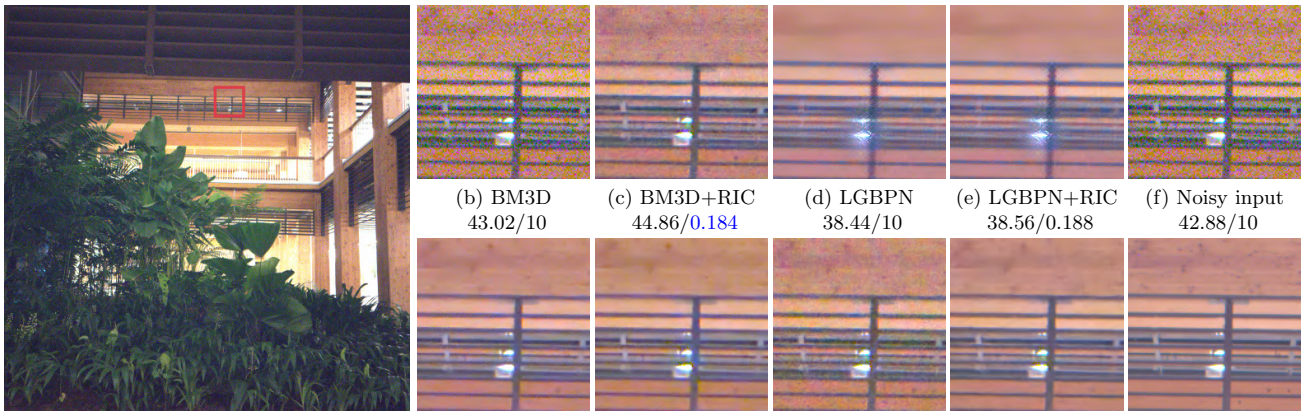
Figure 5. Rate-Distortion curves for various datasets across different cameras and SNR levels.



(a) Noisy input sample from the SID dataset captured with the SonyA7S2.

(b) BM3D 39.97/14 (c) BM3D+RIC 40.08/0.135 (d) LGBPN 34.07/14 (e) LGBPN+RIC 33.92/0.034 (f) Noisy input 33.50/14

(g) Ne2Ne 41.61/14 (h) Ne2Ne+RIC 40.49/0.083 (i) RIC 41.03/1.606 (j) Ours 41.17/0.030 (k) Ground truth PSNR (dB)/bpp



(a) Noisy input sample from the FDRIC dataset captured with the OV64B.

(b) BM3D 43.02/10 (c) BM3D+RIC 44.86/0.184 (d) LGBPN 38.44/10 (e) LGBPN+RIC 38.56/0.188 (f) Noisy input 42.88/10

(g) Ne2Ne 45.27/10 (h) Ne2Ne+RIC 44.88/0.452 (i) RIC 44.89/0.200 (j) Ours 49.14/0.144 (k) Ground truth PSNR (dB)/bpp

Figure 6. Visual comparison across various cameras. (*Zoom in for better view.*)

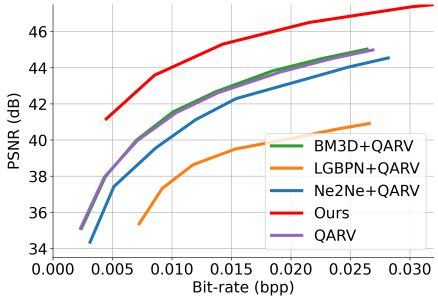


Figure 7. Rate-distortion curves for FDRIC dataset with the ratio of 100, using QARV as the baseline compressor.

Table 2. Ablation study on the selection of λ_P and λ_{cov}, withBD-rate calculations based on FDRIC datasets.

λ_P/λ_D	λ_{cov}	BD-Rate↓ (%)
5 / 0.18	0.2	0
0.1 / 0.18	0.2	48.88
30 / 0.18		28.64
5 / 0.18	0.0	5.35
	0.02	4.45
	2	3.80

denoiser. Visual comparisons in Fig. 6 show that our method attains better quality with lower bit rates.

6.3. Ablation study

We present several ablation studies to validate our method. These include testing different compressor architectures, examining hyperparameters λ_P and λ_{cov}, comparing with the SOTA supervised methods, evaluating inference speed, and analyzing the impact of training across various SNR levels. The evaluation metrics are the RD curve and Bjøntegaard-delta-rate (2001)(BD-Rate).

Different compressor architectures. To verify the generalization capability of our proposed framework, we evaluated it using the recent SOTA Q-VAE based compressor, QARV (2023). As shown in Fig. 7, our method consistently outperforms both QARV and two-stage baselines by a significant margin. This suggests that our approach is not confined to a specific compressor architecture, such as VAE, and can be seamlessly integrated into various compression frameworks.

Impact of hyperparameters. The hyperparameters λ_P and λ_{cov} in the loss function are critical for balancing the trade-off between compression rate and denoising performance. As shown in Table 2, a too small λ_P results in insufficient noise model constraints, leading to inaccurate noise extraction. Conversely, an excessively large λ_P generates overwhelming gradients from the noise regularization loss, adversely affecting compressor convergence. A similar issue arises with λ_{cov} when set either too low or too high.

Table 3. Comparison with supervised methods using BD-rate calculated on the SID dataset with the ratio of 100.

	Ours	SID+RIC	ELD+RIC
BD-Rate↓ (%)	0	8.35	59.50

Table 4. Comparison of encoding complexity and latency on the FDRIC dataset. * notes that FLOPs and parameters are not summarized. "N.A." indicates that RD curve does not intersect with ours.

Method	FLOPs (G)	Params (M)	Latency (s)	BD-rate (%)
Ours	1001	27.42	0.567	0
BM3D + RIC	1001*	27.42*	290.629	234.55
Ne2Ne + RIC	2830	28.68	0.796	317.69
LGBPN + RIC	172253	31.63	92.670	N.A.

Comparison with supervised methods. We benchmarked our method against supervised approaches that were trained with noise-clean paired data from SID (2018) and synthetic noise based on the noise model proposed in ELD (2020). The SID dataset's pretrained model served as the initial stage denoiser, combined with RIC for compression, denoted as SID+RIC or ELD+RIC. As indicated in Table 3, our unsupervised method surpasses even those supervised methods trained on paired data from the SID dataset. This underscores our method's ability to effectively separate noise from the signal without requiring paired data.

Encoding complexity and latency. We also compare the encoding complexity of our method with two staged methods. As shown in Table 4, our method achieves an impressive -234.55% BD-rate improvement compared to traditional BM3D methods and is 500 times faster than BM3D+RIC. Furthermore, we reduce FLOPs by 64% while achieving a remarkable 317.69% BD-rate improvement compared to Ne2Ne+RIC. Notably, our method eliminates the need for a denoiser to preprocess the input image, making it significantly more efficient than the two-stage approach.

7. Conclusion

In this work, we introduced a novel self-supervised framework for joint denoising and compression of raw images, effectively addressing the challenges of noise characteristics in raw imaging. By selectively compressing the noise-free component of the raw input and discarding the unwanted noise using a self-supervised technique, we significantly improve the coding efficiency while maintaining image quality. To further evaluate the performance of the proposed method, we curate a full-day dataset of raw images with calibrated noise parameters and reference images. The results on both existing benchmarks and the proposed dataset demonstrate the effectiveness of our method.

Impact Statement

This paper presents work whose goal is to advance the field of raw image compression. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgements

This research is partially supported by the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity Research & Development Programme (Development of Secured Components & Systems in Emerging Technologies through Hardware & Software Evaluation < NRF-NCR25-DeSNTU-0001 >).

References

Abdelhamed, A., Lin, S., and Brown, M. S. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1692–1700, 2018.

Abdelhamed, A., Afifi, M., Timofte, R., and Brown, M. S. Ntire 2020 challenge on real image denoising: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 496–497, 2020.

Ballé, J., Laparra, V., and Simoncelli, E. P. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

Bégaint, J., Racapé, F., Feltman, S., and Pushparaja, A. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.

Bjontegaard, G. Calculation of average psnr differences between rd-curves. *ITU SG16 Doc. VCEG-M33*, 2001.

Bross, B., Wang, Y.-K., Ye, Y., Liu, S., Chen, J., Sullivan, G. J., and Ohm, J.-R. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31 (10):3736–3764, 2021.

Chen, C., Chen, Q., Xu, J., and Koltun, V. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3291–3300, 2018.

Chen, L., Chu, X., Zhang, X., and Sun, J. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022.

Cheng, K. L., Xie, Y., and Chen, Q. Optimizing image compression via joint learning with denoising. In *European Conference on Computer Vision*, pp. 56–73. Springer, 2022.

Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8): 2080–2095, 2007.

Duan, Z., Lu, M., Ma, J., Huang, Y., Ma, Z., and Zhu, F. Qarv: Quantization-aware resnet vae for lossy image compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Ershov, E., Savchik, A., Shepelev, D., Banić, N., Brown, M. S., Timofte, R., Koščević, K., Freeman, M., Tesalin, V., Bocharov, D., et al. Ntire 2022 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1287–1300, 2022.

Feng, H., Wang, L., Wang, Y., and Huang, H. Learnability enhancement for low-light raw denoising: Where paired real data meets noise modeling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1436–1444, 2022.

Hasinoff, S. W., Sharlet, D., Geiss, R., Adams, A., Barron, J. T., Kainz, F., Chen, J., and Levoy, M. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016.

Huang, T., Li, S., Jia, X., Lu, H., and Liu, J. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14781–14790, 2021.

Joens, M. S., Huynh, C., Kasuboski, J. M., Ferranti, D., Sigal, Y. J., Zeitvogel, F., Obst, M., Burkhardt, C. J., Curran, K. P., Chalasani, S. H., et al. Helium ion microscopy (him) for the imaging of biological samples at sub-nanometer resolution. *Scientific reports*, 3(1):3514, 2013.

Lee, S.-Y. and Ortega, A. A novel approach of image compression in digital cameras with a bayer color filter array. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, volume 3, pp. 482–485. IEEE, 2001.

Levin, N., Kyba, C. C., Zhang, Q., de Miguel, A. S., Román, M. O., Li, X., Portnov, B. A., Molthan, A. L., Jechow, A., Miller, S. D., et al. Remote sensing of night lights: A review and an outlook for the future. *Remote Sensing of Environment*, 237:111443, 2020.

Li, Z., Lu, M., Zhang, X., Feng, X., Asif, M. S., and Ma, Z. Efficient visual computing with camera raw snapshots. *arXiv preprint arXiv:2212.07778*, 2022a.

Li, Z., Yi, S., and Ma, Z. Rendering nighttime image via cascaded color and brightness compensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 897–905, 2022b.

Lu, M., Chen, F., Pu, S., and Ma, Z. High-efficiency lossy image coding through adaptive neighborhood information aggregation. *arXiv preprint arXiv:2204.11448*, 2022.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Sullivan, G. J., Ohm, J.-R., Han, W.-J., and Wiegand, T. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012.

Tong, K., Wu, Y., Li, Y., Zhang, K., Zhang, L., and Jin, X. Qvrf: A quantization-error-aware variable rate framework for learned image compression. *arXiv preprint arXiv:2303.05744*, 2023.

Wallace, G. K. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

Wang, Y., Wan, R., Yang, W., Li, H., Chau, L.-P., and Kot, A. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2604–2612, 2022.

Wang, Y., Yu, Y., Yang, W., Guo, L., Chau, L.-P., Kot, A. C., and Wen, B. Beyond learned metadata-based raw image reconstruction. *arXiv preprint arXiv:2306.12058*, 2023a.

Wang, Y., Yu, Y., Yang, W., Guo, L., Chau, L.-P., Kot, A. C., and Wen, B. Exposurediffusion: Learning to expose for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12438–12448, 2023b.

Wang, Y., Yu, Y., Yang, W., Guo, L., Chau, L.-P., Kot, A. C., and Wen, B. Raw image reconstruction with learned compact metadata. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18206–18215, 2023c.

Wang, Z., Fu, Y., Liu, J., and Zhang, Y. Lg-bpn: Local and global blind-patch network for self-supervised real-world denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18156–18165, 2023d.

Wei, K., Fu, Y., Yang, J., and Huang, H. 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*, pp. 2758–2767, 2020.

Make Raw Image Compression Noise-Free: A Self-Supervised Approach — Supplementary Material

In this supplement, we first elaborate on the distribution and calibration of the camera noise model. Then, we provide a brief proof for Eq. (12).

A. Distribution of the camera noise model

Following ELD (Wei et al., 2020) and PMN (Feng et al., 2022), the noise components in Eq. (2) follow specific distributions:

$$\begin{aligned} n_{shot} &\sim \mathcal{P}\left(\frac{x}{k}\right) \cdot k - x, \\ n_{read} &\sim \mathcal{N}(0, \sigma_{read}^2), \\ n_{row} &\sim \mathcal{N}(0, \sigma_{row}^2), \\ n_{fp} &= \text{ISO} \cdot n_{fp_k} + n_{fp_b}, \end{aligned} \tag{26}$$

where k represents the overall system gain linked to the ISO setting. $\mathcal{P}$ and $\mathcal{N}$ denote the Poisson and Gaussian distributions. The terms $n_{fp_k}, n_{fp_b} \in \mathbb{R}^{H \times W}$ are pixel-wise dark frame noise components. Following assumptions made in studies like ELD (Wei et al., 2020) and PMN (Feng et al., 2022), the relationships between k, σ_{read}, σ_{row}, and ISO are given as:

$$\begin{aligned} k &= a_k \cdot \text{ISO} + b_k, \\ \log\left(\sigma_{read}\right) &= a_{read} \cdot \log(k) + b_{read}, \\ \log\left(\sigma_{row}\right) &= a_{row} \cdot \log(k) + b_{row}. \end{aligned} \tag{27}$$

The set of parameters n_{fp_k}, n_{fp_b}, a_k, b_k, a_{read}, b_{read}, a_{row}, b_{row}, specific to each camera, can be calibrated using a series of flat-frame and dark-frame images captured at various ISO levels.

B. Calibration of the camera noise model paramenters

The calibration process involves three steps: initially, k_i is calibrated for each ISO_i using flat frames; then, n_{fp_i}, σ_{read_i}, and σ_{row_i} are calibrated with dark frames for each ISO. Finally, ISO-related parameters a_k, b_k, n_{fp_k}, n_{fp_b}, a_{read}, b_{read}, a_{row}, and b_{row} are fitted using the calibrated noise parameters from the first two steps across various ISOs. Specifically,

- First, to calibrate k_i at each ISO_i, we capture 25 flat frames under consistent lighting for each exposure time Exp_j, calculating mean and variance for each color block. This yields 24 mean-variance pairs per exposure time Exp_j. With three different exposure times per ISO_i, we gather 72 mean-variance pairs per ISO. The n_{shot}, modeled as $\mathcal{N}(x, x \cdot k)$ where x is the mean and $x \cdot k$ the variance, allows us to calibrate k from the mean-variance relationship. Points with a mean value beyond $1/4$ saturation are excluded due to clipping effects.
- Next, after calibrating k_i, we capture 100 dark frames at each ISO_i in a dark room to calibrate n_{fp_i}, σ_{read_i}, and σ_{row_i}. The mean of these dark frames gives n_{fp_i}, representing fixed pattern noise. Subtracting n_{fp_i} from all dark frames, we calculate variance across rows for n_{row} and total frame variance for n_{read}.
- Finally, by repeating the steps above for different ISO levels and obtaining a set of parameters $\{n_{fp_i}, \sigma_{read_i}, \sigma_{row_i}\}$, we fit a_k, b_k, n_{fp_k}, n_{fp_b}, a_{read}, b_{read}, a_{row}, and b_{row} based on these parameters and equations Eq. (26),27.

We develop an Andriod application to semi-automatically collect the aforementioned flat and dark frames. The calibration app and corresponding calibration codes will be released upon acceptance.

C. Proof of $I(X; \tilde{X}) < I(\tilde{X}; \tilde{X})$

The derivation of Eq. (12) stems from mutual information principles:

$$I(X; \tilde{X}) = H(\tilde{X}) - H(\tilde{X}|X) = H(\tilde{X}) - H(X + N|X) \leq H(\tilde{X}) = I(\tilde{X}; \tilde{X}), \tag{28}$$

where $H(X + N|X) > 0$ due to the presence of signal-related noise, n_{shot}, which is unavoidable.