# Not All Noise Is Harmful: Towards Perception-Aware and Controllable Joint RAW Image Denoising and Demosaicing

**Qianjun Huang**[1] **Qingguo Liu**[1] **Hui Zeng**[2] **Kai Zhang**[1*] **Jian Yang**[1]
[1] Nanjing University [2] OPPO Research Institute
qianjun.huang@smail.nju.edu.cn kaizhang@nju.edu.cn

## Abstract

RAW image denoising and demosaicing play a critical role in the early stages of the image signal processing (ISP) pipeline. Previous methods handle them sequentially, which can introduce error propagation. Unified restoration, such as joint denoising and demosaicing (JDD), has become a widely adopted paradigm and achieves impressive performance. Nonetheless, under constraints of model capacity and low-light conditions, these methods tend to apply overly aggressive denoising, causing notable smearing effects and the erosion of fine textures. Researchers attempt to moderate the denoising strength to better preserve details, but this often results in noticeable artifacts. These observations suggest a key insight: Not All Noise Is Harmful—retaining an appropriate amount of noise can be preferable to over-smoothing and is often aligned with user aesthetics. Moreover, most existing methods are mainly trained on paired clean–degraded data, yielding fixed, hard-to-control denoising behavior that fails to meet evolving user demands. In this paper, we propose **JDD-PRO**, a **P**erception-awa**R**e and contr**O**llable framework for RAW image JDD. Built on a controllable JDD training scheme and a human visual system (HVS)-inspired perceptual adapter, our method allows users to tune the denoising strength at inference to strike their desired balance between noise removal and detail preservation. Comprehensive experiments on real-world datasets validate the effectiveness and robustness of our proposed method. Code is available at `https://github.com/hqj9994ever/JDD-PRO`.

## 1 Introduction

With the rapid advancement of modern camera imaging systems, image restoration has become a key component in enhancing visual quality and improving the user experience. In digital photography, restoration typically involves two fundamental tasks: denoising and demosaicing. These tasks work together to suppress sensor noise, reconstruct missing color information, and correct optical distortions, ultimately aiming to recover clean, sharp, and high-fidelity images from degraded RAW.

In recent years, the rapid progress of deep learning has led to substantial improvements in the performance of these restoration tasks. Most existing methods (1; 2; 3; 4) are designed for a single task or apply separate restoration modules sequentially, which leads to stage-wise error accumulation and increased solution complexity.

As an alternative, joint solutions have become the de facto choice. For instance, SGNet (7) proposes a self-guided joint denoising and demosaicing (JDD) model that uses the green channels as guidance to reconstruct missing values. To address degraded low-resolution RAW images, JDNDMSR (8) employs an end-to-end CNN that jointly performs demosaicing, denoising, and super-resolution. DR-

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

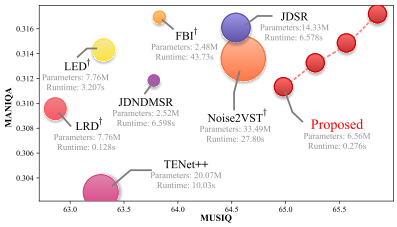
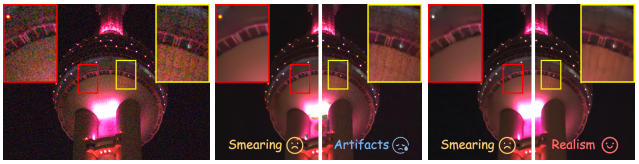



Figure 1: Quantitative comparison of RAW image denoising and JDD methods on the OPPO Find X7 Ultra. Circle size indicates the model parameters. † indicates the AHD (5) demosaicing algorithm is applied.




(a) Noisy Image (b) JDSR (6) (c) Ours

Figure 2: Denoising comparison between JDSR (6) and JDD-PRO. When the denoising strength matches the noise level, both methods show smearing effects, see (b, left) and (c, left). When the denoising strength is reduced, JDSR shows noticeable artifacts, while JDD-PRO preserves more details, see (b, right) and (c, right).

Net (9) proposes a dual-encoder JDD model that combines residual estimation with pixel modulation tailored to the Quad-Bayer pattern. These advances demonstrate the promise of joint restoration.

Nevertheless, real-world camera imaging remains challenging. Under practical constraints such as limited photon budget, sensor size, and computational resources, RAW images often exhibit severe noise, especially in low-light scenes. In practice, most methods often prioritize comprehensive noise removal (10; 6; 4; 8; 11). However, limited model capacity, compounded by high noise levels under low light, makes such aggressive denoising prone to severe loss of fine-grained details (Fig. 2(b), left). To preserve details, existing methods (12) typically moderate the denoising strength. However, this compromise frequently introduces visible artifacts (Fig. 2(b), right). These observations motivate a key insight: Not All Noise is Harmful—preserving an appropriate amount of noise can better maintain textures and avoid over-smoothing-induced smearing. Moreover, most existing methods rely mainly on paired clean–degraded supervision, which tends to lock in a largely fixed denoising behavior and limits user control, failing to meet modern user demands (*e.g.*, some users prefer smoother denoising results, while others favor the film-like grain and perceived authenticity of film photography). Taken together, these observations underscore the need for a JDD framework that can flexibly balance noise removal and detail preservation.

In this paper, we propose JDD-PRO, a perception-aware and controllable framework for RAW image JDD. Specifically, to mitigate smearing effects while preserving fine-grained details and rich textures, we introduce a novel controllable JDD (CTRL-JDD) training scheme. Unlike conventional methods that rely merely on noisy-clean image pairs, the proposed CTRL-JDD explicitly incorporates noise level-to-denoising strength mismatches during training. This design enables user-preferred control over the balance between noise removal and detail preservation.

Moreover, motivated by the human visual system (HVS) (13), we note that color noise, compared with grayscale noise, is more likely to disrupt the structural and chromatic consistency essential for stable perception (14; 15). Therefore, preserving luminance noise while suppressing chromatic noise is more desirable from a perceptual perspective. To this end, we propose a perceptual adapter to remove chromatic noise and thus enhance perceptual quality. As shown in Fig. 2(c) (right), JDD-PRO yields richer textures and better perceptual quality.

Additionally, to address the highly signal-dependent noise in RAW images (16) and spatially varying demosaicing difficulty, we design an efficient RAW JDD network with Degradation-aware Convolution (DAConv). Conditioned on the noise prior, DAConv predicts content-adaptive kernel weights to handle spatially varying degradations.

Extensive experiments on diverse real-world datasets demonstrate that our method outperforms state-of-the-art methods while maintaining high efficiency (see Fig. 1).

To sum up, our main contributions are:

- We propose JDD-PRO, a perception-aware and controllable framework for RAW image joint denoising and demosaicing.
- We introduce a controllable JDD training scheme to enable user control over denoising strength and effectively suppress smearing artifacts. Moreover, we propose an HVS-inspired perceptual adapter to further enhance perceptual quality.

- We design an efficient and robust RAW JDD network featuring a Degradation-aware Convolution to handle regionally uneven degradation in RAW images.
- Extensive experiments on real-world datasets demonstrate the effectiveness and robustness of our method.

## 2 Related Works

In recent years, significant advances have been made in RAW image restoration (17; 1; 18; 2; 4; 19; 7; 8; 10; 20; 21), with methods broadly falling into two groups: single-task methods and joint solutions.

**Single Tasks.** Brooks *et al.* (1) proposed a method that inverts the image processing pipeline to synthesize realistic RAW data from internet images, enabling denoising (DN) models to train on synthetic data. Jin *et al.* (11) proposed a calibration-free RAW image denoising pipeline for extreme low-light conditions. It adapts to new cameras using only a few paired samples and minimal fine-tuning. DualDn (4) performs denoising in both the RAW and sRGB domains and incorporates a differentiable ISP for joint training. Noise2VST (19) learns a model-free variance-stabilizing transform (VST) from a noisy image and a Gaussian denoiser, enabling Gaussian-trained networks to denoise real images without retraining. Lee *et al.* (22) proposed a unified demosaicing (DM) model for Bayer and non-Bayer color filter arrays (CFAs) using lightweight CFA-adaptive filters and an inference-time meta-learning variant.

**Joint Solutions.** Kokkinos *et al.* (18) proposed a novel joint denoising and demosaicing (JDD) method that unifies classical image regularization, large-scale optimization, and deep learning into a principled framework. SGNet (7) self-guides JDD by first estimating green channels and adding a density-map for frequency-aware restoration. Xing *et al.* (8) proposed an end-to-end CNN for joint demosaicing, denoising, and super-resolution (JDDSR). TENet++ (10) studies the effects of pipelines on the mixture problem of learning-based DN, DM, and SR, in both sequential and joint solutions. Dong *et al.* (20) proposed an iterative JDD framework for Bayer-pattern spike cameras with temporal alignment, noise estimation, and color prior.

## 3 The Proposed Method

An overview of the proposed JDD-PRO method is shown in Fig. 3. We describe the method from three aspects: Sec. 3.2 presents the controllable JDD training scheme, Sec. 3.3 introduces the HVS-inspired perceptual adapter, and Sec. 3.4 details the efficient network design.

### 3.1 Preliminaries

**Differentiable bidirectional ISP.** Following (1; 4), we build a differentiable and parameterized bidirectional ISP for training data synthesis and end-to-end learning, in which major camera-dependent operations are parameterized using camera metadata collected from a large number of devices. More details are provided in the appendix.

**Degradation Simulation.** We inject simulated noise and blur in the reverse ISP to synthesize degraded RAW images. Following (1; 23; 24), we model shot noise and read noise using Poisson and Gaussian distributions (25), respectively. The degraded RAW $\mathbf{R}$ is synthesized as

$$\mathbf{R} \sim K \cdot \mathcal{P}\left(\frac{\mathbf{R}^*}{K}\right) + \mathcal{N}(0, \sigma_{\mathtt{r}}^2), \tag{1}$$

where $\mathcal{P}(\cdot)$ and $\mathcal{N}(\cdot)$ denote the Poisson and Gaussian distributions, respectively. $\mathbf{R}^*$ is the clean RAW, $K$ is the total gain and $\sigma_{\mathtt{r}}^2$ denotes the variance of Gaussian noise. $K$ and $\sigma_{\mathtt{r}}^2$ are sampled from the following distributions:

$$\begin{aligned} \log(K) &\sim \mathcal{U}\big(\log(K_{\mathtt{min}}), \log(K_{\mathtt{max}})\big), \\ \log(\sigma_{\mathtt{r}}^2) \mid \log(K) &\sim \mathcal{N}\big(\alpha \cdot \log(K) + \beta, \sigma^2\big). \end{aligned} \tag{2}$$

We set $K_{\mathtt{min}}$ and $K_{\mathtt{max}}$ to $0.0001$ and $0.04$, respectively, and choose $\alpha \in [1.0, 3.0]$, $\beta \in [-0.5, 1.6]$, and $\sigma = 0.3$ to simulate a wide range of noise levels.

To simulate lens blur, we convolve the intermediate linear RGB image with an isotropic Gaussian kernel (26) during the reverse ISP, with kernel size 5 and variance 0.5.

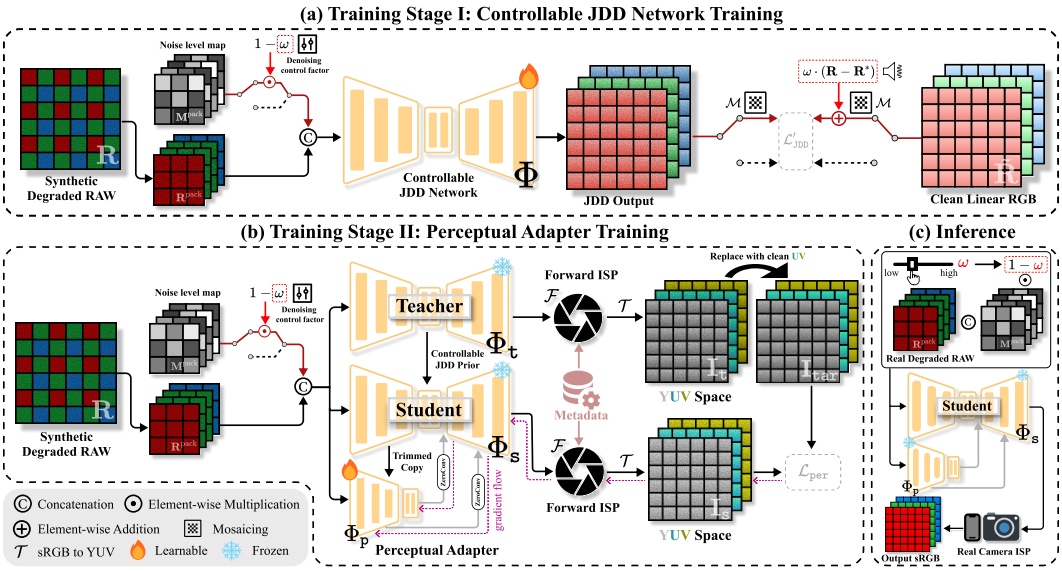



Figure 3: The pipeline of the proposed JDD-PRO. In the first stage, we develop a controllable JDD training scheme to learn a controllable JDD network. In the second stage, we train an HVS-inspired perceptual adapter to suppress chromatic noise and enhance perceptual quality. At inference, the denoising strength can be flexibly adjusted via the control factor $\omega$.

### 3.2 Controllable JDD Training Scheme

Building upon the aforementioned bidirectional ISP and degradation modeling, a clean sRGB image $\mathbf{I}^*$ can be transformed into a degraded RAW $\mathbf{R}$ by applying the reverse ISP flow combined with simulated noise and blur.

According to (12), the noise level map serves as a control mechanism for denoising strength. As the Poisson distribution asymptotically approaches the Gaussian distribution (25), Eq. (1) can be approximated as

$$\mathbf{R} \sim \mathcal{N}(\mathbf{R}^*, K \cdot \mathbf{R}^* + \sigma_{\mathtt{r}}^2). \tag{3}$$

Thus, the noise level map $\mathbf{M}$ can be approximated by the standard deviation of $\mathbf{R}$:

$$\mathbf{M} = \sqrt{K \cdot \mathbf{R}^* + \sigma_{\mathtt{r}}^2}. \tag{4}$$

Since $\mathbf{R}^*$ is not available in practical applications, we use $\mathbf{R}$ as a substitute to estimate $\mathbf{M}$. Then, we apply pixel unshuffle to pack the noisy RAW $\mathbf{R}$ and noise level map $\mathbf{M}$ into four-channel $\mathbf{R}^{\mathtt{pack}}$ and $\mathbf{M}^{\mathtt{pack}}$, which are concatenated and fed into the RAW JDD network. Conventionally, the loss function used to optimize the RAW JDD network is defined as:

$$\mathcal{L}_{\mathtt{JDD}} = \mathcal{L}_{\mathtt{hln}}(\Phi(\mathbf{R}^{\mathtt{pack}}, \mathbf{M}^{\mathtt{pack}}), \tilde{\mathbf{R}}), \tag{5}$$

where $\mathcal{L}_{\mathtt{hln}}$ denotes the hard-log loss in (27), which applies a heavier penalty to highlight regions in the RAW domain where pixel values are generally low. It is defined as $\mathcal{L}_{\mathtt{hln}}(\mathbf{I}, \hat{\mathbf{I}}) = -\frac{1}{CHW}||\log(1 - |\mathbf{I} - \hat{\mathbf{I}}| + \epsilon)||_1$, where $\epsilon$ is a small constant and $||\cdot||_1$ denotes $\ell_1$ distance. $\Phi$ represents the RAW JDD network, and $\tilde{\mathbf{R}}$ is the intermediate clean linear RGB generated during the reverse ISP. However, this optimization paradigm, relying solely on noisy-clean image pairs, lacks the flexibility to regulate denoising strength. To address this limitation, we introduce a denoising control factor $\omega$ to model mismatches between noise level and denoising strength during training:

$$\mathcal{L}'_{\mathtt{JDD}} = \begin{cases} \mathcal{L}_{\mathtt{hln}}\Big(\mathcal{M}\big(\Phi(\mathbf{R}^{\mathtt{pack}}, (1-\omega)\cdot\mathbf{M}^{\mathtt{pack}})\big), \mathcal{M}(\tilde{\mathbf{R}}) + \omega \cdot (\mathbf{R} - \mathbf{R}^*)\Big), & s \geq \tau, \\ \mathcal{L}_{\mathtt{hln}}\Big(\Phi(\mathbf{R}^{\mathtt{pack}}, \mathbf{M}^{\mathtt{pack}}), \tilde{\mathbf{R}}\Big), & s < \tau. \end{cases} \tag{6}$$

Here, the threshold $\tau$ splits the supervision into two cases. For $s \geq \tau$, we explicitly model mismatches between noise level and denoising strength by injecting noise (*i.e.*, $\mathbf{R} - \mathbf{R}^*$) into the clean linear

RGB $\tilde{\mathbf{R}}$ to construct the training target, while sampling $\omega$ to endow the model with controllable denoising strength. As $\mathbf{R}-\mathbf{R}^*$ is Bayer-sampled whereas $\tilde{\mathbf{R}}$ is three-channel, we apply the mosaicing operator $\mathcal{M}$ to project the supervision onto Bayer sampling sites, thereby avoiding demosaicing-induced interpolation bias and color artifacts in the training target and yielding physically consistent supervision. Since supervision in the Bayer domain lacks demosaicing guidance, for $s < \tau$, we resort to the conventional paired clean–degraded supervision scheme to explicitly impart demosaicing knowledge to the model. During training, $\omega \sim \mathcal{U}(0, 0.6)$ and $s \sim \mathcal{U}(0, 1)$, with $\tau$ fixed at 0.5. Notably, exposure to a wide range of noise level maps (*i.e.*, $(1-\omega)\cdot\mathbf{M}^{\texttt{pack}}$) during training enables the model to generalize reasonably well to control values outside the training range $[0, 0.6]$, allowing the denoising strength to be adjusted beyond this range at inference time. A smaller $\omega$ yields smoother results, while a larger $\omega$ preserves more noise.

### 3.3 HVS-inspired Perceptual Adapter

Under the above training scheme, the JDD network $\Phi$ gains controllable denoising capability, although this control is mainly limited to the suppression of color noise. As shown in Fig. 4, at the same noise level, the human visual system (HVS) (13) is more sensitive to color noise than to grayscale noise. This observation motivates us to suppress chromatic noise while preserving luminance noise for improved perceptual quality. Since the luminance–chrominance space (*i.e.*, YUV (28)) is defined on the standard sRGB domain, a straightforward solution is to employ an auxiliary network in the sRGB domain for chrominance denoising. However, such a two-stage design increases inference cost and complicates deployment. To this end, we propose a perceptual adapter for one-stage perceptual enhancement.

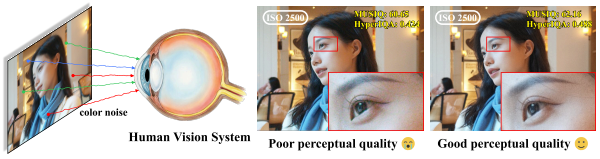



Figure 4: The human visual system is highly sensitive to color noise. At the same noise level, ISO 2500, preserving only luminance noise yields better perceptual quality than preserving color noise (zoom in for a better view).

Specifically, we freeze the controllable JDD network $\Phi$ trained in the first stage and use it as the teacher model $\Phi_{\texttt{t}}$. Given the noisy RAW $\mathbf{R}^{\texttt{pack}}$ and the noise-level map $\mathbf{M}^{\texttt{pack}}$, we scale the latter by $(1-\omega)$ and feed both into $\Phi_{\texttt{t}}$, following the conditioning scheme in Eq. (6). The output is then mapped to the sRGB space via the forward ISP. We then convert it to YUV space and replace its UV channels with those of the clean sRGB image $\mathbf{I}^*$. The resulting image serves as the training target, which can be formulated as

$$\begin{aligned} \mathbf{I}_{\texttt{t}} &= [\mathbf{Y}_{\texttt{t}}, \mathbf{U}_{\texttt{t}}, \mathbf{V}_{\texttt{t}}] = \mathcal{T}\circ\mathcal{F}\big(\Phi_{\texttt{t}}\big(\mathbf{R}^{\texttt{pack}}, (1-\omega)\cdot\mathbf{M}^{\texttt{pack}}\big)\big)\,, \\ \mathbf{I}_{\texttt{tar}} &= [\mathbf{Y}_{\texttt{t}}, \mathbf{U}^*, \mathbf{V}^*], \quad [\mathbf{Y}^*, \mathbf{U}^*, \mathbf{V}^*] = \mathcal{T}(\mathbf{I}^*), \end{aligned} \tag{7}$$

where $\mathcal{T}$ denotes the sRGB-to-YUV transformation, $\mathcal{F}$ denotes the forward ISP, and $\mathbf{I}_{\texttt{tar}}$ is the training target.

Then, instead of training a student network from scratch, we initialize the student model $\Phi_{\texttt{s}}$ with the teacher weights and keep it frozen. This allows the student to perform perceptual enhancement directly on top of the teacher model, rather than being constrained by its performance ceiling. Inspired by ControlNet (29), we propose a perceptual adapter $\Phi_{\texttt{p}}$, implemented as a lightweight trimmed copy (30) of the encoder and middle block, to modulate the base student model. This design preserves the controllable JDD prior in the student model while enabling output modulation with low training cost. Given the same input as the teacher model, the output of the student model is

$$\begin{aligned} \mathbf{I}_{\texttt{s}} &= [\mathbf{Y}_{\texttt{s}}, \mathbf{U}_{\texttt{s}}, \mathbf{V}_{\texttt{s}}] = \mathcal{T}\circ\mathcal{F}\big(\Phi_{\texttt{s}}\big(\mathbf{R}^{\texttt{pack}}, (1-\omega)\cdot\mathbf{M}^{\texttt{pack}}, \mathbf{F}_{\texttt{p}}\big)\big)\,, \\ \mathbf{F}_{\texttt{p}} &= \Phi_{\texttt{p}}\big(\mathbf{R}^{\texttt{pack}}, (1-\omega)\cdot\mathbf{M}^{\texttt{pack}}\big)\,, \end{aligned} \tag{8}$$

where $\mathbf{F}_{\texttt{p}}$ is the control signal provided by $\Phi_{\texttt{p}}$. Thus, the final loss function is defined as

$$\mathcal{L}_{\texttt{per}} = ||\mathbf{I}_{\texttt{tar}} - \mathbf{I}_{\texttt{s}}||_1. \tag{9}$$

### 3.4 Network Architecture

Our RAW JDD network is built upon NAFNet (31), a strong restoration backbone (32). As shown in Fig. 5(a), the encoder incorporates Degradation-aware Convolution (DAConv) to inject noise priors,




Figure 6: Qualitative comparison of denoising performance under varying settings of the denoising control factor $\omega$.

which are extracted from the noise level map. Here, we present only the encoder block. The middle and decoder blocks follow a similar design. For efficiency, the entire network is mainly implemented with point-wise and depth-wise convolutions. More architectural details are provided in the appendix.

**Degradation-aware Convolution.** Unlike additive Gaussian noise, signal-dependent shot and read noise in the RAW domain exhibits strong spatial variability due to abrupt signal changes. Also, demosaicing in texture-rich regions is more challenging than in flat areas and often introduces color and zipper artifacts. The use of content-agnostic kernels in existing CNN-based restoration methods (33; 17) constrains their performance. Inspired by (34), we propose a DAConv module. Fig. 5(b) illustrates the detailed architecture of DAConv. Given the degraded feature $\mathbf{X} \in \mathbb{R}^{H\times W\times C_1}$ and noise level prior $\mathbf{X}^{\mathtt{np}} \in \mathbb{R}^{H\times W\times C_2}$, we first fuse the features by concatenation followed by a series of point-wise convolutions and large-kernel depth-wise convolutions (7×7 `DConv`) to capture large-field contextual spatial information and generate spatially adaptive weights $\mathbf{W} \in \mathbb{R}^{H\times W\times D}$. The entire process can be formalized as:

$$\mathbf{W} = \mathtt{PConv}^2\left(\mathtt{DConv7}\left(\mathtt{PConv}\left(\mathtt{Concat}(\mathbf{X}, \mathbf{X}^{\mathtt{np}})\right)\right)\right), \tag{10}$$

where `Concat` represents the concatenation and $w_{\mathtt{i}} \in \mathbb{R}^D$ is the weight vector generated for $x_{\mathtt{i}}$. To reduce the computational cost, we divide $x_{\mathtt{i}}$ into $G$ groups along the channel dimension, with each group containing $C_1/G$ channels. We set $D = 9G$ and reshape $w_{\mathtt{i}}$ into $w^*_{\mathtt{i}} \in \mathbb{R}^{G\times 3\times 3}$. Subsequently, a dynamic convolution between $\mathbf{W}^*$ and $\mathbf{X}$ is performed to obtain $\mathbf{Y} \in \mathbb{R}^{H\times W\times C_1}$:

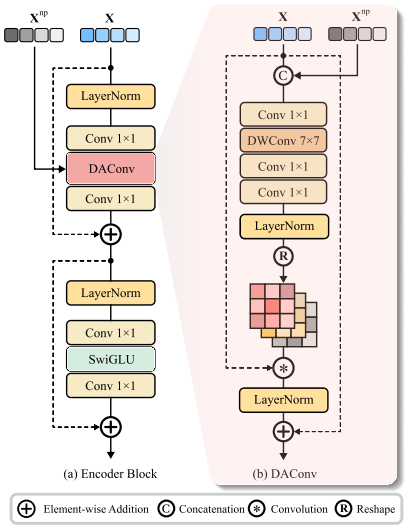



Figure 5: Illustration of the RAW JDD network architecture.

$$y_{\mathtt{ic}} = w_{\mathtt{ig}} \circledast \mathcal{E}_{3\times 3}(x_{\mathtt{ic}}), \tag{11}$$

where $\mathcal{E}_{3\times 3}(x_{\mathtt{ic}})$ denotes the 3×3 neighborhood centered at $x_{\mathtt{ic}}$ and $\circledast$ denotes the convolution operation. Finally, a residual connection is applied to obtain the output $\hat{\mathbf{X}} = \mathbf{X} \oplus \mathbf{Y}$, where $\oplus$ represents element-wise addition.

# 4 Experiments

## 4.1 Implementation Details

We conduct training on high-quality images from the Waterloo Exploration Database (36), DIV2K (37), and Flickr2K (38). The test set comprises (i) smartphone captures from OPPO Find X7 Ultra, iPhone 13 Pro, Xiaomi 12, and Huawei Mate 30, and (ii) the DND (39) benchmark. During training, the patch size and batch size are set to 256×256 and 48, respectively. We use the Adam optimizer with an initial learning rate of $1 \times 10^{-4}$, which is reduced by half at iterations 100K, 150K, 250K, and 350K for both training stages. The framework is implemented in PyTorch and trained on NVIDIA RTX 4090 GPUs.

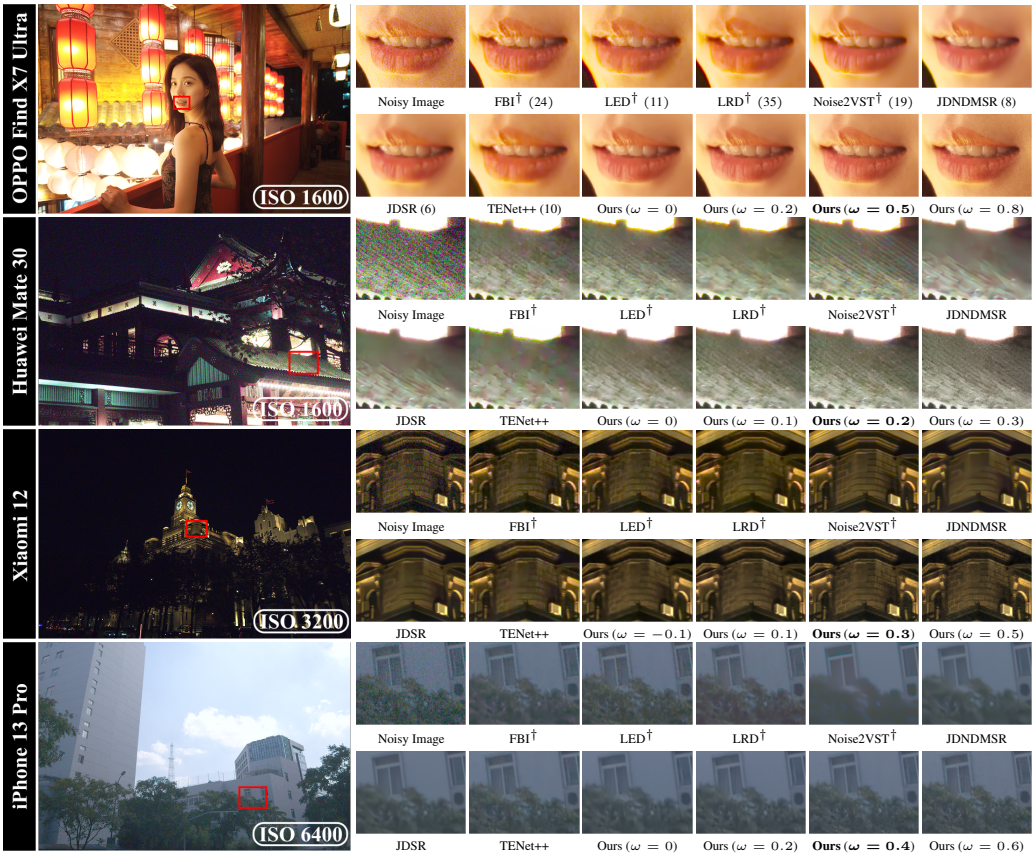



Figure 7: Qualitative comparison across four smartphone brands. † indicates the AHD (5) demosaicing algorithm is applied.

### 4.2 Discussion on the Denoising Control Factor

To evaluate the influence of the denoising control factor $\omega$ on noise suppression and its effect on the smearing effect, we conduct experiments on two smartphone photos from OPPO Find X7 Ultra. Fig. 6 illustrates the denoising results under varying values of $\omega$. In the portrait scenario (first row), an aggressive denoising setting (*e.g.*, $\omega = 0$) results in excessive smoothing of facial features, such as wrinkles around the eyes and skin textures. As $\omega$ increases to $0.5$, fine-grained textures are better preserved, yielding visually more realistic results. Conversely, when $\omega$ is set too high (*e.g.*, $0.8$), insufficient denoising leaves residual noise, which adversely affects the perceptual quality. In the outdoor scenario, this trend is more evident on texture-rich regions (*e.g.*, trees in the second row). Low $\omega$ results in severe smearing, with blended pixels impairing perceptual clarity. As $\omega$ increases, leaf structures and interstitial details become progressively more discernible.

### 4.3 Evaluation on Smartphone-Captured Images

We benchmark against four SOTA RAW denoising methods (FBI (24), LED (11), LRD (35), NoiseVST (19)) and three joint methods (JDNDMSR (8), JDSR (6), TENet++ (10)). JDNDMSR, JDSR, and TENet++ are retrained on the Pixel-Shift200 (10) dataset to meet the JDD task. For fairness, the denoising-only methods are paired with the classical AHD (5) demosaicing algorithm.

**Qualitative Comparisons.** Fig. 7 presents qualitative comparisons on images captured by various smartphone brands. In the portrait scenario (first scene), JDD-PRO ($\omega = 0.5$) achieves the most faithful texture reconstruction on the lips. In low-light conditions (second and third scenes), where severe noise and blur obscure the signal, competing methods exhibit either insufficient denoising or pronounced smearing, while JDD-PRO flexibly balances texture preservation and noise suppression

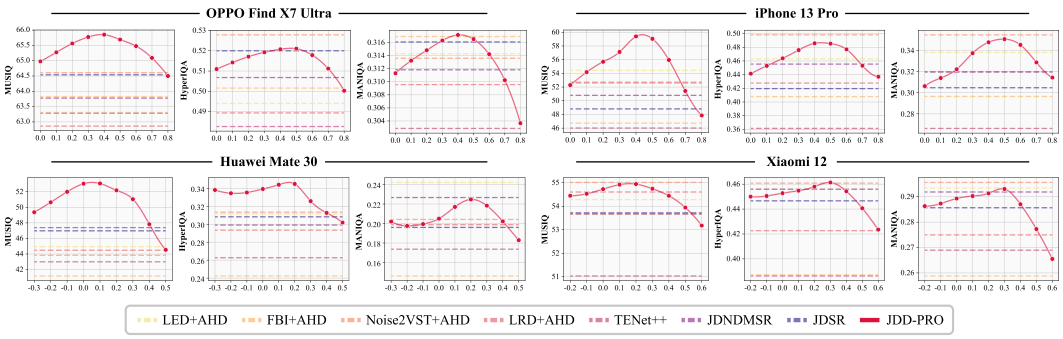



Figure 8: Quantitative comparison across four smartphone brands. The horizontal axis represents the value of $\omega$.

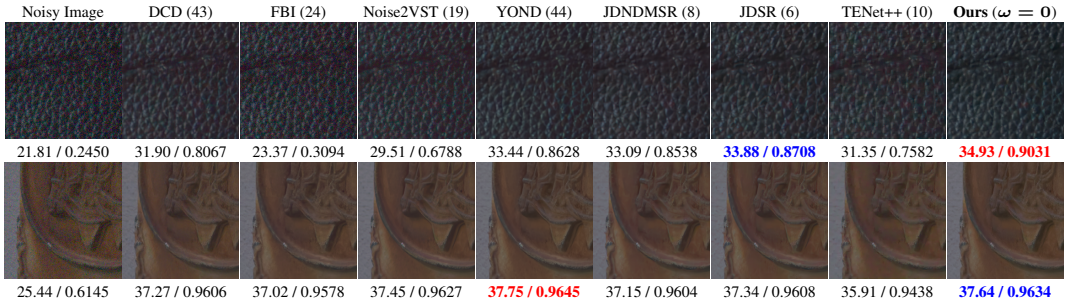



Figure 9: Qualitative comparison on the DND (39) benchmark.

by tuning $\omega$. On iPhone 13 Pro, competing methods exhibit over-smoothing and color casts, while ours restores the richest textures and faithful colors.

Moreover, we observe that different images achieve their best trade-off at different values of $\omega$, which is intuitive. Since $\omega$ controls the percentage of noise retained rather than its absolute magnitude, smaller $\omega$ is preferred under heavy noise, whereas larger $\omega$ works better when the noise is mild.

**Quantitative Comparisons.** Since there is no ground truth for smartphone-captured images, we assess restoration quality on images from four smartphone brands using MUSIQ (40), HyperIQA (41), and MANIQA (42). As illustrated in Fig. 8, adaptive tuning of $\omega$ enables our method to achieve consistently competitive performance across all four smartphone datasets.

### 4.4 Evaluation on DND Benchmark

We compare our method on the DND benchmark with four SOTA RAW denoisers (DCD (43), FBI, Noise2VST, YOND (44)) and the three JDD methods. We found that linear demosaicing outperforms AHD on the DND benchmark, so we apply it to the denoising-only baselines. Fig. 9 shows qualitative comparisons: with $\omega = 0$ (the best-performing setting), our method surpasses the second-best by 1.05 dB in PSNR and 0.0323 in SSIM on the first scene, and is slightly inferior to YOND on the second. Quantitative results are reported in Table 1, where our method achieves the best overall performance.

Table 1: Quantitative comparison on the DND (39) benchmark.

| | | RAW | | sRGB | | | |
|---|---|---|---|---|---|---|---|
| | **Method** | PSNR | SSIM | PSNR | SSIM | HyperIQA | MANIQA |
| DN | DCD (43) | **48.244** | **0.9798** | 39.096 | 0.9470 | 0.3305 | 0.2683 |
| DN | FBI (24) | 47.505 | 0.9704 | 38.568 | 0.9181 | 0.3359 | 0.2544 |
| DN | Noise2VST (19) | 47.944 | 0.9783 | 39.009 | 0.9442 | 0.3520 | 0.2689 |
| DN | YOND (44) | **48.568** | **0.9814** | 39.694 | 0.9553 | **0.3682** | **0.2864** |
| JDD | JDNDMSR (8) | – | – | 37.517 | 0.9322 | 0.3272 | 0.2729 |
| JDD | JDSR (6) | – | – | 39.574 | 0.9527 | 0.3525 | 0.2693 |
| JDD | TENet++ (10) | – | – | 39.733 | 0.9559 | 0.3676 | 0.2447 |
| JDD | Ours ($\omega = -0.1$) | – | – | **39.762** | **0.9584** | 0.3528 | 0.2862 |
| JDD | Ours ($\omega = 0.0$) | – | – | **39.986** | **0.9606** | 0.3642 | **0.2865** |
| JDD | Ours ($\omega = 0.1$) | – | – | 39.734 | 0.9583 | **0.3763** | 0.2757 |

### 4.5 Ablation Studies

We conduct ablation studies on the Xiaomi 12 dataset to evaluate the CTRL-JDD training scheme, perceptual adapter, and DAConv. The analysis is performed under the setting $\omega = 0.3$, which

Table 2: Quantitative results of ablation studies on the CTRL-JDD training scheme, perceptual adapter, and DAConv.

| Variant | MUSIQ | HyperIQA | MANIQA | PaQ2PiQ |
|---|---|---|---|---|
| *CTRL-JDD Ablation* | | | | |
| (a) *w/o* $\mathcal{M}$ in Eq. 6 | 54.641 | 0.4439 | 0.2911 | **70.404** |
| (b) *w/o* $s < \tau$ in Eq. 6 | 53.770 | 0.4396 | 0.2880 | 70.126 |
| *Perceptual Adapter Ablation* | | | | |
| (c) *w/o* $\Phi_{\text{p}}$ | 53.674 | 0.4474 | 0.2917 | 70.168 |
| (d) *w/o* CTRL-JDD Prior | 54.111 | 0.4481 | **0.2960** | 70.182 |
| (e) *w/* WB&CCM | **54.799** | **0.4628** | 0.2938 | 70.275 |
| *DAConv Ablation* | | | | |
| (f) *w/o* DAConv | 53.356 | 0.4256 | 0.2834 | 70.013 |
| **Ours** | **54.833** | **0.4641** | **0.2943** | **70.282** |

Table 3: Comparison of parameters (M), and of runtime (s) and MACs (G) at different image resolutions.

| | Params. | $1024^2$ | | $2048^2$ | | $4096^2$ | |
|---|---|---|---|---|---|---|---|
| | | Time | MACs | Time | MACs | Time | MACs |
| FBI (24) | **2.488** | 0.395 | 1.22K | 1.598 | 4.90K | 71.03 | 19.6K |
| LED (11) | 7.760 | 0.235 | **54.82** | 1.021 | **219.3** | **4.307** | **877.2** |
| Noise2VST (19) | 33.49 | 24.17 | – | 27.15 | – | 28.04 | – |
| JDNDMSR (8) | **2.528** | **0.080** | 671.8 | 2.182 | 2.68K | 8.660 | 10.74K |
| JDSR (6) | 14.33 | 0.225 | 3.87K | **0.664** | 15.5K | Out of Memory | |
| TENet++ (10) | 20.07 | 0.553 | 5.26K | 3.265 | 21.0K | 34.57 | 84.1K |
| **Ours** | 6.569 | **0.046** | **75.88** | **0.235** | **303.5** | **1.084** | **1.214K** |

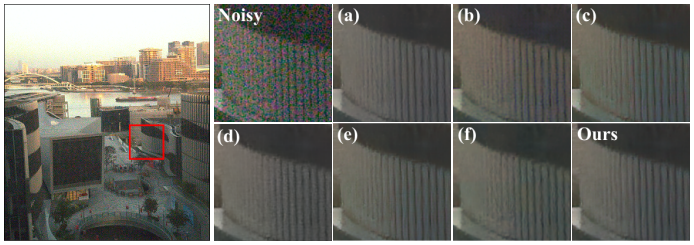



Figure 10: Qualitative results of ablation studies on the CTRL-JDD training scheme, perceptual adapter, and DAConv.

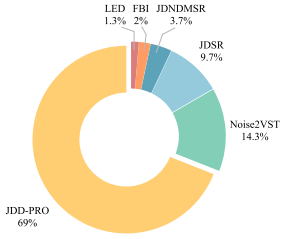



Figure 11: Illustration of user study results.

corresponds to the best observed performance. More visual ablation comparisons under multiple $\omega$ settings are provided in the appendix.

**CTRL-JDD.** The controllable JDD training scheme is the cornerstone of our method. To investigate its impact, we design two variants: (i) removing the mosaicing operation in Eq. 6, *i.e.*, applying all supervision in the three-channel linear RGB domain; and (ii) removing the case of $s < \tau$, *i.e.*, applying all supervision in the Bayer domain. The former constructs training targets by adding Bayer-sampled noise, processed through a demosaicing algorithm (*e.g.*, Malvar (45)), to clean linear RGB images. This forces the model to learn undesirable demosaicing priors, resulting in unsmooth edges (see Fig. 10(a)). The latter lacks explicit demosaicing guidance, causing the model to generate inaccurate colors in missing-pixel regions, with Fig. 10(b) showing obvious color shifts. For these reasons, both variants lead to inferior performance, as reported in Table 2.

**Perceptual Adapter.** We design three variants: (i) *w/o* $\Phi_{\text{p}}$: removing the perceptual adapter; (ii) *w/o* CTRL-JDD Prior: training the student model from scratch; and (iii) *w/* WB&CCM: providing $\Phi_{\text{p}}$ with additional white-balance (WB) (1) and color correction matrix (CCM) (1) conditions. Without $\Phi_{\text{p}}$, the model tends to retain perceptually unpleasant chromatic noise (see Fig. 10(c)). Without the CTRL-JDD prior, the student is trained only with sRGB-domain losses, which are insufficient for the JDD task, leading to performance degradation (see Fig. 10(d)). Since WB and CCM are closely related to ISP color transformations, incorporating them as additional conditions would intuitively benefit noise color modulation. However, we find that such conditioning brings no performance gain. We attribute this to the limited capacity of the perceptual adapter, for which the abstract WB and CCM parameters introduce extra learning difficulty. Moreover, blind modulation is more practical for deployment. Quantitative results are shown in Table 2.

**DAConv.** We replace DAConv with the simple channel attention module in (31), while keeping the parameter count nearly unchanged. Table 2 shows that DAConv brings a clear performance gain. As illustrated in Fig. 10(f), the variant without DAConv suffers from severe degradation.

### 4.6 Computational Efficiency

We compare model size, inference time, and MACs across different image resolutions against state-of-the-art methods. As reported in Table 3, our proposed method achieves the fastest inference and

the second-lowest computational cost. Moreover, as resolution increases, most competing methods incur substantial growth in runtime and compute (*e.g.*, JDSR runs out of memory at a resolution of $4096 \times 4096$, and FBI requires 71.03 seconds per image), whereas our method keeps both runtime and computation within a reasonable range, highlighting its practical relevance.

### 4.7 User Study

Beyond quantitative and qualitative evaluations, we conduct a user study to assess the practical utility of our method, as shown in Fig. 11. We recruit 30 participants from universities, including researchers working on low-level vision and photography enthusiasts. Participants evaluate 10 images randomly selected from the four smartphone datasets used in this paper. For each image, participants are shown the ISP-processed noisy image, the JDD results produced by competing methods, and our results under four different $\omega$ settings. For each image, each participant casts exactly one vote for the method whose result best meets their expectations. The preference score of each method is computed as the proportion of votes it receives across all participant–image evaluations. Our method achieves the highest preference score of 69%, outperforming all competing methods.

## 5 Conclusion and Limitations

In this paper, we presented JDD-PRO, a perception-aware and controllable framework for joint RAW image denoising and demosaicing. By introducing a controllable JDD training scheme together with an HVS-inspired perceptual adapter, JDD-PRO effectively balanced noise suppression and detail preservation under diverse restoration requirements. Extensive experiments on both synthetic and real-world datasets demonstrated that the proposed method achieved superior perceptual quality and strong robustness compared with existing state-of-the-art approaches. Despite these promising results, JDD-PRO still has limitations. In particular, the control parameter $\omega$ currently requires manual adjustment to accommodate different user preferences and application scenarios. Developing an adaptive or content-aware control strategy that automatically determines suitable perceptual restoration levels remains an important direction for future work.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62572234), the Gusu Innovation and Entrepreneurship Leading Talent Program (Grant No. ZXL20254324), the Suzhou Key Technologies Project (Grant No. SGY2023136), and OPPO Research Fund.

## References

[1] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, and J. T. Barron, "Unprocessing images for learned raw denoising," in *CVPR*, 2019, pp. 11 036–11 045.

[2] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Cycleisp: Real image restoration via improved data synthesis," in *CVPR*, 2020, pp. 2696–2705.

[3] S. Xu, Z. Sun, J. Zhu, Y. Zhu, X. Fu, and Z.-J. Zha, "Demosaicformer: Coarse-to-fine demosaicing network for hybridevs camera," in *CVPRW*, 2024, pp. 1126–1135.

[4] R. Li, Y. Wang, S. Chen, F. Zhang, J. Gu, and T. Xue, "Dualdn: Dual-domain denoising via differentiable isp," in *ECCV*. Springer, 2024, pp. 160–177.

[5] K. Hirakawa and T. W. Parks, "Adaptive homogeneity-directed demosaicing algorithm," *IEEE TIP*, vol. 14, no. 3, pp. 360–369, 2005.

[6] R. Zhou, R. Achanta, and S. Süsstrunk, "Deep residual network for joint demosaicing and super-resolution," *Color and Imaging Conference*, vol. 26, no. 1, pp. 75–75, 2018.

[7] L. Liu, X. Jia, J. Liu, and Q. Tian, "Joint demosaicing and denoising with self guidance," in *CVPR*, 2020, pp. 2240–2249.

[8] W. Xing and K. Egiazarian, "End-to-end learning for joint image demosaicing, denoising and super-resolution," in *CVPR*, 2021, pp. 3507–3516.

[9] B. Zheng, H. Li, Q. Chen, T. Wang, X. Zhou, Z. Hu, and C. Yan, "Quad bayer joint demosaicing and denoising based on dual encoder network with joint residual learning," in *AAAI*, vol. 38, no. 7, 2024, pp. 7552–7561.

[10] G. Qian, Y. Wang, J. Gu, C. Dong, W. Heidrich, B. Ghanem, and J. S. Ren, "Rethinking learning-based demosaicing, denoising, and super-resolution pipeline," in *ICCP*. IEEE, 2022, pp. 1–12.

[11] X. Jin, J.-W. Xiao, L.-H. Han, C. Guo, R. Zhang, X. Liu, and C. Li, "Lighting every darkness in two pairs: A calibration-free pipeline for raw denoising," in *ICCV*, 2023, pp. 13 275–13 284.

[12] K. Zhang, W. Zuo, and L. Zhang, "Ffdnet: Toward a fast and flexible solution for cnn-based image denoising," *IEEE TIP*, vol. 27, no. 9, pp. 4608–4622, 2018.

[13] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on hvs," in *Proceedings of the second international workshop on video processing and quality metrics*, vol. 4. IST/SPIE, 2006, pp. 1–6.

[14] X. Song, G. M. Johnson, and M. D. Fairchild, "Minimizing the perception of chromatic noise in digital images," in *Color and Imaging Conference*, vol. 12. Society of Imaging Science and Technology, 2004, pp. 340–346.

[15] T. Seybold, Ö. Cakmak, C. Keimel, and W. Stechele, "Noise characteristics of a single sensor camera in digital color image processing," in *Color and imaging conference*, vol. 22. Society for Imaging Science and Technology, 2014, pp. 53–58.

[16] F. Luisier, T. Blu, and M. Unser, "Image denoising in mixed poisson–gaussian noise," *IEEE TIP*, vol. 20, no. 3, pp. 696–708, 2010.

[17] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *CVPR*, 2018, pp. 3291–3300.

[18] F. Kokkinos and S. Lefkimmiatis, "Iterative joint image demosaicking and denoising using a residual denoising network," *IEEE TIP*, vol. 28, no. 8, pp. 4177–4188, 2019.

[19] S. Herbreteau and M. Unser, "Self-calibrated variance-stabilizing transformations for real-world image denoising," in *ICCV*, 2025, pp. 10 496–10 506.

[20] Y. Dong, R. Xiong, J. Zhao, J. Zhang, X. Fan, S. Zhu, and T. Huang, "Joint demosaicing and denoising for spike camera," in *AAAI*, vol. 38, no. 2, 2024, pp. 1582–1590.

[21] S. Guo, Z. Liang, and L. Zhang, "Joint denoising and demosaicking with green channel prior for real-world burst images," *IEEE TIP*, vol. 30, pp. 6930–6942, 2021.

[22] H. Lee, D. Park, W. Jeong, K. Kim, H. Je, D. Ryu, and S. Y. Chun, "Efficient unified demosaicing for bayer and non-bayer patterned image sensors," in *ICCV*, 2023, pp. 12 750–12 759.

[23] M. V. Conde, F. Vasluianu, and R. Timofte, "Toward efficient deep blind raw image restoration," in *ICIP*. IEEE, 2024, pp. 1725–1731.

[24] J. Byun, S. Cha, and T. Moon, "Fbi-denoiser: Fast blind image denoiser for poisson-gaussian noise," in *CVPR*, 2021, pp. 5768–5777.

[25] A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian, "Practical poissonian-gaussian noise modeling and fitting for single-image raw-data," *IEEE TIP*, vol. 17, no. 10, pp. 1737–1754, 2008.

[26] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *ICCV*, 2021, pp. 4791–4800.

[27] R. Berdan, B. Besbinar, C. Reinders, J. Otsuka, and D. Iso, "Reraw: Rgb-to-raw image reconstruction via stratified sampling for efficient object detection on the edge," in *CVPR*, 2025, pp. 11 833–11 843.

[28] M. A. Ansari and D. K. Singh, "Significance of color spaces and their selection for image processing: a survey," *Recent Advances in Computer Science and Communications*, vol. 15, no. 7, pp. 946–956, 2022.

[29] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023, pp. 3836–3847.

[30] F. Yu, J. Gu, Z. Li, J. Hu, X. Kong, X. Wang, J. He, Y. Qiao, and C. Dong, "Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild," in *CVPR*, 2024, pp. 25 669–25 680.

[31] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *ECCV*. Springer, 2022, pp. 17–33.

[32] M. Conde, R. Timofte, Z. Lu, X. Kong, X. Xing, F. Wang, S. Han, M. Park, T. Hao, Y. He *et al.*, "Ntire 2025 challenge on raw image restoration and super-resolution," in *CVPR*, 2025, pp. 1148–1171.

[33] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *ECCV*. Springer, 2020, pp. 492–511.

[34] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, "Lsnet: See large, focus small," in *CVPR*, 2025, pp. 9718–9729.

[35] F. Zhang, B. Xu, Z. Li, X. Liu, Q. Lu, C. Gao, and N. Sang, "Towards general low-light raw noise synthesis and modeling," in *ICCV*, 2023, pp. 10 820–10 830.

[36] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE TIP*, vol. 26, no. 2, pp. 1004–1016, 2016.

[37] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *CVPRW*, 2017, pp. 126–135.

[38] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPRW*, 2017, pp. 136–144.

[39] T. Plotz and S. Roth, "Benchmarking denoising algorithms with real photographs," in *CVPR*, 2017, pp. 1586–1595.

[40] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *ICCV*, 2021, pp. 5148–5157.

[41] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *CVPR*, 2020, pp. 3667–3676.

[42] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "Maniqa: Multi-dimension attention network for no-reference image quality assessment," in *CVPR*, 2022, pp. 1191–1200.

[43] Y. Zou, C. Yan, and Y. Fu, "Iterative denoiser and noise estimator for self-supervised image denoising," in *ICCV*, 2023, pp. 13 265–13 274.

[44] H. Feng, L. Wang, Y. Huang, T. Li, L. Zhu, and H. Huang, "Yond: Practical blind raw image denoising free from camera-specific data dependency," *IEEE TPAMI*, 2026.

[45] H. S. Malvar, L.-w. He, and R. Cutler, "High-quality linear interpolation for demosaicing of bayer-patterned color images," in *ICASSP*, vol. 3. IEEE, 2004, pp. iii–485.

[46] S. Wolf, "Color correction matrix for digital still and video imaging systems," Institute for Telecommunication Sciences, Tech. Rep., 2003.

[47] M. D. Grossberg and S. K. Nayar, "What is the space of camera response functions?" in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2003, pp. II–602.

[48] Z. Zhang, Y. Jiang, W. Shao, X. Wang, P. Luo, K. Lin, and J. Gu, "Real-time controllable denoising for image and video," in *CVPR*, 2023, pp. 14 028–14 038.

[49] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *CVPR*. IEEE, 2018, pp. 1692–1700.

## Abstract

This appendix is organized into four parts. First, we introduce our differentiable bidirectional ISP, which supports data synthesis and end-to-end training. Second, we describe the architecture of the RAW JDD network in detail. Third, we present additional experiments, including more visualizations of ablation studies, comparisons between one-stage and two-stage architectures, user studies, demosaicing comparisons, and comparisons with controllable denoising methods. Finally, we provide additional qualitative comparisons with state-of-the-art methods.


## A Differentiable bidirectional ISP

Following (1; 4), we design a differentiable and parameterized bidirectional ISP for training data synthesis and end-to-end learning. As illustrated in Fig. 12, The pipeline consists of five stages with corresponding forward and approximate inverse operations.

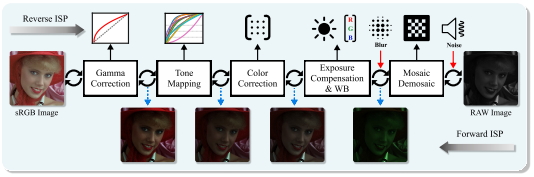



Figure 12: Detailed breakdown of differentiable bidirectional ISP.

**Mosaic & Demosaic.** For mosaicing, two of the three color channels at each pixel are discarded to form an R-G-G-B Bayer pattern. Demosaicing is omitted in the forward ISP, as it is performed by the RAW JDD network.

**Exposure Compensation & White Balance.** For the former, we employ a standard brightness modulation function

$$f(x, \lambda) = x \odot 2^{\lambda}, \quad f^{-1}(y, \lambda) = y \oslash 2^{\lambda}, \tag{12}$$

where $\odot$ and $\oslash$ denote element-wise multiplication and division, respectively, and we define $\lambda \in [-0.5, 5.0]$ to cover the full spectrum of lighting conditions. For the latter, we adopt empirically derived gain ranges from multiple cameras, with R gain $\in [0.9, 3.1]$ and B gain $\in [1.2, 4.2]$

**Color Correction.** For color correction which features a $3\times3$ Color Correction Matrix (CCM) (46), we collect 112 CCMs from camera metadata and perform random sampling and interpolation during training.

**Tone Mapping.** We parameterize nonlinear S-curve tone mapping with 201 representative curves derived from camera metadata (47), randomly sampled during training. To address the non-differentiability of discrete curve lookups, we design a uniform-grid linear interpolation algorithm to support differentiable tone mapping. The detailed algorithm implementation is presented in Algorithm 1. The $\delta$ in Algorithm 1 is set to $1e^{-6}$.

**Gamma Correction.** For gamma correction, we apply a standard power-law curve (1) to nonlinearly encode linear RGB values, approximately accounting for the nonlinear sensitivity of human vision.

$$\Gamma(x) = \max(x, \epsilon)^{1/2.2}, \quad \Gamma^{-1}(y) = \max(y, \epsilon)^{2.2}. \tag{13}$$

## B Details of the RAW JDD Network

As stated in Sec. 3.4, the JDD network adopts a NAFNet-style (31) architecture. Fig. 13 illustrates the detailed architecture of the JDD network. It consists of 4 encoder layers, 1 middle layer, and 4 decoder layers. Specifically, the four encoder layers contain [1, 2, 4, 4] encoder blocks respectively; the middle layer contains 4 middle blocks, and the four decoder layers each contain 1 decoder block. The DAConv$^*$ denotes the degradation-aware convolution without noise-prior guidance.

## C More Experiments

### C.1 More Visualizations of Ablation Studies

In Sec. 4.5, we only show ablation comparisons under a fixed $\omega$. Here, we present qualitative ablation comparisons under multiple $\omega$ settings to more comprehensively validate the effectiveness of each

**Algorithm 1:** Differentiable Bidirectional Tone Mapping via Uniform-Grid Linear Interpolation

**Input:** Control points $\{(\mathrm{ToneCurveX}_i, \mathrm{ToneCurveY}_i)\}$, step $\delta > 0$
**Output:** Operators $T : [0,1] \to [0,1]$ and $T^{-1} : [0,1] \to [0,1]$
**Precompute:**;
$K \leftarrow \lfloor 1/\delta \rfloor$; define uniform grid $\xi_k \leftarrow k\delta$, $k = 0, \ldots, K$;
Fit a cubic spline $f$ through the control points and an inverse spline $f^{-1}$;
Sample $y_k \leftarrow f(\xi_k)$ and $\tilde{y}_k \leftarrow f^{-1}(\xi_k)$ for $k = 0, \ldots, K$;

**Function** $T(x \in [0,1])$**:**
  $k \leftarrow \min(\lfloor x/\delta \rfloor, K-1)$;
  $\alpha \leftarrow x/\delta - k$;
  **return** $(1-\alpha)y_k + \alpha y_{k+1}$;

**Function** $T^{-1}(z \in [0,1])$**:**
  $k \leftarrow \min(\lfloor z/\delta \rfloor, K-1)$;
  $\beta \leftarrow z/\delta - k$;
  **return** $(1-\beta)\tilde{y}_k + \beta \tilde{y}_{k+1}$;

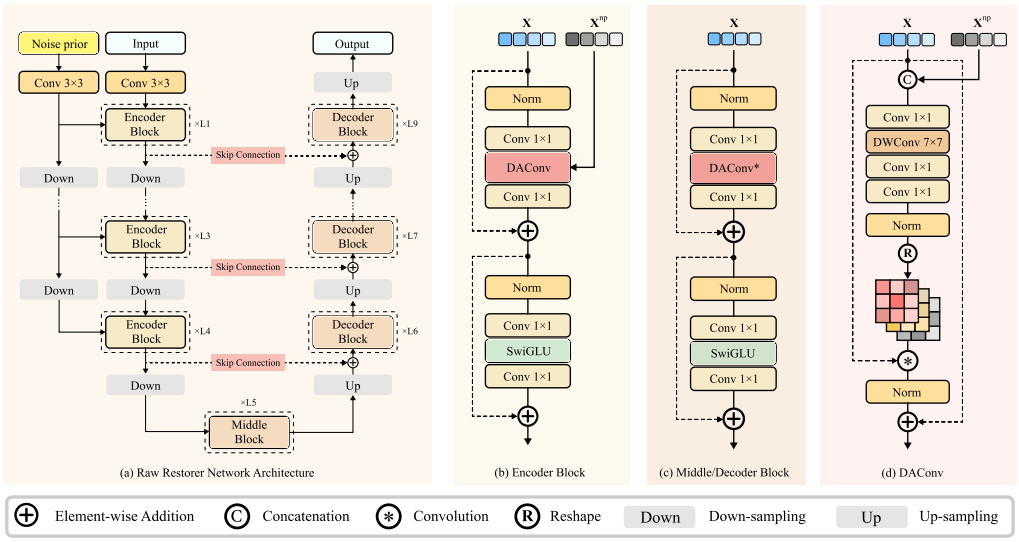



Figure 13: Detailed illustration of the overall architecture in the RAW JDD network.

component. The visualization results are shown in Fig. 14. Variants (a)–(f) correspond to *w/o* $\mathcal{M}$ in Eq. 6, *w/o* $s < \tau$ in Eq. 6, *w/o* $\Phi_{\mathrm{p}}$, *w/o* CTRL-JDD Prior, *w/* WB&CCM, and *w/o* DAConv, respectively. (a) fits the undesirable demosaicing artifacts in the ground truth, causing the retained noise to appear as scattered white speckles and lack realism. (b) lacks demosaicing supervision and therefore suffers from evident color casts. (c) preserves chrominance noise, leading to degraded perceptual quality. (d) is constrained by the teacher model and produces over-smoothed results under low-$\omega$ settings. (e) achieves performance comparable to Ours. (f) lacks spatially anisotropic degradation awareness, resulting in pronounced degradation and distortion.

### C.2 One-stage vs. Two-stage

We adopt a one-stage architecture to achieve perceptual-aware controllable JDD. Alternatively, the same functionality can also be realized by introducing an additional sRGB-domain network to remove chromatic noise in the sRGB domain. We conduct experiments on the DND dataset to compare the strengths and weaknesses of these two architectures. As shown in Fig. 15, benefiting from the additional network in the sRGB domain, the two-stage architecture removes demosaicing-induced color artifacts near edges more effectively, whereas our one-stage architecture exhibits more noticeable color artifacts, as indicated by the arrows in Fig. 15. As a result, the two-stage architecture

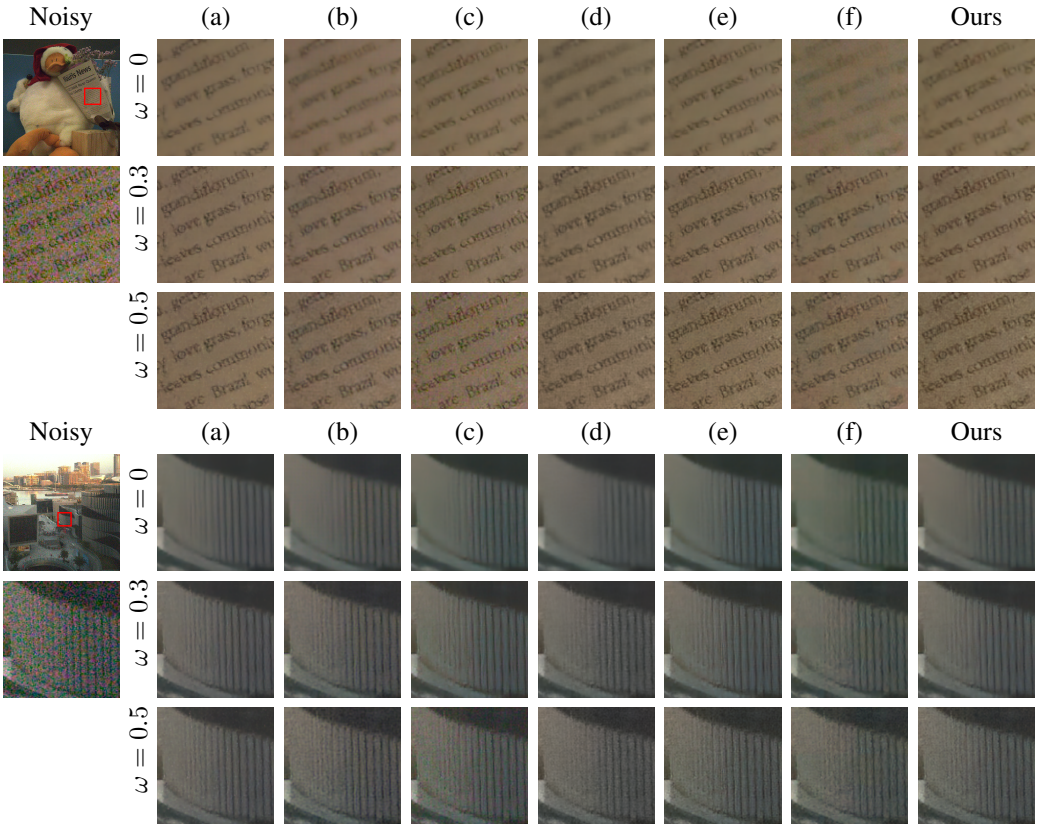



Figure 14: More qualitative comparisons of ablation studies (zoom in for a better view).

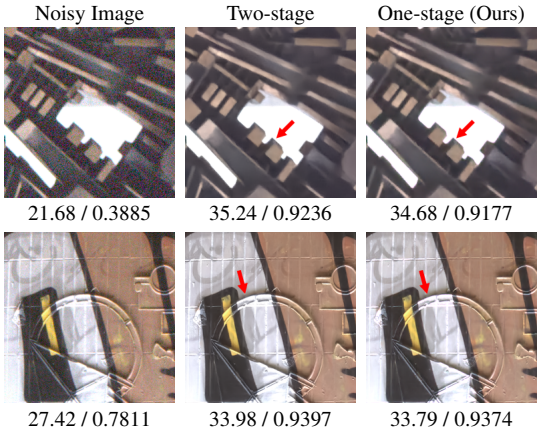



Figure 15: Qualitative comparison between two-stage and one-stage architectures. Zoom in for a better view.

Table 4: Quantitative comparison between two-stage and one-stage architectures. MACs and runtime are measured at a resolution of 2048 $\times$ 2048.

| **Architecture** | PSNR | SSIM | Params (M) | MACs (G) | Time (s) |
|---|---|---|---|---|---|
| Two-stage | **40.020** | **0.9610** | 6.664 | 537.3 | 0.370 |
| One-stage (Ours) | 39.986 | 0.9606 | **6.569** | **303.5** | **0.235** |

achieves better quantitative performance. Table 4 also shows that the two-stage architecture performs slightly better in terms of PSNR and SSIM. However, with a comparable number of parameters, the one-stage architecture is significantly superior to the two-stage architecture in inference speed and computational cost, making it better suited to practical deployment.

### C.3 Demosaicing Comparison

In this experiment, we focus on the demosaicing quality and provide visual comparisons with three joint denoising and demosaicing (JDD) methods. The qualitative comparison results are presented in Fig. 16. We unexpectedly find that our proposed perceptual adapter also improves the model's

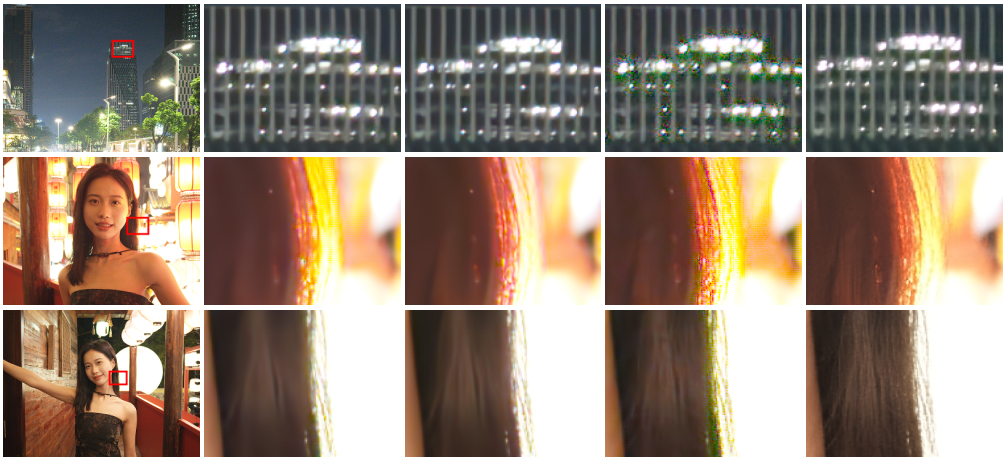



Figure 16: Qualitative comparison of demosaicing performance with three JDD methods.

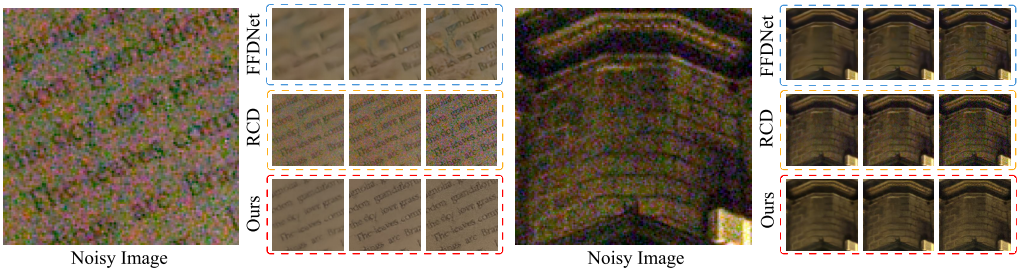



Figure 17: Qualitative comparison with controllable denoising methods.

demosaicing performance. As observed, compared with other JDD methods, our method produces fewer demosaicing-induced color artifacts.

### C.4 Comparison with Controllable Denoising Methods

Additionally, we qualitatively compare our method with two controllable denoising methods, FFDNet (12) and RCD (48), with results shown in Fig. 17. For our method, we show results at $\omega = 0.0, 0.1$, and $0.3$. For FFDNet, we do not directly use the AWGN-pretrained model. Since FFDNet operates in the sRGB domain, we transform the RAW-domain synthesized noisy images and noise-level maps into sRGB through the ISP and retrain FFDNet using its original degraded-clean training strategy. At test time, we also control its denoising strength using $\omega$, and present results at $\omega = 0.0, 0.1$, and $0.3$. RCD is trained on the real-noise SIDD (49) dataset, so we directly use its pretrained weights. As RCD employs a different control mechanism without $\omega$, we show three denoising levels, denoted as strong, middle, and mild. Our method demonstrates markedly superior controllable denoising performance.

## D More Visualization

We provide more visual comparisons to verify the proposed method's superiority against other state-of-the-art methods on four smartphone datasets. The visual comparisons are presented in Fig. 18.

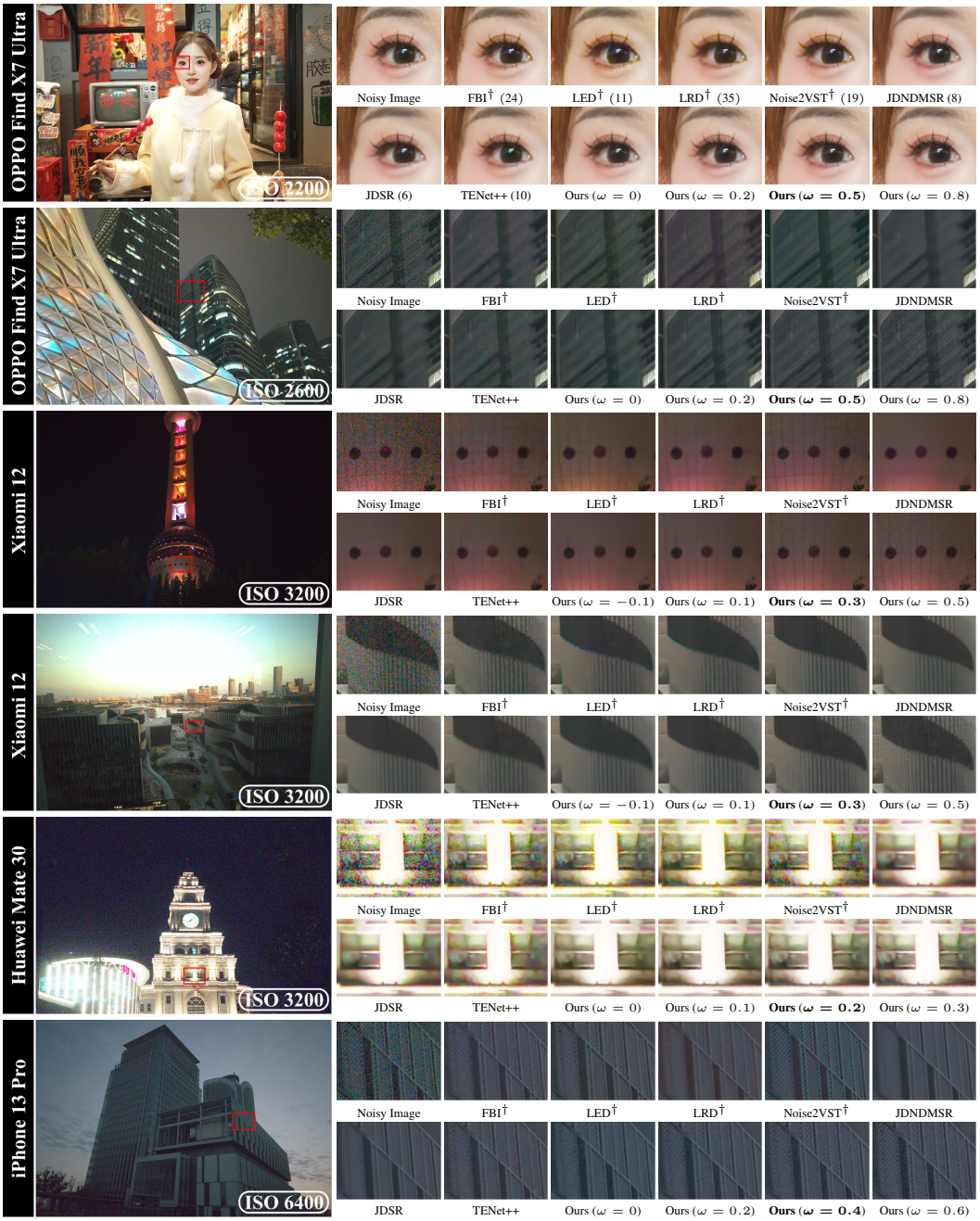



Figure 18: Qualitative comparison across four smartphone brands. † indicates the AHD (5) demosaicing algorithm is applied.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the direction and specific contributions of the proposed work.

   Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This work discusses its limitations in the final section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theorems, formulas, and proofs in the paper are properly numbered and cross-referenced for clarity and consistency.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides a detailed description of the JDD-PRO architecture and clearly specifies the datasets used in the experiments, ensuring reproducibility of the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: After the acceptance of the paper, we will release the code, model parameters, and dataset of the proposed method.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: These details are given in the experimental section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper provides detailed descriptions of the datasets used and experimental setup, but does not include error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the experimental section, we describe the compute resources used for both training and testing.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research complies with the ethical standards specified by NeurIPS.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work improves existing methods in the field, and therefore, will have a positive impact on the industry.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: There is no risk of the paper being misused.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Our paper clearly indicates the cited literature and authors, and all used resources and content are properly acknowledged. The relevant licenses and terms of use are explicitly mentioned and strictly adhered to.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces a new method, and the related code and trained models will be made publicly available after the paper is accepted, along with detailed documentation.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Sec. 4.7 describes the user study setup, evaluation procedure, and instructions provided to participants.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: The user study involved only viewing processed images and expressing preferences regarding their visual quality. No foreseeable risks to participants were identified.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method of this study does not involve LLM related technologies.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.