FIBERR2R: RGB-TO-RAW RECONSTRUCTION WITH LEARNED FORWARD ISP SENSITIVITY GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

RGB-to-RAW reconstruction aims to infer sensor-domain RAW measurements from RGB images, but the image signal processor (ISP) can apply lossy transformations that make the corresponding RAW measurements difficult to recover. To address this challenge, we propose FiberR2R, a three-stage RGB-to-RAW reconstruction framework with learned forward ISP sensitivity guidance. FiberR2R first learns a differentiable forward ISP model from paired RAW–RGB data, then predicts an RGB-conditioned RAW Anchor, and finally trains a residual refiner around the Anchor. We use squared vector–Jacobian product (VJP) responses to provide estimates of the local RGB response energy of the learned forward ISP. These scores jointly parameterize Anchor-centered perturbation sampling and complementary residual supervision. At inference, FiberR2R predicts the Anchor RAW and applies one bounded residual refinement without evaluating the learned forward ISP or its VJPs. Across five camera-specific datasets, FiberR2R has the highest SSIM on all five datasets and the highest PSNR on four, with only 11.51M parameters and 10.38 GFLOPs. These results support learned forward ISP sensitivity as an effective training guidance for RGB-to-RAW reconstruction task.

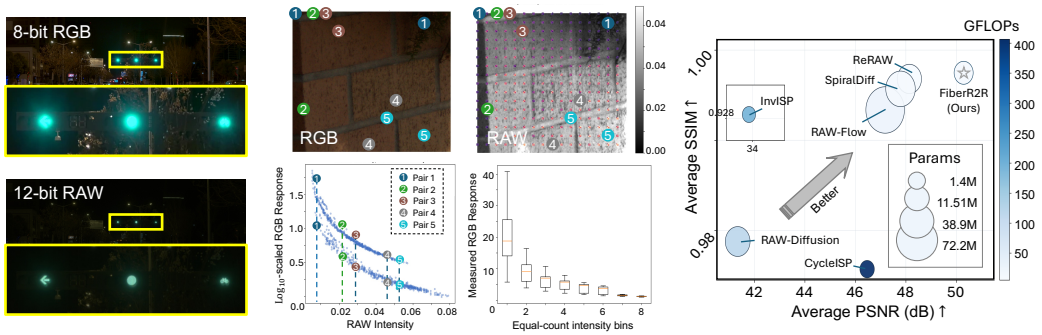
1 INTRODUCTION

Deep learning models have achieved remarkable success in various computer vision tasks. However, their performance strongly depends on the quality of training data. Most existing annotated datasets adopt 8-bit RGB images. These images are lossy outputs produced by the image signal processor (ISP) from RAW sensor measurements, and the processing can distort information useful for low-light and high-dynamic-range imaging. For example, as shown in Figure 6(a), nighttime 8-bit RGB images often contain traffic light glare, which significantly reduces detection accuracy (Yang et al., 2026). In contrast, RAW measurements offer an approximately linear response to scene radiance with higher bit depth. They retain sensor-domain details that are attenuated or even irreversibly distorted during ISP processing (Brooks et al., 2019). In recent years, benefiting from these advances, using RAW data directly has shown great potential for image restoration (Punnappurath et al., 2026; Kim et al., 2024), low-light imaging (Chen et al., 2025; Yue et al., 2025; Liu et al., 2026a), and recognition tasks (Li et al., 2025; 2024).

Nevertheless, to the best of our knowledge, large-scale annotated RAW datasets remain scarce. And collecting and annotating RAW data from scratch requires substantial time and labor. Therefore, transforming annotated RGB images into corresponding RAW measurements (RGB-to-RAW) offers an effective alternative, allowing us to reuse RGB labels to train RAW-domain deep models without expensive re-annotation. And these models can then benefit from the aforementioned advantages of RAW data (e.g., high bit depth), improving task performance in the RAW domain.

Recent RGB-to-RAW methods span convolutional reconstruction (Berdan et al., 2025), flow-matching-based transport (Liu et al., 2026b), and diffusion-based conditional generation (Reinders et al., 2025; Yue et al., 2026), all of which have demonstrated strong reconstruction performance. However, RGB-to-RAW reconstruction is inherently an ill-posed problem, as practical ISP operations including quantization, clipping, denoising, and tone mapping discard information such that a single RGB image can correspond to multiple distinct RAW measurements, and reconstruction difficulty varies across pixel locations. Existing methods such as SpiralDiff (Yue et al., 2026) adapts the diffusion noise to the intensity of the intermediate signal, allowing the model to learn how

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Nighttime traffic light (b) Local forward ISP sensitivity (c) Reconstruction performance comparison

Figure 1: **Motivation for FiberR2R.** (a) For night scenes with traffic lights, RGB exhibits glare, whereas RAW with high bit depth can capture valid signal. (b) Pixels with identical RAW intensity can yield drastically different RGB responses to equal-magnitude RAW perturbations, which motivates sensitivity-aware guidance beyond intensity-based weighting. (c) Across five camera-specific datasets, our FiberR2R achieves competitive reconstruction quality.

RAW reconstruction should vary across different intensity values. This strategy is reasonable when pixel intensity reliably reflects the reconstruction difficulty. Nevertheless, we observe in Figure 6(b) that RAW coordinates with identical intensities can yield drastically different RGB responses under equal-magnitude perturbations. This demonstrates that intensity alone is insufficient to characterize per-pixel RGB-to-RAW reconstruction difficulty.

To address this, we perturb one RAW coordinate at a time by a fixed magnitude and measure the resulting variation in the corresponding RGB coordinate. We directly implement this procedure using a pre-trained forward ISP model (Section 3.2). If the perturbation barely changes the RGB output, nearby RAW candidates are hard to distinguish from the observation; if it induces a notable RGB response, the perturbation can be readily identified. We term the magnitude of this response under the learnable model as learned forward ISP sensitivity. We should note that, throughout this paper, the term *learned forward ISP* refers to the differentiable RAW-to-RGB model learned from paired RAW-RGB data. It functions purely as a training-time differentiable utility for sensitivity computation and rendering constraints rather than a real physical camera ISP, and our core contribution also does not demand perfect mimicry of physical ISP behavior.

Therefore, we propose **FiberR2R**, an RGB-to-RAW reconstruction model that uses coordinate-wise sensitivity from the Jacobian of a learned forward ISP to guide training. Here, Fiber refers to the motivating set of RAW measurements that produce approximately identical RGB observations. But our model learns a deterministic RAW estimate rather than representing or sampling this entire set. In practice, FiberR2R first learns a camera-specific differentiable forward ISP model using paired RAW–RGB data and an RGB-conditioned RAW Anchor estimation model. Around the Anchor, squared vector–Jacobian products estimate the RGB response energy associated with each RAW coordinate under the learned differentiable forward ISP model. The residual branch is initialized at zero, and deployment requires only the Anchor and a one-step bounded residual refinement.

In summary, this paper makes three contributions:

- We introduce local coordinate-wise sensitivity of a learned forward ISP model as a training guidance for RGB-to-RAW reconstruction. Squared coordinate-wise vector–Jacobian product estimates provide a tractable measure of the local RGB response energy with respect to individual RAW coordinates.
- We propose FiberR2R, a three-stage RGB-to-RAW reconstruction framework consisting of a learned differentiable forward ISP model, an RGB-conditioned RAW Anchor, and a sensitivity-guided residual refiner. The learned forward ISP model is used during training to provide rendering-consistency and sensitivity guidance, whereas inference requires only the Anchor and one-step bounded residual refinement.
- Across five camera-specific datasets, FiberR2R achieves the best SSIM on all five and the best PSNR on four, while using only 11.51M parameters and 10.38 GFLOPs. Ablation

studies further demonstrate the empirical utility of sensitivity guidance for RGB-to-RAW reconstruction task.

2 Related Work

Learned ISPs and RAW-Domain Vision. DeepISP (Schwartz et al., 2019) showed that low-level and high-level operations in the RAW-to-RGB pipeline can be learned jointly within a unified framework. Subsequent work extended learned forward imaging to bidirectional RGB–RAW conversion. CycleISP (Zamir et al., 2020) used dual-branch supervised learning for mutual RGB–RAW conversion and synthesized realistic sensor noise from reconstructed RAW data, while InvISP (Xing et al., 2021) employed a reversible network for bidirectional mapping and introduced a differentiable JPEG simulator to model lossy compression.

As RAW data has become more accessible for machine vision, inverse ISP research has expanded beyond image restoration toward sensor-domain representations for downstream perception. Ljungbergh et al. (2023) proposed that learned nonlinear processing of Bayer RAW can improve object detection and proposed a downsampling strategy that preserves the Bayer pattern. More recently, CycleR2R (Li et al., 2024) combined cycle-consistent ISP modeling with randomized illumination variations to remove the need for paired RGB–RAW training data and generate multiple plausible RAW data from a single RGB observation.

High-Fidelity RGB-to-RAW Reconstruction. More recently, RAW-Diffusion (Reinders et al., 2025) injects multi-scale RGB features into an iterative RAW denoising process. ReRAW (Berdan et al., 2025) combines gamma-space multi-head prediction with luminance-aware stratified sampling to improve reconstruction. RAW-Flow (Liu et al., 2026b) formulates conversion between aligned RGB and RAW latent spaces as deterministic transport. SpiralDiff (Yue et al., 2026) introduces a signal-dependent diffusion noise schedule and supports multi-camera generation through camera-specific LoRA adapters. These methods realize RGB-to-RAW reconstruction through conditional generation, luminance-aware sampling, latent-space transport, or camera-conditioned modeling.

3 Method

3.1 Problem Formulation

The prevailing RGB-to-RAW reconstruction paradigms seek high-fidelity recovery of RAW measurements using only paired RGB–RAW data, without prior knowledge of the physical ISP parameters. Formally, let $\boldsymbol{x} \in [0,1]^{4\times H\times W}$ denote a normalized packed Bayer RAW image. Its four channels correspond to the four sampling phases of the 2×2 Bayer pattern and are ordered as (R, G_1, G_2, B). Let $\boldsymbol{y} \in [0,1]^{3\times 2H\times 2W}$ denote the corresponding RGB image, and let c denote the camera index. We formulate RGB-to-RAW reconstruction as a data-driven prediction problem and predict one deterministic RAW estimate conditioned on $(\boldsymbol{y}, c)$. Specifically, rather than recovering or reproducing the physical camera ISP, our FiberR2R realizes this by decomposing the predictor into an RGB-conditioned Anchor A_ψ and a residual refiner R_θ:

$$\boldsymbol{a} = A_\psi(\boldsymbol{y}, c), \qquad \widehat{\boldsymbol{x}} = \mathrm{clip}_{[0,1]}\left(\boldsymbol{a} + R_\theta(\boldsymbol{a}, \boldsymbol{y}, c)\right). \tag{1}$$

The Anchor provides a coarse RAW estimate α, while the residual refiner learns a local correction around it. For training, Stage 1 introduce a learned differentiable forward ISP model F_ϕ for RGB-rendering consistency and local sensitivity guidance. Stage 2 freezes F_ϕ and learns the Anchor A_ψ. Stage 3 freezes both modules and learns the residual refiner R_θ. Thus, F_ϕ is used as a training-time surrogate rather than as a model of the physical camera ISP, and it is not evaluated at inference.

3.2 Learned Differentiable Forward ISP and RAW Anchor

In Stage 1, we learn a differentiable forward ISP model $F_\phi(\boldsymbol{x}, c)$ from paired RAW–RGB data. This differentiability allows gradients to propagate from RGB back to RAW space, which is required for the local sensitivity analysis in Stage 3. As shown in Figure 2, given packed Bayer RAW, the model first reconstructs a linear RGB image through differentiable interpolation and then applies

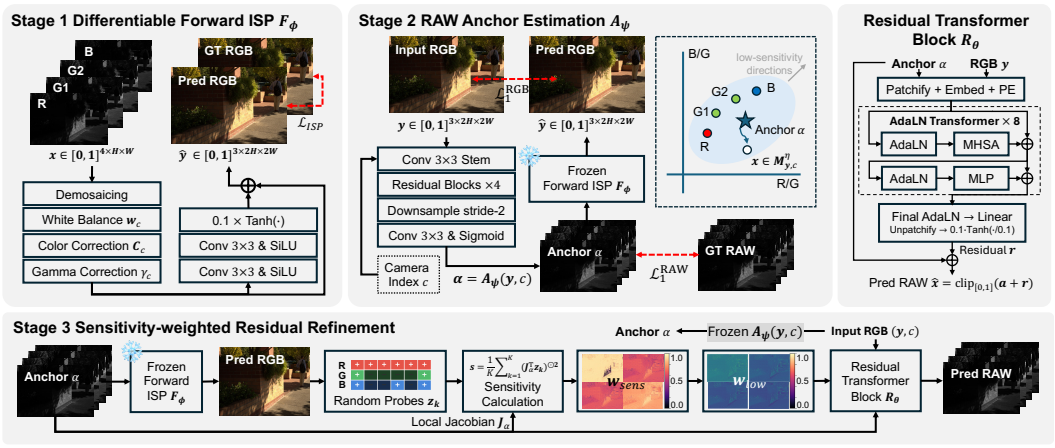


Figure 2: **The FiberR2R pipeline.** Stage 1 learns a differentiable forward ISP model F_ϕ. Stage 2 freezes F_ϕ and estimates an RGB-conditioned RAW Anchor A_ψ. Stage 3 freezes both modules and uses VJP-based sensitivity weights for neighborhood sampling and residual supervision.

global camera transformations such as white balance, color correction and gamma correction. We parameterize these global camera transformations as:

$$\boldsymbol{w}_c = \exp(\widetilde{\boldsymbol{w}}_c), \qquad \boldsymbol{C}_c = \boldsymbol{I} + 0.1\tanh(\widetilde{\boldsymbol{C}}_c), \qquad \gamma_c = 1.6 + 1.2\text{sigmoid}(\widetilde{\gamma}_c). \tag{2}$$

Here, $\widetilde{\boldsymbol{w}}_c$, $\widetilde{\boldsymbol{C}}_c$, and $\widetilde{\gamma}_c$ are learnable camera-indexed embeddings for the white-balance gains, color-correction matrix, and gamma exponent, respectively. Since global camera transformations above are insufficient to model all spatially varying ISP effects, we further employ a shallow convolutional network to predict a bounded residual in RGB space. And we train such learned forward ISP model in Stage 1 using both image and gradient fidelity:

$$\mathcal{L}_{\text{ISP}} = \|F_\phi(\boldsymbol{x}, c) - \boldsymbol{y}\|_1 + \lambda_\nabla \sum_{d\in\{h,w\}} \|\nabla_d F_\phi(\boldsymbol{x}, c) - \nabla_d \boldsymbol{y}\|_1, \tag{3}$$

where ∇_d denotes a finite-difference (FD) spatial gradient and we set $\lambda_\nabla = 0.25$. The first term matches rendered color and brightness, while the second enforces consistency in edges and local textures. After Stage 1, we freeze F_ϕ used for re-ISP processing and train a deterministic RGB-to-RAW Anchor $\boldsymbol{a} = A_\psi(\boldsymbol{y}, c)$ in Stage 2:

$$\mathcal{L}_{\text{Anchor}} = \|\boldsymbol{a} - \boldsymbol{x}\|_1 + \lambda_{\text{render}} \|F_\phi(\boldsymbol{a}, c) - \boldsymbol{y}\|_1. \tag{4}$$

The first term enforces fidelity to the paired RAW target, while the second encourages the Anchor to reproduce the RGB under the learned forward ISP and we set $\lambda_{\text{render}} = 0.1$. Architecturally, A_ψ is a lightweight convolutional network with a 3×3 stem, residual blocks, stride-2 downsampling, and a four-channel sigmoid head. A camera embedding provides scale–shift conditioning.

3.3 Sensitivity-Weighted Residual Refinement

Stage 3 characterizes how local RAW perturbations affect the rendered RGB around the Anchor. The learned forward ISP model in Stage 1 makes the local RGB response differentiable and directly usable as a training guidance. Specifically, for a small RAW perturbation $\boldsymbol{\Delta}$ in a locally smooth neighborhood, a first-order Taylor expansion gives

$$F_\phi(\boldsymbol{a} + \boldsymbol{\Delta}, c) - F_\phi(\boldsymbol{a}, c) = \boldsymbol{J}_{\boldsymbol{a}}\boldsymbol{\Delta} + \mathcal{O}(\|\boldsymbol{\Delta}\|_2^2), \qquad \boldsymbol{J}_{\boldsymbol{a}} = \left.\frac{\partial F_\phi(\boldsymbol{x}, c)}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{a}}.$$

Consequently,

$$\|F_\phi(\boldsymbol{a} + \boldsymbol{\Delta}, c) - F_\phi(\boldsymbol{a}, c)\|_2^2 \approx \boldsymbol{\Delta}^\top \underbrace{(\boldsymbol{J}_{\boldsymbol{a}}^\top \boldsymbol{J}_{\boldsymbol{a}})}_{\boldsymbol{G}_{\boldsymbol{a}} \succeq 0} \boldsymbol{\Delta}. \tag{5}$$

The positive-semidefinite matrix $\boldsymbol{G}_{\boldsymbol{a}}$ describes local RGB response energy under the learned forward ISP. The eigenvectors of $\boldsymbol{G}_{\boldsymbol{a}}$ describe joint perturbation directions, and the corresponding

eigenvalues quantify their local response energy. Our subsequent diagonal statistic focuses on individual RAW coordinates. Explicitly constructing $\boldsymbol{G_a}$ is prohibitive at image resolution. We instead estimate its coordinate-wise energy using K independent Rademacher probes $\boldsymbol{z}_k \in \{-1, +1\}^{D_y}$, where $D_y = 12HW$. For each RGB-space probe, the reverse-mode product $\boldsymbol{J_a^\top z}_k$ is a vector–Jacobian product (VJP). Averaging the squared VJP responses gives a finite-K estimate of the coordinate-wise RGB response energy:

$$\boldsymbol{s} = \frac{1}{K}\sum\nolimits_{k=1}^{K} (\boldsymbol{J_a^\top z}_k)^{\odot 2}, \qquad \mathbb{E}_{\boldsymbol{z}}[\boldsymbol{s} \mid \boldsymbol{a}, c] = \mathrm{diag}(\boldsymbol{G_a}). \tag{6}$$

The probes are sampled in RGB space, and $(\cdot)^{\odot 2}$ denotes elementwise squaring. Thus, the VJP and $\boldsymbol{s}$ lie in $\mathbb{R}^{D_x}$, and s_j estimates the squared norm of the local RGB derivative with respect to RAW coordinate j. Using $\bar{s} = \max(\mathrm{mean}(\boldsymbol{s}), 10^{-8})$, we define normalized high-response weight $\boldsymbol{w}_{\mathrm{sens}}$ and their low-response weight $\boldsymbol{w}_{\mathrm{low}}$:

$$\boldsymbol{w}_{\mathrm{sens}} = \frac{\boldsymbol{s}}{\boldsymbol{s} + \tau\bar{s}}, \qquad \boldsymbol{w}_{\mathrm{low}} = \mathbf{1} - \boldsymbol{w}_{\mathrm{sens}}. \tag{7}$$

Here, $\tau > 0$ is a smoothing hyperparameter and $\mathbf{1} \in \mathbb{R}^{D_x}$ is the all-ones vector. These weights are computed from the frozen learned forward ISP at the Anchor and used during Stage 3 to guide sampling and weighted residual supervision.

Rather than training the refiner only at the Anchor, we sample source states from a bounded Anchor-centered neighborhood. Let $\boldsymbol{u} \sim \mathcal{U}([-1, 1]^{D_x})$ be an elementwise uniform perturbation, let $\boldsymbol{h} = \min_{\mathrm{elem}}(\boldsymbol{a}, \mathbf{1} - \boldsymbol{a})$, and define

$$\boldsymbol{q} = \min_{\mathrm{elem}}(\sigma_s \boldsymbol{w}_{\mathrm{low}} \odot (\boldsymbol{a} + \epsilon_s), \boldsymbol{h}), \qquad \boldsymbol{x}_s = \begin{cases} \boldsymbol{a}, & \text{with probability } p_0, \\ \boldsymbol{a} + \boldsymbol{q} \odot \boldsymbol{u}, & \text{otherwise.} \end{cases} \tag{8}$$

Here, $\sigma_s > 0$ controls the perturbation scale, $\epsilon_s \geq 0$ offsets the intensity factor, and $p_0 \in [0, 1]$ is the probability of retaining the unperturbed Anchor. In implementation, the Bernoulli decision is sampled once per local mini-batch and shared by its samples; Eq. 8 gives the corresponding per-sample marginal distribution. Because the $min(\cdot)$ is elementwise and both arguments are non-negative, $\boldsymbol{q} \leq \boldsymbol{h}$, so every source state remains in $[0, 1]$. Moreover, $\mathbb{E}[\boldsymbol{u}] = 0$ keeps the resulting distribution centered at the Anchor, with $\mathbb{E}[\boldsymbol{x}_s \mid \boldsymbol{a}] = \boldsymbol{a}$. The low-response weight modulates an intensity-scaled neighborhood, subject to the available margin $\boldsymbol{h}$. Appendix A.6 derives its expected local RGB response energy and the scope of this coordinate-wise approximation.

Let the target residual be $\boldsymbol{d} = \boldsymbol{x} - \boldsymbol{x}_s$. We train a camera-specific Transformer on packed-RAW patches and aligned RGB features to predict $\boldsymbol{d}$ from $(\boldsymbol{x}_s, \boldsymbol{y}, c)$. Its output is bounded to keep the predicted refinement local:

$$R_\theta(\boldsymbol{x}_s, \boldsymbol{y}, c) = \delta \tanh\left(\widetilde{R}_\theta(\boldsymbol{x}_s, \boldsymbol{y}, c)/\delta\right), \tag{9}$$

where $\widetilde{R}_\theta$ is the unconstrained residual preactivation, $\tanh(\cdot)$ is applied elementwise, and $\delta = 0.1$ bounds each residual coordinate to $[-\delta, \delta]$. The final residual projection is initialized to zero, so the initial correction is zero: the prediction equals $\boldsymbol{x}_s$ during training and the Anchor at inference. Let $\boldsymbol{r} = R_\theta(\boldsymbol{x}_s, \boldsymbol{y}, c)$ and let $\ell_\beta(\cdot)$ denote the elementwise Smooth-L_1 loss. We separately normalize the high-response and low-response losses:

$$\mathcal{L}_{\mathrm{sens}} = \frac{\mathrm{mean}(\boldsymbol{w}_{\mathrm{sens}} \odot \ell_\beta(\boldsymbol{r}, \boldsymbol{d}))}{\max(\mathrm{mean}(\boldsymbol{w}_{\mathrm{sens}}), \varepsilon)}, \quad \mathcal{L}_{\mathrm{low}} = \frac{\mathrm{mean}(\boldsymbol{w}_{\mathrm{low}} \odot \ell_\beta(\boldsymbol{r}, \boldsymbol{d}))}{\max(\mathrm{mean}(\boldsymbol{w}_{\mathrm{low}}), \varepsilon)}. \tag{10}$$

The means are taken over all RAW coordinates, and $\varepsilon > 0$ is a numerical stabilizer. We perform separate normalization to ensure both components contribute stable gradient magnitude regardless of the fraction of high/low-response pixels in an image. Moreover, we additionally penalize errors in the spatial mean of each CFA sampling phase:

$$\mathcal{L}_\mu = \mathrm{mean}\left[\ell_\beta\left(\underset{H,W}{\mathrm{mean}}(\boldsymbol{x}_s + \boldsymbol{r}), \underset{H,W}{\mathrm{mean}}(\boldsymbol{x})\right)\right]. \tag{11}$$

The inner mean is computed independently for the four CFA phase planes, while the outer mean averages the resulting losses. This term penalizes phase-wise brightness and color-balance errors; it does not impose spatial or frequency-domain consistency. Thus, the complete refiner objective is

$$\mathcal{L}_{\mathrm{Refiner}} = \lambda_{\mathrm{sens}}\mathcal{L}_{\mathrm{sens}} + \lambda_{\mathrm{low}}\mathcal{L}_{\mathrm{low}} + \lambda_\mu\mathcal{L}_\mu. \tag{12}$$

For the default configuration, we use $\lambda_{\mathrm{sens}} = 1.0$, $\lambda_{\mathrm{low}} = 1.0$, $\lambda_\mu = 0.1$.

Table 1: Quantitative comparison of full-resolution reconstruction quality on five datasets. Blue and orange mark the best and second-best results.

Method	Params (M)	GFLOPs	FiveK Canon		FiveK Nikon		NOD Nikon		NOD Sony		PascalRAW	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
InvISP	1.41	184.87	26.74	0.9052	23.92	0.8190	44.10	0.9835	40.63	0.9685	34.36	0.9655
CycleISP	3.14	406.59	38.11	0.9485	39.92	0.9819	55.89	0.9993	48.35	0.9508	50.05	0.9981
RAW-Diffusion	25.01	116.82	39.14	0.9838	40.27	0.9887	44.38	0.9828	41.38	0.9538	41.52	0.9848
ReRAW	23.77	5.01	41.08	0.9939	41.10	0.9931	55.94	0.9993	52.06	0.9986	50.50	0.9991
RAW-Flow	72.19	26.84	41.39	0.9882	41.56	0.9913	54.33	0.9981	52.14	0.9921	50.47	0.9971
SpiralDiff	38.89	10.81	42.59	0.9935	41.66	0.9923	55.95	0.9994	53.12	0.9987	51.61	0.9986
FiberR2R (Ours)	11.51	10.38	42.47	0.9948	42.36	0.9948	57.99	0.9995	54.91	0.9989	53.70	0.9994

4 EXPERIMENTS

Our experiments evaluate our FiberR2R from three perspectives: (1) overall reconstruction performance against existing methods; (2) the contributions of the RGB-conditioned Anchor and residual refiner; (3) the accuracy of VJP-based learned forward ISP sensitivity estimation and its effectiveness for RGB-to-RAW reconstruction task.

4.1 EXPERIMENTAL SETUP

Datasets and comparison scope. We evaluate RGB-to-RAW reconstruction on five camera-specific datasets: FiveK (Bychkovsky et al., 2011) (Canon EOS 5D and Nikon D700), NOD (Morawski et al., 2022) (Nikon D750 and Sony RX100M7), and PascalRAW (Omid-Zohoor et al., 2014). For FiveK and NOD, we use the official splits adopted by SpiralDiff (Yue et al., 2026); for PascalRAW, we use the split provided by RAW-Flow (Liu et al., 2026b). The comparison includes InvISP (Xing et al., 2021), CycleISP (Zamir et al., 2020), RAW-Diffusion (Reinders et al., 2025), ReRAW (Berdan et al., 2025), RAW-Flow (Liu et al., 2026b), and SpiralDiff (Yue et al., 2026). Each baseline is retrained from its official publicly released implementation with reported settings, and both training and evaluation of all methods are performed on NVIDIA H20 GPUs.

Training and inference protocol. We train one FiberR2R model independently for each camera-specific dataset using normalized 64×64 RGB patches and aligned 32×32 four-channel Bayer RAW patches. Training follows the three stages in Section 3; in Stage 3, each local mini-batch retains the Anchor with probability 0.75, and otherwise receives bounded, sensitivity-modulated perturbations. All stages use AdamW with cosine annealing and exponential moving averages (EMA). At inference, the best validation checkpoint processes overlapping full-resolution windows, whose predictions are fused with Hann weights; the model then clips the onepass output to [0, 1]. For FiberR2R, the target RAW is used only during training, and no iterative refinement is performed. Further implementation details are provided in Appendix A.2.

4.2 COMPARISON WITH BASELINES

Table 1 reports full-resolution RGB-to-RAW reconstruction results on the five camera-specific datasets. Among the compared methods, FiberR2R obtains the best SSIM on all five datasets and the best PSNR on four. On FiveK Canon, it trails SpiralDiff by only 0.12 dB in PSNR but has higher SSIM (0.9948 versus 0.9935). FiberR2R has only 11.51M parameters, about 70% fewer than SpiralDiff and fewer than RAW-Flow, with a cost of 10.38 GFLOPs. This highlights that our FiberR2R attains competitive reconstruction quality with substantially reduced model size and computational overhead. Figure 3 shows five representative scenes from all five camera-specific datasets. Each method panel places predicted RAW beside its absolute error map, and the bottom RGGB Q–Q plots compare empirical per-phase quantiles; proximity to the diagonal indicates agreement with ground-truth RAW marginals. This demonstrates that FiberR2R not only reduces per-pixel reconstruction error but also faithfully preserves the marginal statistical properties of real RAW measurements.

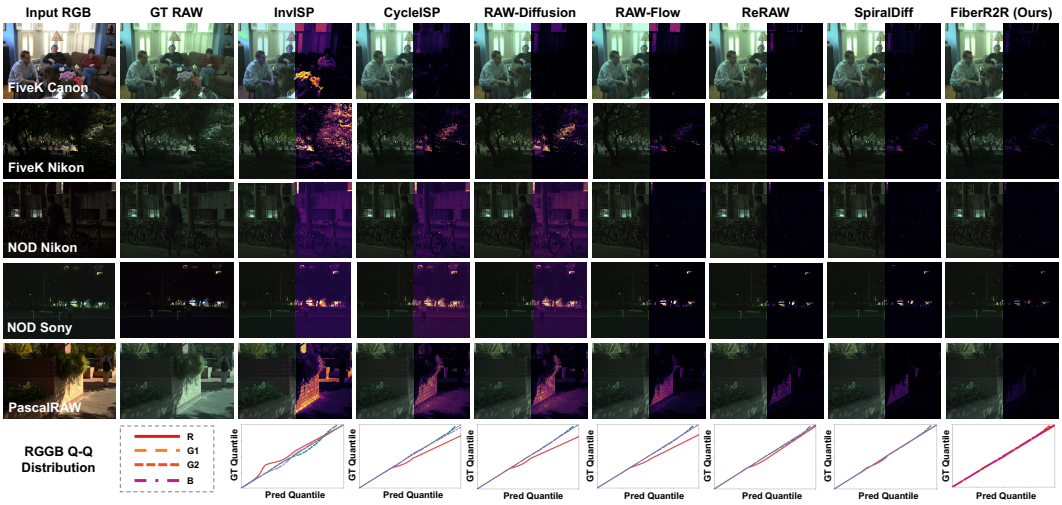


Figure 3: Qualitative comparison of RGB-to-RAW reconstruction. Each method panel shows predicted RAW (left) and absolute error (right); the bottom plots show RGGB Q–Q curves, where proximity to the diagonal indicates agreement with ground-truth marginal quantiles. Across these displayed examples, FiberR2R shows small absolute errors and Q–Q curves close to the diagonal.

4.3 Contribution of the Anchor and Residual Refiner

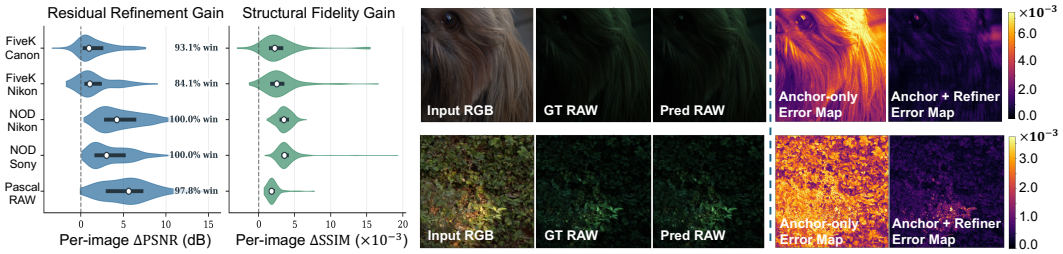
Table 2 isolates the residual refiner while keeping the learned forward ISP and Anchor fixed. Relative to the Anchor-only variant, refinement improves RAW PSNR by 1.65, 1.71, 4.72, 3.58, and 5.46 dB and SSIM by 0.0020–0.0039 across the five datasets. Re-ISP PSNR improves on four datasets and decreases slightly on PascalRAW. In Figure 4(a), PSNR improves on 84.1–100% of test images and SSIM gains are similarly shifted above zero; Figure 4(b) shows lower error magnitude after refinement. These results verify that the bounded residual refiner substantially boosts RAW reconstruction quality on top of the deterministic Anchor prediction.

Table 2: Contribution of the residual refiner on five camera-specific datasets. Forward ISP columns compare $F_\phi(\boldsymbol{x}, c)$ with the paired RGB image; re-ISP PSNR compares each predicted RAW rendered by frozen F_ϕ with that RGB image.

Dataset	Forward ISP RGB			Anchor-only RAW			Refiner RAW			Gain	
	PSNR	SSIM	L1	PSNR	SSIM	re-ISP PSNR	PSNR	SSIM	re-ISP PSNR	PSNR	SSIM
FiveK Canon	34.91	0.9727	0.0186	40.82	0.9919	34.77	42.47	0.9948	35.13	+1.65	+0.0029
FiveK Nikon	34.29	0.9743	0.0205	40.65	0.9918	32.63	42.36	0.9948	34.19	+1.71	+0.0030
NOD Nikon	35.43	0.9197	0.0132	53.27	0.9959	35.11	57.99	0.9995	36.38	+4.72	+0.0036
NOD Sony	32.34	0.9003	0.0182	51.33	0.9950	33.32	54.91	0.9989	33.70	+3.58	+0.0039
PascalRAW	35.49	0.9673	0.0135	48.24	0.9974	35.94	53.70	0.9994	35.82	+5.46	+0.0020

4.4 Ablation Study on Learned VJP Sensitivity Estimation

To evaluate the contribution of VJP guidance, we fix the learned forward ISP, Anchor, refiner architecture, and optimization schedule, and compare several RAW-coordinate guidance strategies. All variants use the same perturbation scale, Anchor-retention probability, training schedule, and inference protocol. *Learned VJP* uses the squared VJP estimate $\boldsymbol{s}$ in Eq. 6 and computes $\boldsymbol{w}_{\text{low}} = \mathbf{1} - \boldsymbol{s}/(\boldsymbol{s} + \tau\bar{s})$. *Intensity* replaces the sensitivity $\boldsymbol{s}$ with the coordinate-wise Anchor magnitude $\boldsymbol{a}$, using $\boldsymbol{w}_{\text{low}} = \mathbf{1} - \boldsymbol{a}/(\boldsymbol{a} + \tau\bar{a})$ with $\bar{a} = \max(\text{mean}(\boldsymbol{a}), 10^{-8})$. *Uniform* removes coordinate variation by setting $\boldsymbol{w}_{\text{low}} = \boldsymbol{w}_{\text{sens}} = 0.5$, while retaining the intensity scaling in the sampler. To disentangle the two uses of VJP guidance, we further report *VJP-sample-only* and *VJP-loss-only*. In *VJP-sample-only*, the learned VJP weights are used only in the Anchor-centered neighborhood



(a) Per-image PSNR / SSIM Gain (b) Visualization of Gains in Anchor-only model via Residual Refiner

Figure 4: Effect of residual refinement beyond the Anchor. (a) Per-image PSNR and SSIM gains; annotations give the fraction of images with positive PSNR gain. (b) Representative predictions and absolute-error maps before and after refinement. The refiner lowers the displayed RAW errors.

sampler in Eq. 8, while the residual loss uses uniform coordinate weights. Conversely, *VJP-loss-only* uses uniform coordinate guidance in the sampler and applies the learned VJP weights only in the residual supervision of Eq. 10. As an analysis-only reference, *FD-reference* replaces the VJP estimate with coordinate-wise finite-difference responses computed through the same frozen F_ϕ and applies the same weighting rule. It therefore provides a surrogate-model reference rather than an estimate of the physical camera ISP.

Table 3: Ablation of RGB-to-RAW reconstruction performance under different RAW-coordinate guidance strategies, including separate VJP-based sampling and loss guidance.

Dataset	Uniform		Intensity		VJP-sample-only		VJP-loss-only		Learned VJP		FD-reference	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
FiveK Canon	40.86	0.9941	41.20	0.9943	41.48	0.9944	41.72	0.9946	42.47	0.9948	42.85	0.9958
FiveK Nikon	41.57	0.9942	41.90	0.9944	42.11	0.9945	42.22	0.9946	42.36	0.9948	42.76	0.9952
NOD Nikon	56.07	0.9982	56.49	0.9984	56.93	0.9987	57.31	0.9990	57.99	0.9995	58.31	0.9995
NOD Sony	52.86	0.9976	53.17	0.9978	53.60	0.9982	54.04	0.9985	54.91	0.9989	55.21	0.9991
PascalRAW	52.11	0.9982	52.44	0.9983	52.80	0.9986	53.12	0.9988	53.70	0.9994	54.02	0.9995

Table 3 shows that the two single-use variants both improve over *Uniform*, with *VJP-loss-only* consistently providing a larger PSNR gain than *VJP-sample-only*. The full *Learned VJP* configuration achieves the best PSNR and SSIM on all five datasets and outperforms either single-use variant, supporting complementary benefits from VJP-guided sampling and residual supervision. The *FD-reference* remains slightly stronger, indicating that the learned VJP captures much, but not all, of the coordinate guidance available from finite differences through the frozen surrogate ISP.

We also assess the learned VJP sensitivity estimation using a response ratio and FD Spearman correlation (Spearman, 1904). Using a shared random direction, we construct low-response and high-response perturbations from $\boldsymbol{w}_{\text{low}}$ and $\boldsymbol{w}_{\text{sens}}$, respectively, and normalize both to equal RAW-space norm. The response ratio r_{res} compares their rendered RGB responses:

$$r_{\text{res}} = \frac{\|F_\phi(\boldsymbol{a} + \boldsymbol{\Delta}_{\text{low}}, c) - F_\phi(\boldsymbol{a}, c)\|_{\text{RMS}}}{\|F_\phi(\boldsymbol{a} + \boldsymbol{\Delta}_{\text{sens}}, c) - F_\phi(\boldsymbol{a}, c)\|_{\text{RMS}}}, \qquad \|\boldsymbol{\Delta}_{\text{low}}\|_2 = \|\boldsymbol{\Delta}_{\text{sens}}\|_2.$$

Here $\|\cdot\|_{\text{RMS}}$ denotes the root-mean-square. The response ratio measures relative RAW–RGB response separation within F_ϕ. Values below one indicate smaller RGB changes for low-response perturbations; lower values indicate greater response separation. FD-Spearman evaluates rank consistency with coordinate-wise finite-difference sensitivities of the same F_ϕ. Figure 5 provides complementary diagnostics of the VJP sensitivity estimate by varying the number of probes $K \in \{1, 4, 8, 16\}$. Increasing K consis-

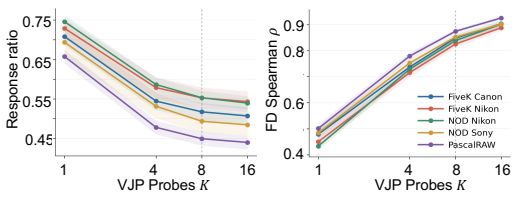


Figure 5: Effect of the VJP probe count K on the response ratio and FD Spearman.

tently reduces the response ratio and increases FD Spearman correlation, indicating that our VJP estimator reliably captures the pixel-wise sensitivity ranking of the frozen learned ISP.

To further examine whether these sensitivities remain informative when evaluated on a separate renderer unused during their estimation, we conduct an additional evaluation using an independent renderer. For each predicted RAW Anchor, we compute coordinate-wise finite-difference sensitivities under this renderer and compare three guidance measures: Learned VJP, Shuffled VJP, and the inverse of the SpiralDiff (Yue et al., 2026) q-sample noise factor.

Shuffled VJP preserves the marginal distribution of the learned VJP sensitivities while disrupting their spatial assignment, and therefore serves as a spatial-control baseline. We report the overall FD-Spearman correlation, the correlation within matched intensity strata, and the low/high-sensitivity response ratio. The latter is defined as the mean renderer response of the lowest-scored 20% of coordinates divided by that of the highest-scored 20%; lower values indicate stronger separation between coordinates assigned low and high sensitivity. The independent renderer is fixed, has no learned parameters, and is used only for evaluation. Its finite-difference sensitivities are computed independently of F_ϕ. Confidence intervals are obtained by image-level bootstrap resampling.

Table 4: Independent renderer evaluation of coordinate-wise sensitivity guidance. VJP denotes learned VJP sensitivity estimation, Shuf. denotes spatially shuffled VJP, and SD-q^{-1} denotes the inverse of the SpiralDiff q-sample noise factor rather than the complete SpiralDiff model.

Dataset	FD-Spearman ↑			Within-intensity ρ ↑			Response ratio ↓		
	VJP	Shuf.	SD-q^{-1}	VJP	Shuf.	SD-q^{-1}	VJP	Shuf.	SD-q^{-1}
FiveK Canon	**0.850** [0.827,0.868]	0.734 [0.705,0.760]	0.523 [0.478,0.566]	**0.791** [0.764,0.815]	0.611 [0.578,0.644]	0.429 [0.387,0.470]	**0.191** [0.168,0.216]	0.242 [0.216,0.268]	0.430 [0.373,0.490]
FiveK Nikon	**0.814** [0.777,0.841]	0.697 [0.649,0.734]	0.536 [0.470,0.595]	**0.705** [0.662,0.743]	0.505 [0.453,0.551]	0.467 [0.403,0.525]	**0.214** [0.174,0.261]	0.268 [0.225,0.316]	0.399 [0.323,0.475]
NOD Nikon	**0.845** [0.838,0.853]	0.646 [0.632,0.660]	0.303 [0.266,0.340]	**0.827** [0.819,0.835]	0.602 [0.589,0.614]	0.258 [0.224,0.290]	**0.205** [0.194,0.216]	0.335 [0.319,0.351]	0.597 [0.555,0.638]
NOD Sony	**0.834** [0.821,0.845]	0.669 [0.652,0.684]	0.368 [0.336,0.401]	**0.802** [0.790,0.814]	0.599 [0.583,0.614]	0.288 [0.262,0.314]	**0.169** [0.158,0.180]	0.262 [0.246,0.281]	0.497 [0.460,0.537]
PascalRAW	**0.828** [0.776,0.871]	0.727 [0.670,0.771]	0.347 [0.288,0.400]	**0.757** [0.698,0.807]	0.576 [0.513,0.628]	0.321 [0.268,0.374]	**0.221** [0.185,0.264]	0.270 [0.228,0.321]	0.635 [0.554,0.715]

As shown in Table 4, learned VJP sensitivity estimation achieves the highest FD-Spearman and intensity-stratified correlations and the lowest low/high-sensitivity response ratio on all five datasets. Relative to Shuffled VJP, its FD-Spearman correlation is higher by 0.101–0.199, its intensity-stratified correlation is higher by 0.180–0.225, and its response ratio is lower by 0.049–0.130. The consistent degradation after spatial shuffling shows that the result depends on the coordinate assignment of the VJP sensitivities, rather than only on their marginal distribution.

Learned VJP also achieves better results than the SpiralDiff q-sample intensity measure. Its advantage remains within matched intensity strata, indicating that the learned forward ISP sensitivity captures coordinate-dependent information beyond the intensity-based guidance. Together, these results provide complementary evidence that the VJP sensitivities estimated from the learned forward ISP remain informative under an independently specified renderer.

5 Conclusion

In this paper, FiberR2R leverages coordinate-wise sensitivity derived from a learned forward ISP to provide training guidance for RGB-to-RAW reconstruction. Squared VJP estimates computed around an RGB-conditioned RAW anchor guide training-time perturbation sampling and residual supervision. During inference, FiberR2R predicts the RAW anchor and applies one-step bounded residual refinement. Across evaluated benchmarks, FiberR2R achieves the highest SSIM on all five camera-specific datasets and the highest PSNR on four, with only 11.51M parameters and 10.38 GFLOPs. When jointly deployed for sampling and loss computation, learned VJP-based sensitivity guidance outperforms intensity-driven and uniform-weight alternatives. These findings demonstrate the efficacy of learned forward ISP sensitivity as training guidance for RGB-to-RAW reconstruction, while offering compact and one-step inference.

REFERENCES

Radu Berdan, Beril Besbinar, Christoph Reinders, Junji Otsuka, and Daisuke Iso. Reraw: Rgb-to-raw image reconstruction via stratified sampling for efficient object detection on the edge. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11833–11843, 2025.

Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. Unprocessing images for learned raw denoising. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11028–11037, 2019.

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

Xianmin Chen, Longfei Han, Peiliang Huang, Xiaoxu Feng, Dingwen Zhang, and Junwei Han. Retinex-RAWMamba: Bridging Demosaicing and Denoising for Low-Light RAW Image Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

Woohyeok Kim, Geonu Kim, Junyong Lee, Seungyong Lee, Seung-Hwan Baek, and Sunghyun Cho. Paramisp: Learned forward and inverse isps using camera parameters. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26067–26076, 2024.

Zhihao Li, Ming Lu, Xu Zhang, Xin Feng, M. Salman Asif, and Zhan Ma. Efficient visual computing with camera raw snapshots. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (7):4684–4701, 2024.

Zhong-Yu Li, Xin Jin, Bo-Yuan Sun, Chun-Le Guo, and Ming-Ming Cheng. Towards raw object detection in diverse conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8859–8868, 2025.

Haoyue Liu, Jinghan Xu, Luxin Feng, Hanyu Zhou, Haozhi Zhao, Yi Chang, and Luxin Yan. Necdiff: Noise-robust event-raw complementary diffusion for seeing motion in extreme darkness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22281–22290, 2026a.

Zhen Liu, Diedong Feng, Hai Jiang, Liaoyuan Zeng, Hao Wang, Chaoyu Feng, Lei Lei, Bing Zeng, and Shuaicheng Liu. Raw-flow: advancing rgb-to-raw image reconstruction with deterministic latent flow matching. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, pp. 7431–7439, 2026b.

William Ljungbergh, Joakim Johnander, Christoffer Petersson, and Michael Felsberg. Raw or cooked? object detection on raw images. In *Image Analysis*, pp. 374–385, 2023.

Igor Morawski, Yu-An Chen, Yu-Sheng Lin, Shusil Dangi, Kai He, and Winston H Hsu. Genisp: Neural isp for low-light machine cognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 630–639, 2022.

Alex Omid-Zohoor, David Ta, and Boris Murmann. PASCALRAW: Raw image database for object detection, 2014. URL `http://purl.stanford.edu/hq050zr7488`.

Abhijith Punnappurath, Luxi Zhao, Ke Zhao, Hue Nguyen, Radek Grzeszczuk, and Michael S Brown. Edit-aware raw reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8418–8427, 2026.

Christoph Reinders, Radu Berdan, Beril Besbinar, Junji Otsuka, and Daisuke Iso. Raw-diffusion: Rgb-guided diffusion models for high-fidelity raw image generation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8431–8443, 2025.

Eli Schwartz, Raja Giryes, and Alex M. Bronstein. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE Transactions on Image Processing*, 28(2):912–923, 2019.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6283–6292, 2021.

Tingyu Yang, Yuan Cheng, and Xiaoyun Yuan. Lucid: Learning unified control for image deflaring and exposure mastery in nighttime photography. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 11, 2026.

Huanjing Yue, Cong Cao, Lei Liao, and Jingyu Yang. Rvideformer: Efficient raw video denoising transformer with a larger benchmark dataset. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):8929–8944, 2025.

Huanjing Yue, Shangbin Xie, Cong Cao, Qian Wu, Lei Zhang, Lei Zhao, and Jingyu Yang. Spiraldiff: Spiral diffusion with lora for rgb-to-raw conversion across cameras. *arXiv preprint arXiv:2603.14885*, 2026.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2693–2702, 2020.

A Appendix

A.1 Real-world Application

In real captured nighttime scenes, the 8-bit RGB image of traffic lights sometimes exhibit prominent glare, rendering road signs illegible. In contrast, the corresponding 12-bit RAW generated by our FiberR2R model can effectively recover high-bit-depth information to eliminate glare while preserving valid details, which significantly improves driving safety for autonomous vehicles.

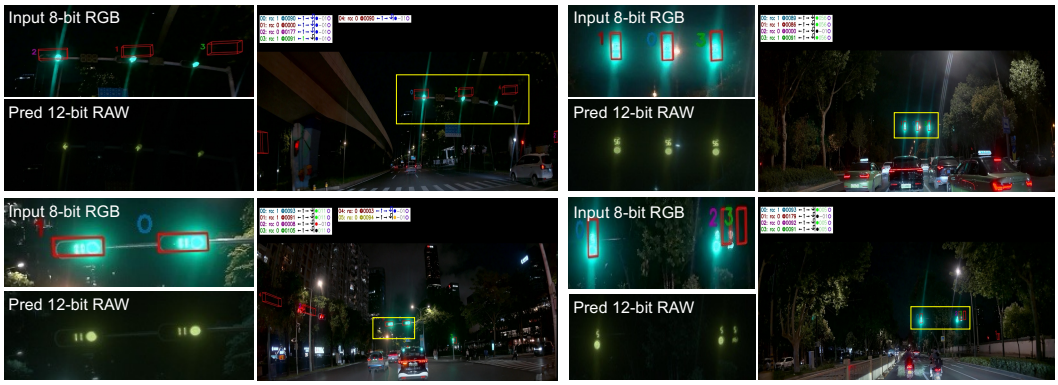


Figure 6: Visual comparison between 8-bit input RGB and the 12-bit RAW generated by FiberR2R for nighttime driving scenarios. The 12-bit output RAW can effectively recover useful information of traffic lights while suppressing glare artifacts.

A.2 Details of the Training and Inference Protocol

We train a separate FiberR2R model for each of the five camera-specific datasets. For each training image, we sample 144 aligned patch pairs. Each pair consists of a 64×64 normalized RGB patch and its corresponding 32×32 four-channel packed Bayer RAW patch.

The model is trained in three stages. In Stage 1, only the camera-conditioned learned forward ISP model is optimized using AdamW with a learning rate of 4×10^{-4} and weight decay of 10^{-4}. The training objective combines an L_1 reconstruction loss with a spatial-gradient loss on the rendered RGB output. In Stage 2, the learned forward ISP model is frozen at its Stage-1 EMA weights, and only the RAW Anchor module is optimized using the RAW reconstruction and RGB re-ISP terms in

Eq. 4, with a learning rate of 3×10^{-4} and weight decay of 10^{-4}. In Stage 3, the learned forward ISP model and Anchor module are frozen, and only the residual refiner is optimized using AdamW with a learning rate of 10^{-4} and weight decay of 0.05. The Stage-3 objective combines the sensitivity-weighted loss $\mathcal{L}_{\text{sens}}$ and the complementary low-response loss $\mathcal{L}_{\text{low}}$ in Eq. 10, each with coefficient 1.0, together with the CFA phase-wise mean loss with weight 0.1. All stages employ a 5-epoch linear warm-up followed by cosine annealing, BF16 mixed-precision training, and an EMA decay rate of 0.9999. Gradient clipping is set to 1.0 in the first two stages and 0.5 in Stage 3. The Stage-1/2/3 epoch schedules for FiveK Canon, FiveK Nikon, NOD Nikon, NOD Sony, and PascalRAW are $180/180/260$, $240/240/340$, $38/38/56$, $48/48/70$, and $240/240/340$, respectively.

At inference, we use the Stage-3 EMA checkpoint selected on the validation split. Each full-resolution RGB image is processed using overlapping 64×64 RGB-space windows with a 16-pixel overlap, and the corresponding packed-RAW predictions are fused using Hann-weighted averaging. FiberR2R first predicts the deterministic Anchor from the input RGB image and then applies a single bounded residual refinement before clipping the final RAW output to $[0, 1]$. Inference requires neither target RAW data nor iterative refinement.

For completeness, additional implementation and evaluation details are all provided in the supplementary code package.

A.3 Diagnostic Protocols

The independent renderer used in Table 4 performs fixed normalized interpolation using the separable kernel $[1, 2, 1]^\top [1, 2, 1]/16$ while preserving the measured mosaic samples at their original locations. It then clips the interpolated output to $[0, 1]$ and applies a gamma transform with $\gamma = 2.2$. The renderer is fixed, contains no learned parameters, and is used only for evaluation. For each test image, the protocol evaluates the full-resolution predicted RAW Anchor. Centered finite differences with step $h = 0.005$ are computed at coordinates sampled with spatial stride four and restricted to values satisfying $a_j - h \in [0, 1]$ and $a_j + h \in [0, 1]$. Consequently, validity filtering near the intensity boundaries can reduce the effective number of evaluated coordinates. Let $\mathcal{G}$ denote the independent renderer and let $\boldsymbol{e}_j$ be the basis vector for RAW coordinate j. The finite-difference RGB response for that coordinate is

$$s_j^{\text{FD}} = \left\| \frac{\mathcal{G}(\boldsymbol{a} + h\boldsymbol{e}_j) - \mathcal{G}(\boldsymbol{a} - h\boldsymbol{e}_j)}{2h} \right\|_2^2, \tag{13}$$

which sums the squared RGB derivative over the rendered image. For the Shuffled-VJP comparison, VJP sensitivities are shuffled among coordinates belonging to the same image region, CFA channel, and intensity group. This preserves the sensitivity distribution within each channel–intensity group while changing its spatial assignment. The within-intensity statistic pools CFA channels within each of four equal-width intensity bins. For each guidance score, the low/high-sensitivity response ratio is the mean finite-difference response of the lowest-scored 20% of coordinates divided by that of the highest-scored 20%; a smaller value indicates clearer separation between coordinates assigned low and high forward ISP sensitivity. Table 3 reports 95% percentile intervals obtained from 1,000 bootstrap resamples over the evaluated image-level diagnostic records. Table 4 uses 2,000 bootstrap resamples over complete test images. Together, these diagnostic protocols evaluate whether the VJP-derived sensitivity ranking agrees with finite-difference RGB responses under both the learned forward ISP and an independently specified renderer.

A.4 Three-Seed Stability on Two Datasets

Table 5 reports RAW and re-ISP PSNR from three random-seed runs on FiveK Canon (Canon EOS 5D) and NOD Nikon (Nikon D750), together with the mean and sample standard deviation across seeds. The RAW PSNR standard deviations are 0.0329 and 0.0084 dB, respectively, and the corresponding re-ISP PSNR standard deviations are 0.0291 and 0.0082 dB. The small standard deviations indicate consistent performance across the three evaluated seeds on these two datasets.

A.5 Bounded-residual saturation audit

The residual refiner uses $\boldsymbol{R}_\theta = \delta \tanh(\tilde{\boldsymbol{R}}_\theta/\delta)$ with $\delta = 0.1$, which mathematically constrains the network-produced residual to satisfy $|R_{\theta,j}| \leq \delta$ before the final $[0, 1]$ clipping operation. This

Table 5: Three-seed stability of FiberR2R on two datasets. Results report per-seed PSNR and mean ± sample standard deviation.

Dataset	Metric	Num. of Images	Full Model (mean ± sample SD)	Seed	Per-seed PSNR
Canon EOS 5D	RAW PSNR (dB)	116	42.4658 ± 0.0329	2027	42.4574
				2028	42.4379
				2029	42.5021
Canon EOS 5D	re-ISP PSNR (dB)	116	35.1285 ± 0.0291	2027	35.1123
				2028	35.1112
				2029	35.1621
NOD Nikon D750	RAW PSNR (dB)	357	57.9850 ± 0.0084	2027	57.9787
				2028	57.9818
				2029	57.9946
NOD Nikon D750	re-ISP PSNR (dB)	357	36.3822 ± 0.0082	2027	36.3915
				2028	36.3793
				2029	36.3758

hard bound applies only to the network output R_θ; the ground-truth residual $|x_j - a_j|$ between target RAW and the Anchor prediction is not constrained and can exceed δ. Given a fixed Anchor, ground-truth residuals larger than δ cannot be fully corrected within one bounded refinement step, creating a reachability limitation.

We perform an inference-only audit on held-out test data using converged checkpoints without retraining. In total we evaluate all test images across five camera-specific datasets. Table 6 reports pixel fractions for two quantities: (1) how often the network-output residual $|R_\theta|$ approaches the tanh-imposed bound; (2) the fraction of pixels where ground-truth residual magnitude exceeds $\delta = 0.1$. Across evaluated test coordinates, learned corrections rarely approach the prescribed bound, indicating that the refiner generally operates in an unsaturated regime. Only 0.041% of coordinates satisfy $|R_\theta| \geq 0.95\delta$, and merely 0.023% satisfy $|R_\theta| \geq 0.99\delta$. Nevertheless, 0.17%–1.71% of test pixels exhibit ground-truth residuals $|x - a| > 0.1$. Given a fixed Anchor, these large residuals cannot be fully eliminated by our single-step bounded correction. This reachability limitation contributes a non-negligible theoretical lower bound to the overall refiner MSE and represents an inherent limitation of our one-step refinement formulation.

Table 6: Audit of residual refiner bound utilization on the test set. Reported values are fractions (in percent) of RAW coordinates satisfying each threshold. $|R_\theta|$: network-output residual before final clipping. $|x - a|$: ground-truth residual between ground-truth RAW and Anchor prediction.

Dataset	$\lvert R_\theta\rvert \geq 0.95\delta$ (%)	$\lvert R_\theta\rvert \geq 0.99\delta$ (%)	$\lvert x - a\rvert > 0.1$ (%)
FiveK Canon	0.0012	0.0001	1.2155
FiveK Nikon	0.1095	0.0539	1.7052
NOD Nikon	0.0363	0.0187	0.1742
NOD Sony	0.0466	0.0292	0.3864
PascalRAW	0.0278	0.0159	0.2768

A.6 Analysis of Coordinate-Wise Sensitivity Guidance

The following analysis characterizes the squared-VJP estimator and its roles in neighborhood sampling and residual supervision. All quantities are defined locally around an Anchor under the learned forward ISP model F_ϕ.

Squared VJP estimator. Condition on $\boldsymbol{a}$ and c, and write $\boldsymbol{J} = \boldsymbol{J}_{\boldsymbol{a}}$ and $\boldsymbol{G} = \boldsymbol{J}^\top \boldsymbol{J}$. For a Rademacher probe with independent entries, $\mathbb{E}_{\boldsymbol{z}}[\boldsymbol{z}\boldsymbol{z}^\top] = \boldsymbol{I}$. If $\boldsymbol{e}_j$ is the basis vector associated with the jth RAW coordinate, then

$$\mathbb{E}_{\boldsymbol{z}}\left[(\boldsymbol{J}^\top \boldsymbol{z})_j^2\right] = \boldsymbol{e}_j^\top \boldsymbol{J}^\top \mathbb{E}_{\boldsymbol{z}}[\boldsymbol{z}\boldsymbol{z}^\top] \boldsymbol{J}\boldsymbol{e}_j = G_{jj} = \|\boldsymbol{J}\boldsymbol{e}_j\|_2^2. \tag{14}$$

Linearity of expectation gives Eq. 6 for any $K \geq 1$. The finite-K estimator is unbiased for the coordinate-wise response energy, although it remains random, and the nonlinear normalization in Eq. 7 need not preserve this unbiasedness. The resulting diagonal estimate characterizes the RGB response energy of individual RAW coordinates; correlated multi-coordinate directions are outside the scope of this coordinate-wise approximation.

Expected response of the sampled neighborhood. Condition on $\boldsymbol{a}$, c, and the realized VJP estimate $\boldsymbol{s}$, which fix the scale vector $\boldsymbol{q}$ in Eq. 8. Although the Bernoulli decision is shared within each local mini-batch during training, its per-sample marginal can be represented by a variable b that is independent of $\boldsymbol{u}$, with $b = 1$ with probability $1 - p_0$ and $b = 0$ otherwise. Then

$$\boldsymbol{\Delta} = \boldsymbol{x}_s - \boldsymbol{a} = b\,\boldsymbol{q} \odot \boldsymbol{u}.$$

The independent uniform entries satisfy $\mathbb{E}[u_j] = 0$ and $\mathbb{E}[u_i u_j] = 1/3$ for $i = j$ and zero otherwise. Consequently,

$$\begin{aligned} \mathbb{E}[\boldsymbol{\Delta}\boldsymbol{\Delta}^\top \mid \boldsymbol{a}, c, \boldsymbol{s}] &= \frac{1 - p_0}{3}\,\mathrm{diag}(\boldsymbol{q}^{\odot 2}), \\ \mathbb{E}[\|\boldsymbol{J}\boldsymbol{\Delta}\|_2^2 \mid \boldsymbol{a}, c, \boldsymbol{s}] &= \frac{1 - p_0}{3} \sum\nolimits_{j=1}^{D_x} q_j^2 G_{jj}. \end{aligned} \tag{15}$$

This identity gives the exact expected energy of the linearized response $\boldsymbol{J}\boldsymbol{\Delta}$: independence and zero-mean perturbations eliminate the off-diagonal terms in expectation. It explains why coordinate-wise sensitivity is directly relevant to the proposed sampling distribution, even though the diagonal approximation does not represent general correlated perturbations. For finite perturbations, the actual nonlinear RGB response is approximated through the local first-order expansion. Because $\boldsymbol{q}$ also contains Anchor-dependent intensity scaling and boundary-aware clipping, the sampled neighborhood jointly reflects forward ISP sensitivity, Anchor intensity, and the available value-domain margin.

Effective weighting of the residual loss. Write $w_j = w_{\mathrm{sens},j}$, $m = \mathrm{mean}(\boldsymbol{w}_{\mathrm{sens}})$, and $\ell_j = \ell_\beta(r_j, d_j)$, with all means taken over the same RAW coordinates as in Eq. 10. Since $w_{\mathrm{low},j} = 1 - w_j$, the two weighted losses combine exactly as

$$\lambda_{\mathrm{sens}}\mathcal{L}_{\mathrm{sens}} + \lambda_{\mathrm{low}}\mathcal{L}_{\mathrm{low}} = \mathrm{mean}_j(\alpha_j \ell_j), \quad \alpha_j = \frac{\lambda_{\mathrm{sens}} w_j}{\max(m, \varepsilon)} + \frac{\lambda_{\mathrm{low}}(1 - w_j)}{\max(1 - m, \varepsilon)}. \tag{16}$$

For the default setting $\lambda_{\mathrm{sens}} = \lambda_{\mathrm{low}} = 1$, when both means exceed ε,

$$\alpha_j = \frac{1}{1 - m} + \frac{1 - 2m}{m(1 - m)} w_j. \tag{17}$$

Thus, $m > 1/2$ assigns larger effective weights to smaller w_j, $m < 1/2$ gives the opposite ordering, and $m = 1/2$ yields $\alpha_j = 2$ for every coordinate. When neither denominator is clamped, separate normalization gives each complementary term unit mean weight before its coefficient is applied. Accordingly, equal coefficients give equal global normalization to the two complementary terms, while the resulting coordinate preference is determined by m. The neighborhood sampler and residual weighting therefore provide two complementary uses of the same forward ISP sensitivity estimate.

Bounded residual refinement. Equation 9 ensures $|r_j| \leq \delta$ for every RAW coordinate. At inference, because $a_j \in [0, 1]$ and clipping projects the corrected estimate onto the same interval,

$$|\widehat{x}_j - a_j| = \left|\mathrm{clip}_{[0,1]}(a_j + r_j) - a_j\right| \leq |r_j| \leq \delta. \tag{18}$$

The residual refiner therefore performs a single bounded, deterministic correction around the RGB-conditioned Anchor.