This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SpiralDiff: Spiral Diffusion with LoRA for RGB-to-RAW Conversion Across Cameras

Huanjing Yue[1,2] Shangbin Xie[1,2] Cong Cao[1] Qian Wu[3] Lei Zhang[3] Lei Zhao[3] Jingyu Yang[1*]
[1]School of Electrical and Information Engineering, Tianjin University
[2]State key laboratory of Smart Power Distribution Equipment and System, Tianjin University [3]Individual
{huanjing.yue, chuancyx, caocong_123, yjy}@tju.edu.cn {youzhagao, Zhaoleiyyy}@gmail.com qian.wu.alex@outlook.com

Abstract

RAW images preserve superior fidelity and rich scene information compared to RGB, making them essential for tasks in challenging imaging conditions. To alleviate the high cost of data collection, recent RGB-to-RAW conversion methods aim to synthesize RAW images from RGB. However, they overlook two key challenges: (i) the reconstruction difficulty varies with pixel intensity, and (ii) multi-camera conversion requires camera-specific adaptation. To address these issues, we propose SpiralDiff, a diffusion-based framework tailored for RGB-to-RAW conversion with a signal-dependent noise weighting strategy that adapts reconstruction fidelity across intensity levels. In addition, we introduce CamLoRA, a camera-aware lightweight adaptation module that enables a unified model to adapt to different camera-specific ISP characteristics. Extensive experiments on four benchmark datasets demonstrate the superiority of SpiralDiff in RGB-to-RAW conversion quality and its downstream benefits in RAW-based object detection. Our code is available at https://github.com/Chuancy-TJU/SpiralDiff.

1. Introduction

As the direct output of a camera sensor, RAW images preserve rich, unprocessed scene information with a linear radiometric response and high dynamic range. These images are subsequently processed by an Image Signal Processor (ISP) to produce RGB images suitable for display. During this transformation, crucial photometric information is inevitably altered or discarded through operations such as demosaicing, white balance, tone mapping, and JPEG compression. As a result, the linearity, dynamic range, and realistic noise characteristics inherent in RAW data are largely lost in RGB images.

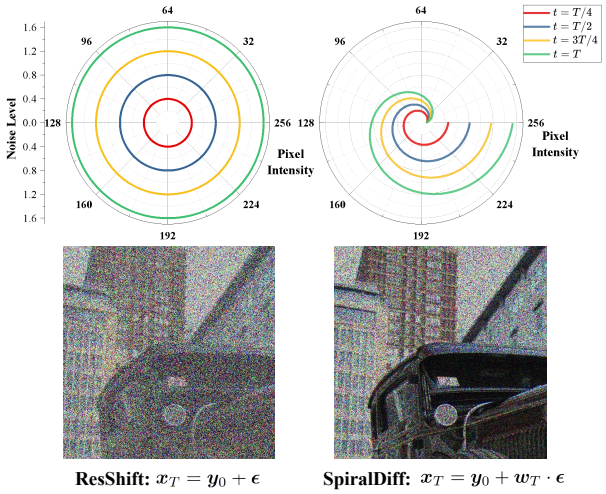


Figure 1. Comparison of noise schedule (top) and visualization of the noisy $\boldsymbol{x}_T$ (bottom) in diffusion: ResShift [45] uses uniform Gaussian noise, whereas our SpiralDiff introduces a signal-dependent noise schedule based on pixel intensity ($\boldsymbol{w}_t$).

Owing to these advantages, recent studies have demonstrated that performing computer vision tasks directly in the RAW domain can yield superior results in denoising [1, 44], low-light enhancement [7, 49], and object detection [15, 25, 41]. Although existing RAW datasets provide valuable resources, they are limited in scale and diversity compared with RGB collections. Moreover, it is a costly and time-consuming process to construct a new RAW dataset for a specific sensor, often requiring extensive data collection and large storage resources. To alleviate this burden, recent studies have explored RGB-to-RAW conversion. It aims to invert the ISP pipeline and reconstruct plausible RAW images from abundantly available RGB inputs without repeatedly collecting new sensor-specific RAW datasets.

Existing RGB-to-RAW conversion methods can be

*This work was supported in part by the National Natural Science Foundation of China under Grant 62472308 and Grant 62231018. Corresponding author: Jingyu Yang

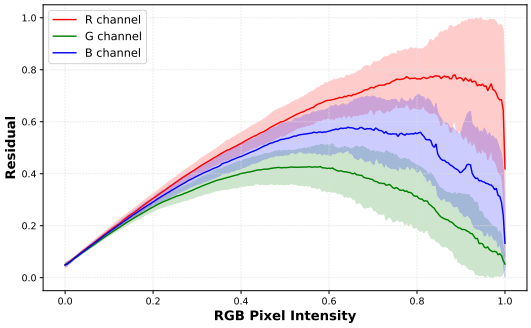


Figure 2. Relationship between RGB pixel intensity and residual magnitude between RGB and RAW images across channels on the FiveK Nikon dataset. Colored lines and shaded regions represent the mean and standard deviation, respectively.

broadly classified into metadata-based and metadata-free approaches. Metadata-based approaches [5, 8, 24, 29, 32] leverage ISP parameters or sampled RAW pixels, but such metadata is rarely accessible in real-world settings. Metadata-free methods [3, 10, 33, 40, 47] bypass this requirement by learning the mapping end-to-end. Although they achieve competitive results, they still overlook two fundamental challenges: (i) The conversion difficulty is inherently signal-dependent, varying with pixel intensity. As shown in Fig. 2, residuals between RAW and RGB exhibit strong dependence on pixel intensity. Specifically, in low-intensity regions, the residuals are small and stable, which enables high-fidelity recovery. In contrast, high-intensity and over-exposed regions suffer from large and uncertain residuals due to non-linear tone mapping and value clipping, making accurate prediction challenging. This indicates that a globally uniform reconstruction strategy in existing methods is suboptimal. Instead, an intensity-aware mechanism is required to adapt reconstruction flexibility according to local difficulty. (ii) Effective multi-camera RGB-to-RAW conversion requires a unified model with camera-specific adaptation. A naive approach to unification is to merge data from multiple cameras into a single training set. However, this leads to performance degradation because the model conflates different ISP characteristics and learns a compromised representation. This suggests a camera-specific adaptation is necessary to preserve the respective characteristics.

Based on the above analysis, we propose SpiralDiff, a diffusion-based framework tailored for RGB-to-RAW conversion. To handle intensity-dependent reconstruction, we introduce a spatially variant noise-weighting strategy within the diffusion process. Specifically, we build a Markov chain for the transition between the RAW image and its RGB counterpart based on ResShift [45, 46], an efficient diffusion framework for image restoration. We then modulate the isotropic Gaussian perturbation with a signal-dependent weight map: low-intensity regions receive relatively small perturbations to preserve fidelity, while high-intensity and over-exposed regions are injected with stronger noise to allow more flexible reconstruction under severe nonlinearity and clipping. An intuitive comparison is illustrated in Fig. 1. The weighting strategy is further extended to a time-variant schedule that evolves with the diffusion steps, aligning the noise injection with the transformation process between RAW and RGB. In addition, to mitigate cross-camera interference when training a unified model, we propose CamLoRA, a camera-aware LoRA module. It consists of several lightweight LoRA layers that learn camera-specific adaptations conditioned on an input camera label. Such a design enables consistent camera-aware conversion within a unified model with minimal additional parameters. Moreover, once the unified model is pretrained, it can be efficiently adapted to a new camera by fine-tuning only one LoRA layer, effectively leveraging the learned priors. Our major contributions are summarized as follows:

- We propose SpiralDiff, a diffusion-based framework tailored for RGB-to-RAW conversion. A signal-dependent noise weighting strategy is introduced to adaptively balance reconstruction fidelity and generative flexibility across regions with different intensity characteristics.
- We introduce a camera-aware adaptation module (CamLoRA) for RGB-to-RAW conversion, enabling the model to generate RAW outputs that reflect device-specific properties conditioned on the input camera label.
- Experiments demonstrate that SpiralDiff surpasses state-of-the-art approaches on four benchmark datasets. In addition, the generated RAW images improve downstream object detection performance.

2. Related Work

RGB-to-RAW Conversion. The increasing need for large-scale datasets in the RAW domain has motivated research into synthesizing RAW images from RGB inputs. Existing approaches can be broadly divided into metadata-dependent and metadata-free methods. The former [5, 8, 29] reconstruct RAW images by leveraging ISP parameters or a subset of RAW data stored alongside the RGB image. However, such metadata is rarely available in real-world scenarios, limiting their practical applicability. Metadata-free RGB-to-RAW conversion aims to learn the inverse ISP mapping directly from data. Early works such as CycleISP [47] and InvISP [40] enforce cycle consistency between RAW and RGB domains to regularize the ill-posed inversion. To make the model interpretable and parameters controllable, MBISPLD [10] replaces handcrafted ISP parameters in UPI [5] with learnable dictionaries trained end-to-end. ReRAW [3] employs a multi-head architecture to estimate RAW intensities. More recently, the generative approach RAW-Diffusion [33] formulates the task as a

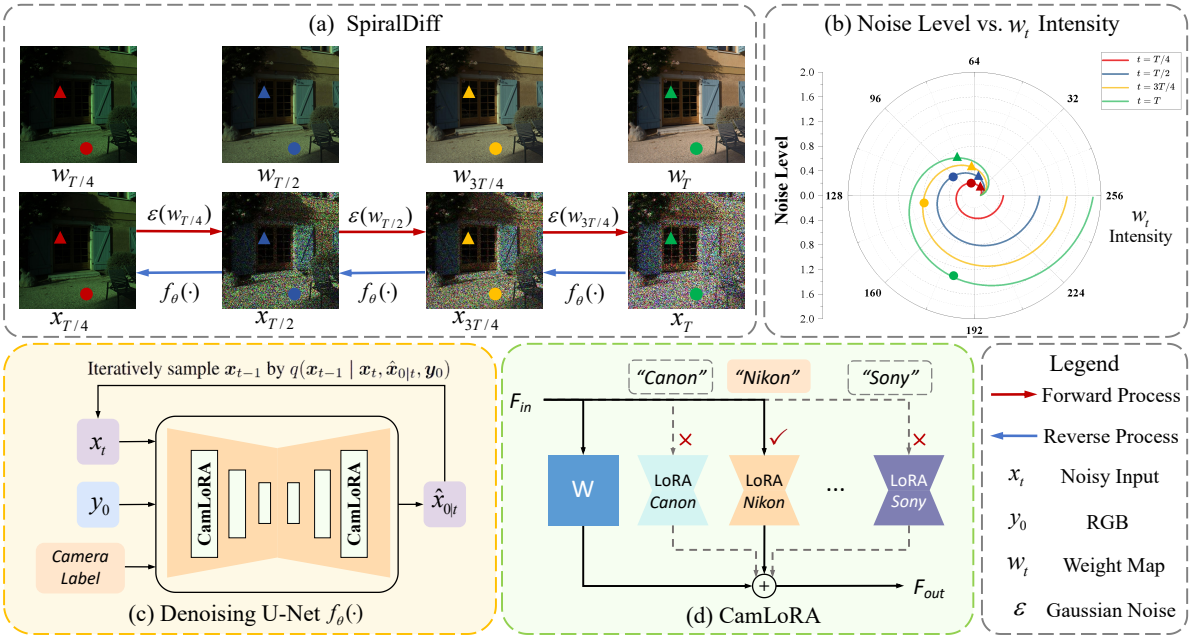


Figure 3. Overview of the proposed SpiralDiff with CamLoRA. (a) SpiralDiff introduces a signal-dependent noise schedule via a weight map set $\{\boldsymbol{w}_t\}_{t=1}^{T}$ that aligns with the RAW-to-RGB conversion process. The noise level depends on local pixel intensity and diffusion timestep t: darker regions (△) receive less noise, and brighter regions (◯) receive more. (b) The spiral structure visualizes how noise scales with $\boldsymbol{w}_t$ for different pixel intensities. (c) shows the framework of SpiralDiff. The noisy image $\boldsymbol{x}_t$, RGB image $\boldsymbol{y}_0$ and camera label are fed into denoising U-Net, which iteratively samples to refine the RAW output. (d) The camera label selects the camera-specific LoRA layer in CamLoRA, enhancing adaptation to each camera's characteristic.

conditional generation process using diffusion models [17]. However, existing methods yield limited results, especially in high-intensity regions, because they treat all pixels uniformly and fail to account for the fact that reconstruction difficulty varies significantly with local intensity. To address this limitation, we propose SpiralDiff for intensity-aware RGB-to-RAW conversion.

Diffusion Models. Diffusion models [17] have gained significant attention in generative modeling due to their remarkable ability to synthesize high-fidelity images. These models operate through two complementary processes: a forward diffusion process that gradually adds Gaussian noise to data until it resembles a standard normal distribution, and a reverse denoising process that iteratively reconstructs the original signal using a neural network trained to predict and remove noise. Owing to their strong generative capacity and flexibility in conditioning, diffusion models have been successfully adapted to various image restoration tasks, including image super-resolution [36, 38, 39, 50], low-light image enhancement [18, 21, 42], and high-dynamic-range imaging [9, 20]. By conditioning on low-quality inputs, these methods are capable of generating high-quality outputs with rich structural and textural details. Among recent advances, ResShift [45] accelerates the sampling process requiring only four steps while achieving competitive performance across multiple image restoration tasks such as super-resolution, image inpainting, and blind face restoration. Given its high efficiency and strong performance on various restoration tasks, ResShift presents a promising foundation for RGB-to-RAW conversion.

Low-Rank Adaptation. As a parameter-efficient fine-tuning strategy, Low-Rank Adaptation (LoRA) [19] decomposes weight updates into low-rank matrices, keeping the pretrained model frozen while optimizing only a small set of trainable parameters. Originally developed for large language models, LoRA has been successfully extended to computer vision tasks for domain adaptation [22, 34] or stylized image generation [4, 14, 30]. Recently, LoRA has shown promise in image restoration. UIR-LoRA [48] and LoRA-IR [2] employ LoRA in universal image restoration as a lightweight alternative to degradation prompts, while PiSA-SR [35] uses it for controllable super-resolution.

3. Methodology

Given an RGB image and its camera label, we aim to convert it into a RAW image using our proposed diffusion model. Our framework is presented in Fig. 3. The reverse process iteratively denoises a noisy RAW estimate,

conditioned on the input RGB image and camera label with model weights dynamically modulated by CamLoRA. In the following, we first outline the baseline diffusion formulation, then detail SpiralDiff and CamLoRA.

3.1. Preliminaries

ResShift [45] is a diffusion model designed for image restoration. Instead of sampling from the pure Gaussian noise, it embeds the low-quality image into the initial state as the prior information, then progressively denoises and transitions to the high-quality image. Based on its residual shifting design, ResShift significantly improves efficiency requiring only four sampling steps. Considering its performance on various restoration tasks, it can be adopted for RGB-to-RAW conversion.

For ease of representation, we denote the RGB image by $\boldsymbol{y}_0$, the RAW image by $\boldsymbol{x}_0$ and their residual by $\boldsymbol{e}_0$, i.e., $\boldsymbol{e}_0 = \boldsymbol{y}_0 - \boldsymbol{x}_0$. ResShift builds up a transition from $\boldsymbol{x}_0$ to $\boldsymbol{y}_0$ by gradually shifting the residual $\boldsymbol{e}_0$. The forward transition is designed as

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{x}_{t-1} + \alpha_t \boldsymbol{e}_0, \kappa^2 \alpha_t \boldsymbol{I}), \quad (1)$$

for $t = 1, 2, \ldots, T$, where the shifting sequence $\{\eta_t\}_{t=1}^T$ is predefined and $\alpha_t = \eta_t - \eta_{t-1}$, $\kappa > 0$ controls the overall noise level, and $\boldsymbol{I}$ is the identity matrix. From this recursive definition, the marginal distribution at step t can be derived as

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0, \boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{x}_0 + \eta_t \boldsymbol{e}_0, \kappa^2 \eta_t \boldsymbol{I}). \quad (2)$$

As t increases, the mean of $\boldsymbol{x}_t$ is shifting gradually from $\boldsymbol{x}_0$ to $\boldsymbol{y}_0$, while the variance (i.e. the noise level) grows proportionally to η_t. Based on Eq. 1 and Eq. 2, the corresponding reverse process is given by

$$\begin{aligned} &q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{y}_0) \\ &\quad = \mathcal{N}(\boldsymbol{x}_{t-1}; \frac{\eta_{t-1}}{\eta_t}\boldsymbol{x}_t + \frac{\alpha_t}{\eta_t}\boldsymbol{x}_0, \kappa^2 \frac{\eta_{t-1}}{\eta_t}\alpha_t \boldsymbol{I}). \end{aligned} \quad (3)$$

3.2. Spiral Diffusion

Although ResShift provides a strong foundation, it still faces challenges inherent to many existing methods: the difficulty of adapting reconstruction strategies across regions with varying pixel intensities. The reason is that commonly used isotropic Gaussian noise introduces uniform perturbations across all pixels, which fails to account for the fact that the reconstruction difficulty varies with local intensity characteristics. Fig. 2 illustrates the relationship between RGB pixel intensity and the residual magnitude between RGB and RAW images across the red, green, and blue channels. The x-axis represents RGB pixel intensity, while the y-axis shows the mean residual magnitude for each channel. The red, green, and blue curves trace these mean values, and the shaded bands around each mean line represent the standard deviation above and below the mean. As pixel intensity increases, the bands widen, indicating that the residual variance is signal dependent. Note that, when the RGB intensities approach saturation, both the mean residual and its variance shrink, since the RAW image itself is also approaching saturation. The main reason is that multiplicative ISP operations, such as digital gains, cause brighter regions to exhibit larger residuals. For dark areas, excessive noise injection will hinder accurate reconstruction. Conversely, in bright or even over-exposed regions, insufficient noise injection limits the model's generation ability. This discrepancy makes it hard to balance reconstruction across different intensity regions with a fixed noise schedule. Based on these observations, we propose an adaptive signal-dependent noise weighting strategy in the forward process of the diffusion model, called SpiralDiff.

Forward Process. We introduce a weight map $\boldsymbol{w}$ to modulate the noise variance in Eq.1. At timestep T, the local noise level of $\boldsymbol{x}_T$ is correlated with $\boldsymbol{y}_0$. Low-intensity regions are perturbed lightly so the model can better recover their details, whereas high-intensity regions receive stronger noise to accommodate their higher reconstruction uncertainty. This design aligns with the analysis discussed above.

However, the static weight map does not adapt to the progress of the conversion. As $t \to 0$, we expect a stable output, where the weight map $\boldsymbol{w}_0$ should correlate to $\boldsymbol{x}_0$, rather than $\boldsymbol{y}_0$. To address this, we extend the static weight map $\boldsymbol{w}$ into a time-varying weight map set $\{\boldsymbol{w}_t\}_{t=1}^T$, defined as

$$\boldsymbol{w}_t = \boldsymbol{x}_0 + \eta_t \boldsymbol{e}_0. \quad (4)$$

The weight map $\boldsymbol{w}_t$ is consistent with the mean term and it can be interpreted as the noise-free intermediate state between $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$, intuitively illustrated in Fig. 3 (a). Consequently, the noise level evolves along with the signal transformation process, enabling a smooth transition from $\boldsymbol{x}_0$ to $\boldsymbol{y}_0$. The transition process is then modeled as:

$$\begin{aligned} &q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{y}_0) \\ &\quad = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{x}_{t-1} + \alpha_t \boldsymbol{e}_0, \kappa^2(\eta_t \boldsymbol{w}_t^2 - \eta_{t-1}\boldsymbol{w}_{t-1}^2)\boldsymbol{I}), \end{aligned} \quad (5)$$

where $\boldsymbol{w}_t^2$ represents the element-wise square of $w_t(p)$ for each pixel p. By iteratively applying this transition, the marginal distribution at timestep t can be expressed as

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0, \boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{x}_0 + \eta_t \boldsymbol{e}_0, \kappa^2 \eta_t \boldsymbol{w}_t^2 \boldsymbol{I}). \quad (6)$$

Thus, the noise level at timestep t is spatially correlated with the weight map $\boldsymbol{w}_t$, which is dependent on the current transition state $\boldsymbol{x}_0 + \eta_t \boldsymbol{e}_0$.

Reverse Process. Building upon the forward process, the corresponding backward transition distribution can be derived from Eq. 5 and Eq. 6, denoted as:

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1}), \tag{7}$$

where the mean $\boldsymbol{\mu}_{t-1}$ and variance $\boldsymbol{\Sigma}_{t-1}$ are given by

$$\begin{aligned} \boldsymbol{\mu}_{t-1} &= \boldsymbol{\gamma}_t(\boldsymbol{x}_t - \alpha_t \boldsymbol{e_0}) + (1-\boldsymbol{\gamma}_t)(\boldsymbol{x}_0 + \eta_{t-1}\boldsymbol{e}_0) \\ \boldsymbol{\Sigma}_{t-1} &= \kappa^2 \boldsymbol{\gamma}_t(\eta_t \boldsymbol{w}_t^2 - \eta_{t-1}\boldsymbol{w}_{t-1}^2)\boldsymbol{I}, \end{aligned} \tag{8}$$

with a blending coefficient $\boldsymbol{\gamma}_t = \frac{\eta_{t-1}\boldsymbol{w}_{t-1}^2}{\eta_t \boldsymbol{w}_t^2}$.

Notably, this backward transition distribution remains structurally aligned with that of ResShift (Eq. 3). On the one hand, when the spatial noise weighting is disabled ($\boldsymbol{w}_t \equiv 1$), the expressions of $\boldsymbol{\mu}_{t-1}$ and $\boldsymbol{\Sigma}_{t-1}$ reduce exactly to those of ResShift, and SpiralDiff degenerates to ResShift. On the other hand, even in the general case, the reverse transition distribution in SpiralDiff preserves the same fundamental guidance mechanism as ResShift. Specifically, in both frameworks, the mean $\boldsymbol{\mu}_{t-1}$ is expressed as a convex combination of two guided components: a *noisy term*, representing the current corrupted state, and a *clean term*, representing the target structure to be recovered. These components are blended under the control of a time-dependent balancing coefficient, which governs their relative influence at each sampling step. In ResShift (Eq. 3), the noisy term is simply $\boldsymbol{x}_t$ and the clean term is $\boldsymbol{x}_0$. The mean term gradually converges toward $\boldsymbol{x}_0$ as t decreases, progressively refining the output image. In SpiralDiff, the noisy term corresponds to $\boldsymbol{x}_t - \alpha_t \boldsymbol{e}_0$, a backward-projected estimate of the intermediate state at step $t-1$ (derived from Eq. 5). The clean term corresponds to $\boldsymbol{x}_0 + \eta_{t-1}\boldsymbol{e}_0$, which represents a forward-advanced approximation targeting the same state $\boldsymbol{x}_{t-1}$ (Eq. 6). The interpolation of these two terms occurs naturally within the local neighborhood of $\boldsymbol{x}_{t-1}$, yielding a more stable sampling trajectory. Crucially, the balancing coefficient $\boldsymbol{\gamma}_t$ in SpiralDiff depends on the spatial weight map $\boldsymbol{w}_t$, enabling a pixel-adaptive blending between the noisy and clean components, in contrast to the scalar coefficient used in ResShift.

Following ResShift, a deep neural network $f_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t)$ with parameter $\boldsymbol{\theta}$ is optimized to predict the target $\boldsymbol{x}_0$. For the detailed derivation, overall pipeline, and loss function, please refer to the supplementary material.

3.3. Camera-aware Low-Rank Adaptation

RAW images exhibit strong camera-dependent characteristics. Training a single conversion model on mixed multi-camera data often leads to interference between devices, resulting in degraded reconstruction quality. To address this, we propose Camera-aware Low-Rank Adaptation (CamLoRA), a lightweight adaptation module that conditions the denoising network on camera identity. As shown in Fig. 3, CamLoRA treats the input camera label as a discrete control signal, so that camera-specific adaptations are isolated in the low-rank branches, leaving the shared backbone free to learn universal features.

CamLoRA works by augmenting each trainable weight matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$ in the backbone with a camera-specific low-rank update:

$$\mathbf{W}_i = \mathbf{W} + \Delta\mathbf{W}_i = \mathbf{W} + \mathbf{B}_i\mathbf{A}_i, \tag{9}$$

where $\mathbf{A}_i \in \mathbb{R}^{r\times k}$ and $\mathbf{B}_i \in \mathbb{R}^{d\times r}$ are low-rank matrices for camera i, and $r \ll \min(d,k)$. This design adds only $r(d+k)$ parameters per camera, ensuring high parameter efficiency. We apply CamLoRA to the query, key, value, and output projection matrices ($\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o$) in all Swin Transformer [26] layers of our backbone. During training, the shared base weights $\mathbf{W}$ are updated using all data to learn general features. In contrast, only the adapter $\Delta\mathbf{W}_i$ corresponding to the current camera label c_i is updated. Following standard LoRA practice [19], we initialize $\mathbf{B}_i$ to zero and $\mathbf{A}_i$ with random Gaussian noise to ensure stable optimization.

4. Experiments

4.1. Datasets

We conduct experiments on two publicly available datasets: the MIT-Adobe FiveK Dataset (FiveK) [6] and the Night Object Detection Dataset (NOD) [28]. FiveK comprises diverse scenes captured under various lighting and exposure conditions. NOD is a large-scale benchmark for low-light object detection in the RAW domain, with annotated images captured at night. Following [33], we select RAW images captured by Canon EOS 5D and Nikon D700 from the FiveK dataset containing 777 and 590 images, respectively, and randomly split them into train/test sets with an 85/15 ratio for each camera. For the NOD dataset, we use RAW images from Nikon D750 and Sony RX100 VII, resulting in 4.0k and 3.2k images, respectively, and adopt the official train-test split. All RAW files are processed with the *rawpy* library to generate the corresponding RGB images. Unlike prior works [3, 33], we use the as-shot white balance coefficients stored in the RAW metadata for ISP simulation instead of auto-estimated white balance, which provides a more faithful color rendering and is beneficial for camera-aware RGB-to-RAW conversion.

4.2. Implementation Details

Following ResShift, we adopt a U-Net architecture equipped with Swin Transformer [26] blocks. Regarding VQGAN [13], which compresses images into a lower resolution to reduce the computational cost for efficient diffusion, we remove it since it is trained on RGB images

Table 1. Performance comparison of different methods trained on separate and combined datasets, measured by PSNR(↑) and SSIM(↑). *+CamLoRA* denotes that CamLoRA is applied to the combined training setting. Best results under each setting are highlighted in **bold**.

Methods	Separate				Combined			
	FiveK Canon	FiveK Nikon	NOD Nikon	NOD Sony	FiveK Canon	FiveK Nikon	NOD Nikon	NOD Sony
CycleISP [47]	37.93 / 0.9913	40.18 / 0.9920	50.11 / 0.9985	46.57 / 0.9975	38.06 / 0.9911	38.86 / 0.9918	48.68 / 0.9984	46.12 / 0.9971
InvISP [40]	36.81 / 0.9814	34.30 / 0.9163	48.29 / 0.9954	44.76 / 0.9922	35.37 / 0.9828	36.60 / 0.9850	47.73 / 0.9948	44.43 / 0.9907
ReRAW [3]	35.08 / 0.9608	33.09 / 0.9265	48.25 / 0.9906	45.50 / 0.9788	32.70 / 0.9684	33.24 / 0.9648	46.39 / 0.9715	45.51 / 0.9817
RAW-Diffusion [33]	39.96 / 0.9890	39.68 / 0.9866	50.52 / 0.9954	47.31 / 0.9908	40.63 / 0.9889	40.35 / 0.9916	49.60 / 0.9950	46.80 / 0.9897
SpiralDiff	**42.82 / 0.9936**	**41.72 / 0.9925**	**53.64 / 0.9990**	**50.46 / 0.9980**	41.99 / 0.9931	43.19 / 0.9947	52.06 / 0.9987	49.52 / 0.9974
+ CamLoRA	—	—	—	—	**42.46 / 0.9934**	**43.82 / 0.9950**	**52.62 / 0.9988**	**50.08 / 0.9977**

Table 2. Quantitative comparison on the over-exposed test set.

Methods	FiveK Canon		NOD Nikon	
	PSNR	SSIM	PSNR	SSIM
CycleISP [47]	27.75	0.9833	35.86	0.9957
InvISP [40]	25.15	0.9699	36.66	0.9905
ReRAW [3]	26.36	0.9664	38.56	0.9691
RAW-Diffusion [33]	30.60	0.9739	40.05	0.9926
SpiralDiff	**31.10**	**0.9856**	**40.79**	**0.9973**

and such a latent representation is ill-suited for RAW data. Therefore, SpiralDiff operates directly in pixel space. We retain the original diffusion hyperparameters from ResShift, including the noise scale κ, the shifting schedule $\{\eta_t\}_{t=0}^{T}$, and the total number of timesteps $T = 4$, ensuring high inference efficiency. The CamLoRA rank is set to $r = 8$ to maintain parameter efficiency, introducing only 2.7% additional parameters ($1.05\,M$) across four camera-specific LoRA adapters.

Each RAW image is first normalized using its black level and white level, then linearly scaled to the range $[-1, 1]$. For the training stage, each RAW image is packed to four-channel *RGGB* format, then randomly cropped to the resolution 256×256, with a batch size set to 8. Its RGB counterpart is cropped to 512×512 to maintain spatial alignment. We use the Adam optimizer [23] with cosine learning rate annealing [27], starting from an initial learning rate of 1×10^{-4} and decaying to 1×10^{-5}. During evaluation, images are inferred in the full resolution to avoid block artifacts. All experiments are implemented in PyTorch [31] and run on a single NVIDIA GeForce RTX 3090 GPU.

4.3. Comparison with State-of-the-Art Methods

We compare our method with four state-of-the-art methods for RGB-to-RAW conversion: CycleISP [47], InvISP [40], ReRAW [3], and RAW-Diffusion [33]. For a fair comparison, we retrain and evaluate all methods using identical data splits. We adopt two settings: *separate* and *combined*. In the separate setting, a dedicated model is trained for each of the four cameras (FiveK Canon, FiveK Nikon, NOD Nikon, and NOD Sony). In the *combined* setting, all camera data is merged into one training set to train a unified model, where CamLoRA uses camera labels and assigns different labels to FiveK Nikon and NOD Nikon since they correspond to different camera sensors. All the methods are trained with the same patch size, except for ReRAW [3], for which we keep its original global-context design. For InvISP [40], we remove its white balance pre-processing on RAW images so that all methods operate on the same data, ensuring the fair comparison.

Quantitative Comparison. Quantitative results on the four camera-specific test sets are reported in Tab. 1, under the *Separate* and *Combined* training settings mentioned before. We evaluate conversion quality using PSNR and SSIM [37]. In the separate setting, our proposed SpiralDiff achieves the best performance across all cameras, outperforming RAW-Diffusion [33] by a large margin. This highlights the advantage of our signal-dependent noise scheme over RAW-Diffusion's use of isotropic Gaussian noise for modeling the RGB-to-RAW mapping. In the *combined* setting, the model trained on the combined dataset usually performs worse than models trained on separate datasets, since one model needs to deal with multiple camera sensors. As shown in Tab. 1, all datasets exhibit performance degradation except for FiveK Nikon. This exception may be attributed to the fact that the separate model trained on FiveK Nikon uses only 502 training images, while the combined dataset includes a significantly larger number of Nikon-style images due to the incorporation of 3,206 NOD Nikon images. Fortunately, when equipped with our proposed CamLoRA, SpiralDiff achieves a clear improvement in the combined setting, denoted as *+CamLoRA*. The results approach those of the separate setting. This demonstrates that our CamLoRA strategy can effectively adapt the model to different sensors according to the input camera label.

To further assess the generative performance of different methods, we construct an over-exposed (OE) test set by selecting images with higher proportions of saturated pixels. We select the top 20% of images from the FiveK dataset and the top 10% from the NOD dataset. The resulting OE test set includes 115 images: 24 from FiveK Canon, 18

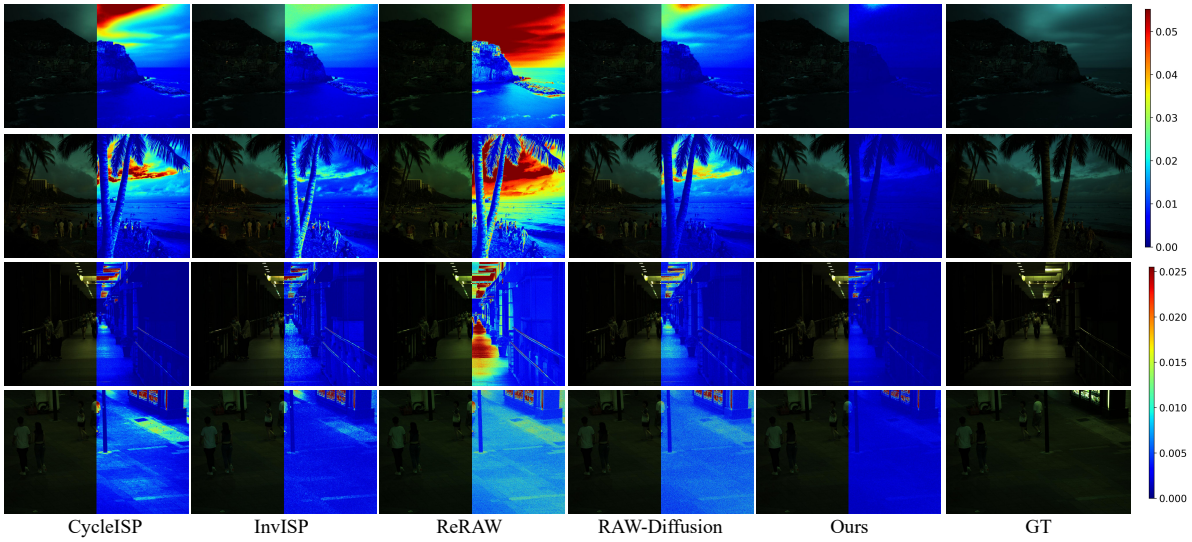


Figure 4. Qualitative comparison with state-of-the-art RGB-to-RAW conversion methods on the FiveK dataset (top two rows) and the NOD dataset (bottom two rows). For each result, the left half is the predicted RAW, and the right half is the error map. The proposed SpiralDiff shows better conversion results, especially in bright regions.

from FiveK Nikon, 40 from NOD Nikon, and 33 from NOD Sony. As shown in Tab. 2, the diffusion-based methods outperform previous methods, and our method outperforms the second best method (RAW-Diffusion) by more than 0.5 dB on PSNR. The full results are presented in the supplementary material.

Qualitative Comparison. Fig. 4 shows qualitative comparisons of reconstructed RAW images by methods trained on the combined dataset. While all methods exhibit higher errors in bright regions, our method suffers least from color distortion and produces the most visually pleasing results. This demonstrates that our method outperforms others in preserving image quality and effectively handling high-intensity regions.

4.4. Ablation Study

First, we verify the effectiveness of the proposed noise weighting strategy by replacing it with other variants. All models are trained in the separate setting. The baseline variant is the original ResShift [45], which utilizes isotropic Gaussian noise without adaptive weighting. The second variant directly uses the input RGB image ($\boldsymbol{y}_0$) to modulate the noise intensity. In contrast, our method uses the time-varying weight map sequence $\boldsymbol{w}_t$ to modulate the noise intensity. As shown in Tab. 3, our method achieves the best results, outperforming the baseline method by more than 1 dB on FiveK Canon dataset. Note that an unsuitable noise weighting strategy, namely $\boldsymbol{y}_0$, performs even worse than the baseline, suggesting that a static weight map cannot model the reconstruction difficulty throughout the sampling process. This further demonstrates that our proposed noise weighting scheme is suitable for the RGB-to-RAW conversion task. Moreover, we conduct a plug-in experiment. We replace the DDPM diffusion process in RAW-Diffusion [33] with our SpiralDiff, while keeping its network architecture and training protocol unchanged, resulting in a +1.57 dB PSNR improvement on the FiveK Canon dataset (see supplementary for details).

Second, we compare CamLoRA with a direct camera embedding approach [16]. It uses a set of trainable embeddings, one per camera, to provide camera-specific conditioning. Specifically, the embedding corresponding to the

Table 3. Ablations on the noise weighting strategies. The baseline uses spatially uniform noise. $\boldsymbol{y}_0$ ($\boldsymbol{w}_t$) means the noise is weighted according to the intensity of $\boldsymbol{y}_0$ ($\boldsymbol{w}_t$).

Variant	FiveK Canon		NOD Nikon	
	PSNR	SSIM	PSNR	SSIM
Baseline	41.40	0.9906	53.48	0.9989
$\boldsymbol{y}_0$	40.06	0.9912	53.42	**0.9990**
$\boldsymbol{w}_t$	**42.82**	**0.9936**	**53.64**	**0.9990**

Table 4. Ablations on three conditioning strategies: *Uncond.* (no camera info), *Embed.* (camera embedding), and *CamLoRA*.

Variant	FiveK Canon		NOD Nikon	
	PSNR	SSIM	PSNR	SSIM
Uncond.	41.99	0.9931	52.06	0.9987
Embed.	40.97	0.9917	51.44	0.9984
CamLoRA	**42.46**	**0.9934**	**52.62**	**0.9988**

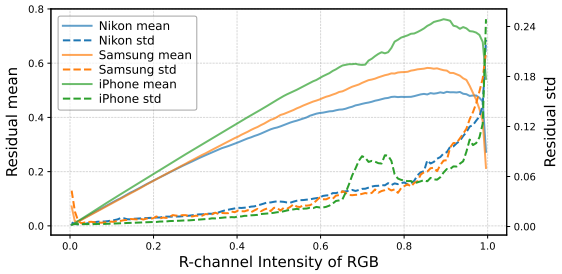


Figure 5. Residual-intensity relationship on rawpy ISP (FiveK Nikon) and real ISP (iPhone and Samsung).

input camera label is injected at the same location as the time embedding to modulate features. We train a unified model with this approach, denoted as *Embed*, and report results in Tab. 4. Interestingly, it performs worse than even the unconditional baseline. In contrast, CamLoRA achieves substantial gains by applying low-rank updates, enabling effective camera adaptation with only a small number of parameters. In addition, we investigate the effect of CamLoRA rank: with only $r = 4$, it already achieves a notable improvement over the unconditional baseline. Please refer to the supplementary material for detailed results.

CamLoRA enables the unified model to adapt to new cameras by training a new camera-specific LoRA branch while keeping shared parameters frozen, allowing the model to effectively reuse the rich priors learned during pre-training. This capability is particularly beneficial in few-shot settings. As shown in Tab. 5, with only 1 or 5 training images from the SID Sony dataset [7], CamLoRA fine-tuning achieves higher PSNR and SSIM than training a model from scratch, demonstrating superior adaptability and data efficiency.

4.5. Evaluation on Real-ISP Dataset

We further evaluate SpiralDiff on a real-ISP dataset from NTIRE Challenge [11], where RGB images are rendered by smartphone ISPs (iPhone X and Samsung S9). The residual-intensity relationship shows a similar increasing uncertainty trend to the rawpy-based setting (Fig. 5). Despite variations in the exact curve shapes across ISPs, this shared behavior is consistent and forms the basis of our design, which is therefore not tied to any specific ISP. Replacing the original DDPM-based diffusion process in RAW-Diffusion with SpiralDiff while keeping the network architecture and training protocol unchanged yields consistent improvements on both subsets (Tab. 6), showing that SpiralDiff remains effective on real-ISP data.

5. Applications

A practical application of RGB-to-RAW conversion is object detection. Training detectors directly on RAW images can potentially exploit richer sensor information and skip the ISP processing to further reduce latency. With our method, existing large-scale RGB object detection datasets can be converted into the RAW domain with their annotations directly reused, avoiding the high cost of collecting and labeling RAW data. We evaluate whether synthesized RAW data improves RAW-domain detection on the NOD benchmark in low-data settings. As shown in Tab. 7, training with only 100 real RAW images yields limited performance. Augmenting the training set with our synthesized RAW images from BDD100K (BDD) [43] and Cityscapes (CS) [12] consistently improves the results. This highlights the value of leveraging large-scale RGB datasets and reduces the need for costly manual RAW annotation.

Table 5. Few-shot RGB-to-RAW conversion on the SID Sony dataset. We compare training from scratch with CamLoRA adaptation under 1-shot and 5-shot settings.

Variant	1-shot		5-shot	
	PSNR	SSIM	PSNR	SSIM
Scratch	39.83	0.9895	43.40	0.9951
CamLoRA	**42.85**	**0.9942**	**43.92**	**0.9954**

Table 6. Plug-in experiment on the real-ISP dataset.

Method	iPhone	Samsung
RAW-Diffusion (DDPM)	22.74 / 0.7503	30.28 / 0.9140
RAW-Diffusion (SpiralDiff)	**30.11 / 0.8810**	**34.07 / 0.9322**

Table 7. Comparison of object detection results on the NOD test set with different training datasets.

Dataset	NOD Nikon			NOD Sony		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
RGB-only	19.1	37.0	18.5	19.7	38.3	18.0
RAW-only	18.4	35.1	17.0	17.6	35.7	15.1
RAW + CS-RAW	24.5	46.2	23.5	25.9	49.3	24.5
RAW + BDD-RAW	**26.7**	**49.5**	**25.1**	**29.0**	**55.1**	**26.5**

6. Conclusion

In this work, we propose SpiralDiff, a diffusion-based framework tailored for RGB-to-RAW conversion. SpiralDiff incorporates a signal-dependent noise schedule that adapts reconstruction based on pixel intensity, allowing the model to handle both low-intensity and high-intensity regions effectively and in a spatially adaptive manner. We further introduce CamLoRA, a parameter-efficient camera-aware adaptation module that enables a single model to adapt to multiple cameras. Experiments demonstrate that our method achieves state-of-the-art RGB-to-RAW conversion performance. In addition, the generated RAW images improve downstream object detection performance.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1692–1700, 2018. 1

[2] Yuang Ai, Huaibo Huang, and Ran He. Lora-ir: taming low-rank experts for efficient all-in-one image restoration. *arXiv preprint arXiv:2410.15385*, 2024. 3

[3] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11833–11843, 2025. 2, 5, 6

[4] Shubhankar Borse, Shreya Kadambi, Nilesh Pandey, Kartikeya Bhardwaj, Viswanath Ganapathy, Sweta Priyadarshi, Risheek Garrepalli, Rafael Esteves, Munawar Hayat, and Fatih Porikli. Foura: Fourier low-rank adaptation. *Advances in Neural Information Processing Systems*, 37:71504–71539, 2024. 3

[5] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11036–11045, 2019. 2

[6] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. IEEE, 2011. 5

[7] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018. 1, 8

[8] Hongjun Chen, Wencheng Han, Huan Zheng, and Jianbing Shen. Rawmamba: Unified srgb-to-raw de-rendering with state space model. *arXiv preprint arXiv:2411.11717*, 2024. 2

[9] Zixuan Chen, Yujin Wang, Xin Cai, Zhiyuan You, Zheming Lu, Fan Zhang, Shi Guo, and Tianfan Xue. Ultrafusion: Ultra high dynamic imaging using exposure fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16111–16121, 2025. 3

[10] Marcos V Conde, Steven McDonagh, Matteo Maggioni, Ales Leonardis, and Eduardo Pérez-Pellitero. Model-based image signal processors via learnable dictionaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 481–489, 2022. 2

[11] Marcos V. Conde, Radu Timofte, Radu Berdan, Beril Besbinar, Daisuke Iso, Pengzhou Ji, Xiong Dun, Zeying Fan, Chen Wu, Zhansheng Wang, Pengbo Zhang, Jiazi Huang, Qinglin Liu, Wei Yu, Shengping Zhang, Xiangyang Ji, Kyungsik Kim, Minkyung Kim, Hwalmin Lee, Hekun Ma, Huan Zheng, Yanyan Wei, Zhao Zhang, Mengyuan Sun, Jing Fang, Meilin Gao, Xiang Yu, Shangbin Xie, Huanjing Yue, and Jingyu Yang. Raw image reconstruction from rgb on smartphones. ntire 2025 challenge report. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 01–15, 2025. 8

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 8

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 5

[14] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, pages 181–198. Springer, 2024. 3

[15] Shani Gamrian, Hila Barel, Feiran Li, Masakazu Yoshimura, and Daisuke Iso. Beyond rgb: Adaptive parallel processing for raw object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5547–5557, 2025. 1

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 7

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[18] Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. *Advances in Neural Information Processing Systems*, 36:79734–79747, 2023. 3

[19] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3, 5

[20] Tao Hu, Qingsen Yan, Yuankai Qi, and Yanning Zhang. Generating content for hdr deghosting from frequency view. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25732–25741, 2024. 3

[21] Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023. 3

[22] Samar Khanna, Medhanie Irgau, David B. Lobell, and Stefano Ermon. ExPLoRA: Parameter-efficient extended pre-training to adapt vision transformers under domain shifts. In *Forty-second International Conference on Machine Learning*, 2025. 3

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[24] Leyi Li, Huijie Qiao, Qi Ye, and Qinmin Yang. Metadata-based raw reconstruction via implicit neural functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18196–18205, 2023. 2

[25] 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 Computer Vision and Pattern Recognition Conference*, pages 8859–8868, 2025. 1

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5

[27] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[28] Igor Morawski, Yu-An Chen, Yu sheng Lin, Shusil Dangi, Kai He, and Winston H. Hsu. Genisp: Neural isp for low-light machine cognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 629–638, 2022. 5

[29] Seonghyeon Nam, Abhijith Punnappurath, Marcus A Brubaker, and Michael S Brown. Learning srgb-to-raw-rgb de-rendering with content-aware metadata. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17704–17713, 2022. 2

[30] Ziheng Ouyang, Zhen Li, and Qibin Hou. K-lora: Unlocking training-free fusion of any subject and style loras. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13041–13050, 2025. 3

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[32] Abhijith Punnappurath and Michael S Brown. Spatially aware metadata for raw reconstruction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 218–226, 2021. 2

[33] 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)*, pages 8431–8443. IEEE, 2025. 2, 5, 6, 7

[34] Linus Scheibenreif, Michael Mommert, and Damian Borth. Parameter efficient self-supervised geospatial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27841–27851, 2024. 3

[35] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2333–2343, 2025. 3

[36] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024. 3

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

[38] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25456–25467, 2024. 3

[39] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13095–13105, 2023. 3

[40] Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6287–6296, 2021. 2, 6

[41] Ruikang Xu, Chang Chen, Jingyang Peng, Cheng Li, Yibin Huang, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Toward raw object detection: A new benchmark and a new model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13384–13393, 2023. 1

[42] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12302–12311, 2023. 3

[43] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 8

[44] Huanjing Yue, Cong Cao, Lei Liao, Ronghe Chu, and Jingyu Yang. Supervised raw video denoising with a benchmark dataset on dynamic scenes. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2298–2307, 2020. 1

[45] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36:13294–13307, 2023. 1, 2, 3, 4, 7

[46] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by residual shifting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[47] 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)*, pages 2693–2702, 2020. 2, 6

[48] Cheng Zhang, Dong Gong, Jiumei He, Yu Zhu, Jinqiu Sun, and Yanning Zhang. Uir-lora: Achieving universal image restoration through multiple low-rank adaptation. *arXiv preprint arXiv:2409.20197*, 2024. 3

[49] Feng Zhang, Bin Xu, Zhiqiang Li, Xinran Liu, Qingbo Lu, Changxin Gao, and Nong Sang. Towards general low-light raw noise synthesis and modeling. In *Proceedings of*

the IEEE/CVF international conference on computer vision, pages 10820–10830, 2023. 1

[50] Leheng Zhang, Weiyi You, Kexuan Shi, and Shuhang Gu. Uncertainty-guided perturbation for image super-resolution diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17980–17989, 2025. 3