ISPDiffuser: Learning RAW-to-sRGB Mappings with Texture-Aware Diffusion Models and Histogram-Guided Color Consistency

Yang Ren[1,4,*], **Hai Jiang**[1,4,*], **Menglong Yang**[1,2,†], **Wei Li**[1,2], **Shuaicheng Liu**[3,4,†]
[1]School of Aeronautics and Astronautics, Sichuan University
[2]Key Laboratory of Advanced Spatial Mechanism and Intelligent Spacecraft, Sichuan University
[3]University of Electronic Science and Technology of China
[4]Megvii Technology
{renyang@stu.,jianghai@stu.,mlyang@,li.wei@}scu.edu.cn, liushuaicheng@uestc.edu.cn

Abstract

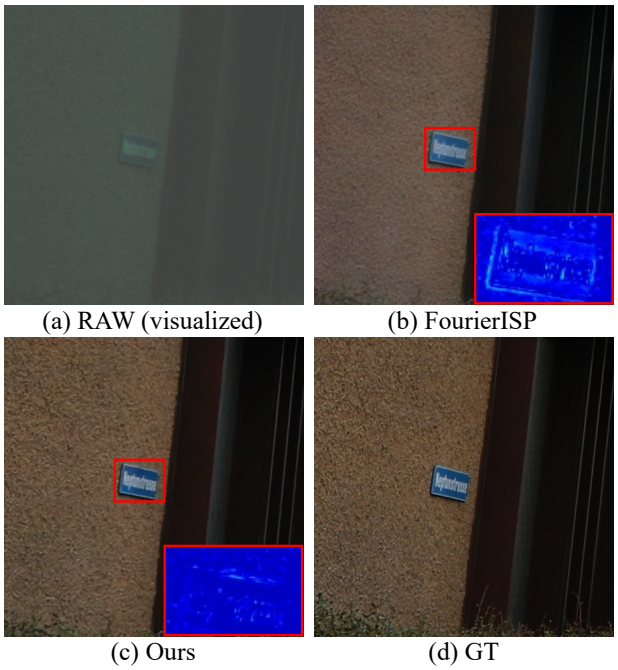
RAW-to-sRGB mapping, or the simulation of the traditional camera image signal processor (ISP), aims to generate DSLR-quality sRGB images from raw data captured by smartphone sensors. Despite achieving comparable results to sophisticated handcrafted camera ISP solutions, existing learning-based methods still struggle with detail disparity and color distortion. In this paper, we present ISPDiffuser, a diffusion-based decoupled framework that separates the RAW-to-sRGB mapping into detail reconstruction in grayscale space and color consistency mapping from grayscale to sRGB. Specifically, we propose a texture-aware diffusion model that leverages the generative ability of diffusion models to focus on local detail recovery, in which a texture enrichment loss is further proposed to prompt the diffusion model to generate more intricate texture details. Subsequently, we introduce a histogram-guided color consistency module that utilizes color histogram as guidance to learn precise color information for grayscale to sRGB color consistency mapping, with a color consistency loss designed to constrain the learned color information. Extensive experimental results show that the proposed ISPDiffuser outperforms state-of-the-art competitors both quantitatively and visually.

Code — https://github.com/RenYangSCU/ISPDiffuser

Introduction

The image signal processor (ISP) pipeline is a fundamental process for converting RAW data captured by camera sensors into sRGB images that are perceivable by the human eye. Traditionally, the ISP pipeline in cameras consists of a series of discrete processes, including demosaicing, denoising, white balance, gamma correction, and color correction, each of which necessitates complex and extensive manual parameter tuning (Ramanath et al. 2005).

With the rapid advancement in mobile photography, smartphones have become the dominant photography tool due to their convenience and portability. However, due to limitations in aperture and sensor size, mobile devices generally produce images with lower quality than those captured by DSLR cameras. To address this disparity, there has been growing interest in developing deep learning-based ISP models that can convert RAW data captured by mobile sensors into sRGB images with DSLR-like quality (Ignatov et al. 2021a). Existing deep ISP solutions focus on either compensating for the misalignment caused by different capture equipment of the training pairs (Zhang et al. 2021) or treating the RAW-to-sRGB mapping solely as a color mapping task (He et al. 2024b; Ignatov, Van Gool, and Timofte 2020a; Dai et al. 2020). Despite the remarkable progress, a notable limitation persists: most of these models are based on convolutional neural networks (CNNs), which are naturally limited by their inherent locality restriction, leading to



(a) RAW (visualized) (b) FourierISP (c) Ours (d) GT

Figure 1: Visual comparison with the previous state-of-the-art method FourierISP (He et al. 2024b). Our approach exhibits better local detail reconstruction (the red boxes show the content difference between generated images and GT images) and global color consistency mapping capabilities.

*These authors contributed equally.
†Corresponding authors.

local detail disparity and global color distortion. As shown in Fig. 1(b) and (d), the previous state-of-the-art method FourierISP (He et al. 2024b) produces blurred details and color distortion to degrade visual quality.

Recently, generative model-based methods (Wang, She, and Ward 2021; He et al. 2024a) have emerged as promising approaches for various low-level vision tasks to achieve better perceptual quality, where diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) have gained significant attention due to their impressive generative capabilities and advantages over previous generative models, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), by avoiding instability and mode-collapse problems. DiffRAW (Yi et al. 2024) proposed a diffusion-based framework for RAW-to-sRGB mappings that utilizes the sRGB images produced by LiteISPNet (Zhang et al. 2021) as the color-position preserving condition, which however would constrain the learned color and detail distributions similar to LiteISPNet. Moreover, diffusion models, despite their superior high-frequency detail information generation capability, are usually biased in low-frequency information generation such as color and illumination (Ning et al. 2024), which presents challenges in handling local detail reconstruction and global color mapping simultaneously for RAW-to-sRGB mappings.

To this end, we proposed a diffusion-based decoupled framework, named ISPDiffuser, which separates RAW-to-sRGB mappings into detail reconstruction in grayscale space and color consistency mapping from grayscale to sRGB to achieve visually satisfactory results. Specifically, we propose a texture-aware diffusion model (TADM) that leverages the generative ability of diffusion models to focus on details reconstruction without concern for color information, in which the encoded feature of the grayscale version of the corresponding sRGB image is taken as input to perform diffusion processes with the guidance of encoded RAW feature and newly proposed texture enrichment loss. Subsequently, we present a histogram-guided color consistency module (HCCM) that employs the color histogram (Han and Ma 2002) as guidance to learn precise color information to transform the gray feature generated by TADM into the sRGB feature with consistent colors as DSLR images, with a color consistency loss formulated to supervise the learned color information. As shown in Fig. 1(c), our method is capable of generating sRGB images with more distinct details and vivid color, being more visually pleasant. Extensive experiments demonstrate that our method outperforms existing state-of-the-art competitors quantitatively and visually.

Our contributions can be summarized as follows:

- We propose a diffusion-based decoupled framework, dubbed ISPDiffuser, which performs detail reconstruction and color consistency mapping separately to achieve visually satisfactory RAW-to-sRGB mappings.
- We propose a texture-aware diffusion model that leverages the generative ability of diffusion models to focus on detail reconstruction, as well as a histogram-guided color consistency module that utilizes the color histogram to learn stable global color mapping.
- Extensive experiments demonstrate that our method outperforms existing state-of-the-art competitors and is capable of generating images with better perceptual quality.

Related Work

Traditional ISP. The traditional Image Signal Processing (ISP) pipeline includes denoising (Dabov et al. 2007; Zhang et al. 2017), demosaicing (Gharbi et al. 2016), white balancing (Cheng et al. 2015), color correction (Kwok et al. 2013; Rizzi, Gatta, and Marini 2003), gamma correction, and tone mapping (Rana et al. 2019; Liu et al. 2021, 2022), aiming to convert RAW images to high-quality sRGB images. To date, camera systems typically utilize manual ISP workflows, necessitating experienced engineers to adjust numerous parameters to achieve satisfactory image quality. Sequential execution of multiple subtasks on proprietary hardware can result in accumulated errors. Compared to digital single-lens reflex (DSLR) cameras, mobile devices face hardware limitations that hinder their ability to capture images of comparable quality to those produced by professional DSLRs.

Learning-based ISP. With the development of deep learning, learning-based approaches (Schwartz, Giryes, and Bronstein 2018; Zamir et al. 2020; Xing, Qian, and Chen 2021) have intensified to address the hardware constraints and manual adjustments required by traditional ISP methods, which most of these efforts aim to capture the ISP process under controlled conditions where RAW and sRGB training pairs are captured from the same device. To this end, Ignatov *et al.* (Ignatov, Van Gool, and Timofte 2020a) set a new challenge by tackling the RAW-to-sRGB mapping problem caused by dual-device capture, which involves addressing spatial misalignment and resolution variations. To tackle these challenges, AWNet (Dai et al. 2020) explored the potential of wavelet transformation and non-local attention mechanisms in the ISP pipeline. LiteISPNet (Zhang et al. 2021) created a global color mapping module to address color inconsistency issues and used an aligned loss to compute optical flow between the predicted sRGB images and ground truth. FourierISP (He et al. 2024b) employed the Fourier prior to separating and refining the color and structural representations. TransformISP (Shekhar Tripathi et al. 2022) used a color-conditional ISP network with a masked aligned loss to refine color and tone mappings.

Diffusion Models in Low-Level Vision. Diffusion models (Sohl-Dickstein et al. 2015) are generative models that employ stochastic diffusion processes based on thermodynamics. To date, diffusion models have been widely used in various low-level vision tasks owing to their powerful generative capabilities such as image editing (Kawar et al. 2023; Zhang et al. 2023), image restoration (Jiang et al. 2023, 2025; Lugmayr et al. 2022; Kawar et al. 2022), and image alignment (Luo et al. 2024; Li et al. 2024; Zhou et al. 2024), while their application to RAW-to-sRGB mappings still needs to be explored. Recently, DiffRAW (Yi et al. 2024) introduced a diffusion-based framework for RAW-to-sRGB mappings that utilizes the sRGB images produced by LiteISPNet (Zhang et al. 2021) as the color-position preserving condition, which would result in the learned color and detail distributions be similar as LiteISPNet. In this work,

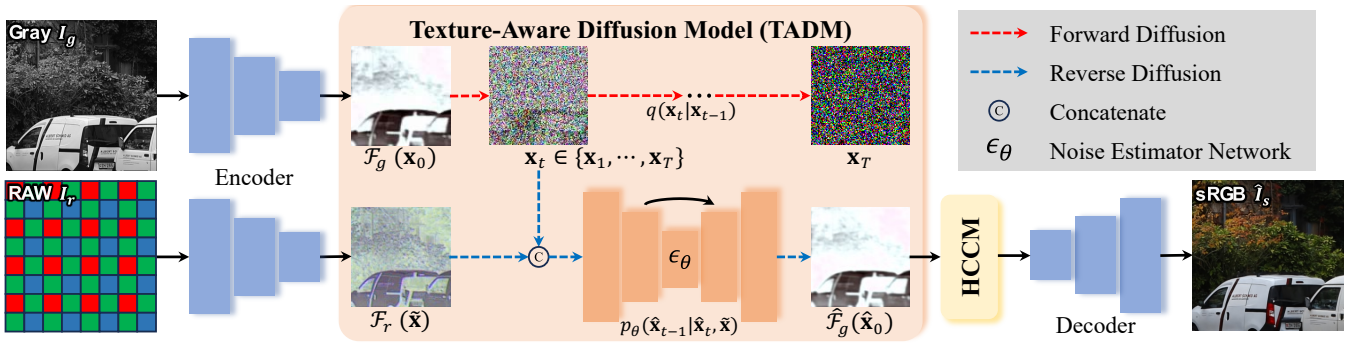


Figure 2: The overall pipeline of our proposed framework. We first employ an encoder $\mathcal{E}(\cdot)$ to convert RAW image I_r and grayscale version I_g of the sRGB image into latent space denoted as $\mathcal{F}_r$ and $\mathcal{F}_g$. The encoded feature $\mathcal{F}_g$ is taken as the input of the proposed texture-aware diffusion model (TADM) to perform the forward diffusion process. With the guidance of the raw feature $\mathcal{F}_r$, we generate the reconstructed gray feature $\hat{\mathcal{F}}_g$ from the noised tensor $\mathbf{x}_t$ during training, which is replaced by randomly sampled Gaussian noise $\hat{\mathbf{x}}_T$ during inference. Finally, we utilize the proposed histogram-guided color consistency module (HCCM) to colorize the generated $\hat{\mathcal{F}}_g$ and subsequently send it to a decoder $\mathcal{D}(\cdot)$ to produce the final sRGB result $\hat{I}_s$.

we proposed a new diffusion-based framework that decouples the RAW-to-sRGB task into grayscale detail reconstruction and color consistency mapping from grayscale to sRGB to achieve more visually satisfactory results.

Methodology

Overview

The overall pipeline of our method is illustrated in Fig. 2. Our approach decouples the RAW-to-sRGB mappings into grayscale detail reconstruction and color-consistent mapping from grayscale to sRGB, aiming to transform the RAW images into high-quality sRGB images. Given a RAW image $I_r \in \mathbb{R}^{H\times W\times 1}$ and the grayscale version $I_g \in \mathbb{R}^{H\times W\times 1}$ of the corresponding sRGB image, we adopt an encoder $\mathcal{E}(\cdot)$, which consists of k cascaded residual blocks where each block downsamples the input by a scale of 2, to transform the input images into latent space denoted as $\mathcal{F}_r \in \mathbb{R}^{\frac{H}{2^k}\times\frac{W}{2^k}\times c}$ and $\mathcal{F}_g \in \mathbb{R}^{\frac{H}{2^k}\times\frac{W}{2^k}\times c}$. Then, we introduce a texture-aware diffusion model (TADM) which leverages the generative ability of diffusion models to transform the RAW feature into the content informative grayscale feature $\hat{\mathcal{F}}_g$ with the guidance of the newly proposed texture preservation loss. Subsequently, we propose a histogram-guided color consistency module (HCCM) to colorize the grayscale feature $\hat{\mathcal{F}}_g$ into the sRGB feature $\hat{\mathcal{F}}_s$, which will be sent to a decoder $\mathcal{D}(\cdot)$ for reconstruction to produce the final sRGB image $\hat{I}_s$.

Texture-Aware Diffusion Model

RAW-to-sRGB mappings have two critical concerns: local detail recovery and global color mapping. Recently, diffusion models have gained attention for their impressive generative ability, while encountering low-frequency generative bias, such as color and exposure (Jiang et al. 2023; Ning et al. 2024). To this end, we present a texture-aware diffusion model (TADM) that focuses on reconstructing the details of sRGB images without attending to the low-frequency color mapping. Our approach follows standard diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) that perform forward diffusion and reverse diffusion processes to generate results with fine-grained details.

Forward Diffusion. We take the gray feature $\mathcal{F}_g$ as the initial input $\mathbf{x}_0$ for the forward diffusion process, in which a predefined variance schedule $\{\beta_1, \beta_2, ..., \beta_T\}$ is employed to progressively convert $\mathbf{x}_0$ into Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ over T steps, which can be formulated as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

where $\mathbf{x}_t$ indicates the noisy data at time-step $t \in [0, T]$. With parameter renormalization, we can obtain $\mathbf{x}_t$ directly from the input $\mathbf{x}_0$ and thereby simplify Eq.(1) into a closed expression as $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$, where $\alpha_t = 1-\beta_t$, $\bar{\alpha}_t = \prod_{i=0}^{t}\alpha_i$, and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Reverse Diffusion. The reverse diffusion process learns the non-Markovian forward processes that gradually denoise a randomly sampled Gaussian noise $\hat{\mathbf{x}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ into a sharp result $\hat{\mathbf{x}}_0$ conforming to the target data distribution. To strengthen the controllability of the generation procedure, we apply the conditional mechanism (Chung, Sim, and Ye 2022) to improve the fidelity of the reconstructed results conditioned on the encoded RAW feature $\mathcal{F}_r$, denoted as $\tilde{\mathbf{x}}$. The reverse diffusion process can be formulated as:

$$p_\theta(\hat{\mathbf{x}}_{t-1}|\hat{\mathbf{x}}_t, \tilde{\mathbf{x}}) = \mathcal{N}(\hat{\mathbf{x}}_{t-1}; \boldsymbol{\mu}_\theta(\hat{\mathbf{x}}_t, \tilde{\mathbf{x}}, t), \sigma_t^2\mathbf{I}), \quad (2)$$

where $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ is the variance and $\boldsymbol{\mu}_\theta(\hat{\mathbf{x}}_t, \tilde{\mathbf{x}}, t) = \frac{1}{\sqrt{\alpha_t}}(\hat{\mathbf{x}}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\hat{\mathbf{x}}_t, \tilde{\mathbf{x}}, t))$ is the mean value.

In the training phase, instead of optimizing the parameters θ of the network $\boldsymbol{\epsilon}_\theta$ to promote the estimated noise vector close to Gaussian noise, we follow (Luo et al. 2024; Li et al. 2024) to generate the sharp gray feature $\hat{\mathbf{x}}_0$, i.e., $\hat{\mathcal{F}}_g$ and employ the content loss $\mathcal{L}_{con}$ for optimization as:

$$\mathcal{L}_{con} = ||\hat{\mathbf{x}}_0 - \mathbf{x}_0||_2, \quad (3)$$

where $\hat{\mathbf{x}}_0$ is estimated from the disturbed noise data as:

$$\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, \tilde{\mathbf{x}}, t\right)\right). \quad (4)$$

Furthermore, we introduce a texture enrichment loss $\mathcal{L}_{tel}$ to enable the reconstructed features to contain detailed texture information similar to the original input. Specifically, we employ the traditional Canny edge detector to extract the corresponding texture maps of the generated feature $\hat{\mathcal{F}}_g$ and the original gray feature $\mathcal{F}_g$ before non-maximum suppression, denoted as $\hat{\mathbf{T}}_g = \text{Canny}(\hat{\mathcal{F}}_g)$ and $\mathbf{T}_g = \text{Canny}(\mathcal{F}_g)$. Thus, the $\mathcal{L}_{tel}$ aims to constrain the texture similarity to prompt the diffusion model to generate $\hat{\mathcal{F}}_g$ with more intricate texture details as:

$$\mathcal{L}_{tel} = ||\hat{\mathbf{T}}_g - \mathbf{T}_g||_1. \tag{5}$$

Overall, the object function used to optimize our TADM is rewritten as $\mathcal{L}_{diff} = \mathcal{L}_{con} + \lambda_1 \mathcal{L}_{tel}$. During inference, we derive the restored feature $\hat{\mathcal{F}}_g$ from the learned distribution through the reverse diffusion process with the implicit sampling strategy (Song, Meng, and Ermon 2020).

Histogram-Guided Color Consistency Model

The Bayer filter array (BFA) captures color information by processing specific color arrangements within individual channels, which makes the RAW to sRGB transformation suffer from color disparities and unstable color mapping. To this end, with our decoupled framework, we introduce a histogram-guided color consistency module (HCCM) that utilizes the color histogram (Han and Ma 2002) as guidance to transform the gray feature generated by our TADM to sRGB feature with vivid color, as illustrated in Fig. 3.

Specifically, the RAW feature $\mathcal{F}_r$ is taken as input to our designed color histogram predictor $\mathcal{CHP}(\cdot)$ to predict the color histogram $\mathcal{H} \in \mathcal{R}^{N \times 256}$ of sRGB distribution, where $N = 3$ indicates the number of color channels in sRGB space (i.e., R, G, and B) and 256 aligns with the range of pixel values. However, since color histogram primarily describes the proportion of different colors across the entire image without accounting for spatial arrangement. Therefore, we utilize the RAW feature along with $\mathcal{H}$ to extract the position-specific color feature $\mathcal{F}_c$ as:

$$\mathcal{F}_c = \text{Conv}(\mathcal{H}) \times \mathcal{F}_r^{'}, \mathcal{H} = \mathcal{CHP}(\mathcal{F}_r), \tag{6}$$

where $\mathcal{F}_r^{'}$ is the reshaped RAW feature to satisfy dimension alignment and $\times$ denotes the matrix multiplication. Subsequently, we adopt a cross-attention layer to leverage the estimated position-specific color feature $\mathcal{F}_c$ to colorize the grayscale feature into sRGB feature $\hat{\mathcal{F}}_s$ enriched with detailed information and consistent color, in which the $\mathcal{F}_c$ is taken as query vector q while key k and value v vectors are calculated from original grayscale feature $\hat{\mathcal{F}}_g$.

To facilitate the $\mathcal{CHP}(\cdot)$ to predict more accurate color histograms for grayscale to sRGB transformation, we design a color consistency loss $\mathcal{L}_{ccl}$ to optimize the estimated $\mathcal{H}$ align with the color histogram $\mathcal{H}_s$ of encoded ground-truth sRGB feature $\mathcal{F}_s$, which is formulated as:

$$\mathcal{L}_{ccl} = ||\mathcal{H} - \mathcal{H}_s||_2. \tag{7}$$

Moreover, a feature loss $\mathcal{L}_{fea}$ is also adopted to constrain the reconstructed sRGB feature as:

$$\mathcal{L}_{fea} = ||\hat{\mathcal{F}}_s - \mathcal{F}_s||_2. \tag{8}$$

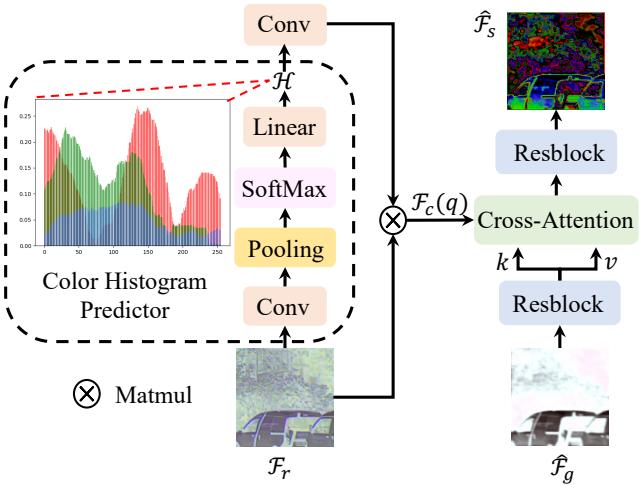


Figure 3: The detailed architecture of our proposed histogram-guided color consistency module.

Overall, the objective function used to optimize the HCCM is formulated as $\mathcal{L}_{hccm} = \mathcal{L}_{fea} + \lambda_2 \mathcal{L}_{ccl}$.

Network Training

Our approach employs a two-stage training strategy. In the first stage, we use paired RAW-sRGB images to train the encoder $\mathcal{E}(\cdot)$ and decoder $\mathcal{D}(\cdot)$, while freezing the parameters of the diffusion model and HCCM. The encoder and decoder are optimized with the content loss $\mathcal{L}_{stage1}$ as:

$$\mathcal{L}_{stage1} = ||I - \mathcal{D}(\mathcal{E}(I))||_2, \tag{9}$$

where I denotes the input RAW, sRGB, and grayscale images. In the second stage, we optimize the TADM and HCCM simultaneously through $\mathcal{L}_{stage2} = \mathcal{L}_{diff} + \mathcal{L}_{hccm}$ while freezing the parameters of the encoder and decoder.

Experiments

Experimental Settings

Implementation Details. We implement the proposed method with PyTorch on four NVIDIA RTX 2080Ti GPUs, where the batch size and patch size are set to 16 and 256×256. We employ the Adam optimizer (Kingma and Ba 2014) for optimization with the initial learning rate set to 1×10^{-4} and decay by a factor of 0.8 in both two stages. The feature downsampling scale k is set to 2. The hyperparameters λ_1 and λ_2 are both set to 0.01. For our TADM, the U-Net (Ronneberger, Fischer, and Brox 2015) architecture is adopted as the noise estimator network with the time step T and sampling step S set to 1000 and 25 for the forward diffusion and reverse diffusion process, respectively.

Datasets. We have conducted experiments on two publicly available benchmarks including ZRR (Ignatov, Van Gool, and Timofte 2020b) and MAI (Ignatov et al. 2021b) datasets. The ZRR dataset contains 46.8k RAW-sRGB image pairs for training and 1.2k pairs for evaluation, where the RAW images are captured by Huawei P20 and the

Method	Time (ms)	ZRR (Original GT)			ZRR (Align GT with RAW)			MAI		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
PyNet	62.7	21.19	0.747	0.193	22.73	0.845	0.152	23.81	0.848	0.139
AWNet-R	55.7	21.42	0.748	0.198	23.27	0.854	0.151	24.53	0.872	0.136
AWNet-D	62.7	21.53	0.749	0.212	23.38	0.850	0.164	24.64	0.866	0.147
MW-ISPNet	110.5	21.42	0.754	0.213	23.07	0.848	0.165	25.02	0.885	0.133
LiteISPNet	23.3	21.55	0.749	0.187	23.76	0.873	0.133	24.90	0.877	0.123
FourierISP	25.0	21.65	**0.755**	0.182	23.93	0.874	0.124	25.37	0.891	0.072
DiffRAW	-	21.31	0.743	**0.145**	-	-	-	-	-	-
ISPDiffuser (Ours)	490.0	**21.77**	0.754	0.157	**24.09**	**0.881**	**0.111**	**25.64**	**0.894**	**0.071**

Table 1: Quantitative comparisons on the ZRR (Ignatov, Van Gool, and Timofte 2020b) and MAI (Ignatov et al. 2021b) test sets. The best results are highlighted in **bold** and the second best results are in underlined. '-' indicates the results are unavailable since the source code of DiffRAW (Yi et al. 2024) has not been publicly released. 'Time' denotes the average time cost when performing inference on the ZRR test set, where the images are with 448×448 resolution.

Method	ZRR		MAI	
	MUS. ↑	TOP. ↑	MUS. ↑	TOP. ↑
PyNet	43.796	0.362	39.823	0.445
AWNet-R	43.441	0.355	40.211	0.441
AWNet-D	45.100	0.362	39.839	0.432
MW-ISPNet	42.448	0.340	40.652	0.449
LiteISPNet	47.310	0.370	40.365	0.445
FourierISP	44.534	0.369	47.614	**0.535**
ISPDiffuser (Ours)	**50.117**	**0.392**	**48.032**	**0.535**

Table 2: Non-reference perceptual metric comparisons on the ZRR (Ignatov, Van Gool, and Timofte 2020b) and MAI (Ignatov et al. 2021b) test sets. The best results are highlighted in **bold** and the second best results are in underlined. 'MUS.'and 'TOP.' denote MUSIQ and TOPIQ.

sRGB images are captured by Canon camera. Meanwhile, the MAI dataset concentrates on mapping the RAW images captured by Sony IMX586 to sRGB distributions of the Fuji camera. Since the test set of the MAI dataset lacks GT sRGB images, we follow (He et al. 2024b) to split the training set into 90% for training (21.7k pairs) and 10% for evaluation (2.4k pairs). Notably, the RAW images in the ZRR dataset are 10-bit, while those in the MAI dataset are 12-bit.

Metrics. We adopt two full-reference distortion metrics PSNR and SSIM (Wang et al. 2004), and a perceptual metric LPIPS (Zhang et al. 2018) for evaluation. Moreover, two non-reference perceptual metrics MUSIQ (Ke et al. 2021) and TOPIQ (Chen et al. 2024) are adopted to measure the visual quality of generated images.

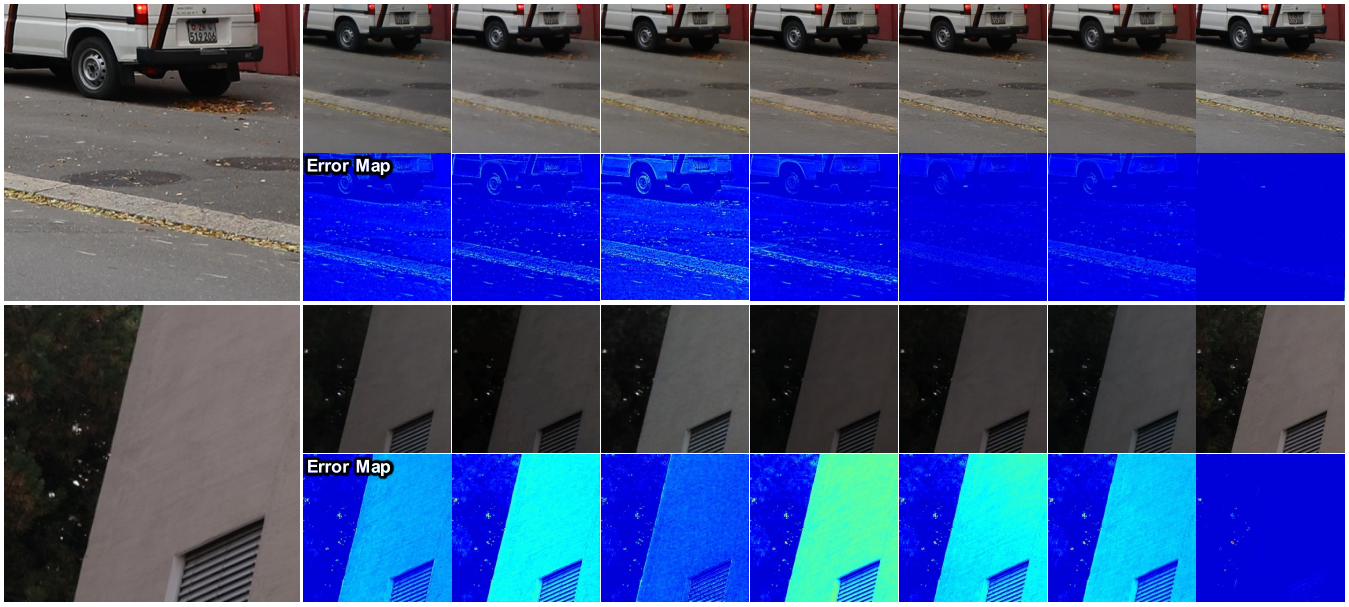
Comparison with Existing Methods

Comparison Methods. In this section, we compare the proposed method with existing state-of-the-art methods, including PyNet (Ignatov, Van Gool, and Timofte 2020b), AWNet (Dai et al. 2020), MW-ISPNet (Ignatov et al. 2020), LiteISPNet (Zhang et al. 2021), FourierISP (He et al. 2024b), and DiffRAW (Yi et al. 2024). Note that we follow (Zhang et al. 2021) which uses two models of AWNet, i.e., AWNet (RAW) denoted as AWNet-R, and AWNet (demosaic) denoted as AWNet-D, for comparison. Moreover, the metrics of DiffRAW are adopted from its associated publication since the source code is unavailable.

Quantitative Comparison. We compare our method with all comparison methods on the ZRR (Ignatov, Van Gool, and Timofte 2020b) and MAI (Ignatov et al. 2021b) test sets. Since the image pairs in the ZRR dataset present spatial misalignment, we follow (Zhang et al. 2021) that calculate the metrics using the original ground truth sRGB images denoted as "Original GT" and adopt the optical flow estimation method PWCNet (Sun et al. 2018) to align the image pairs for evaluation as "Align GT with RAW".

As reported in Table 1, our method outperforms the second-best method FourierISP by 0.12dB and 0.16dB in terms of PSNR under both evaluation modalities of the ZRR dataset. For SSIM and LPIPS, our method is slightly inferior to FourierISP and DiffRAW under the setting of "Original GT" on the ZRR dataset due to the misalignment between original RAW-sRGB image pairs, while achieving the best results with 0.007 and 0.013 improvements in another mode after alignment, which indicates our method can produce images with satisfactory visual quality and is more appropriate for RAW-to-sRGB mappings. For the MAI dataset, our method obtains state-of-the-art performance in all three metrics with 0.27dB improvement in terms of PSNR, 0.003 improvement in SSIM, and 0.001 improvement in LPIPS, respectively, which proves the strong generalization ability of our method. To further validate the effectiveness of our method, we adopt two non-reference perceptual metrics to measure the visual quality of our generated sRGB images on the ZRR and MAI test sets. As illustrated in Table 2, our method achieves the best performance in both two metrics on the ZRR and MAI test sets with a remarkable improvement in MUSIQ, which proves that our method is capable of generating images with better perceptual quality.

Qualitative Comparison. We present qualitative comparisons of our method and competitive methods on the ZRR and MAI test sets, as illustrated in Fig. 4. For better visualization, we provide the error maps beneath each corresponding image that indicate the content discrepancies between the generated sRGB images and GT images. As we can see previous methods, such as PyNet and both variants of AWNet, struggle to preserve color accuracy and detail



(a) GT (b) PyNet (c) AWNet-R (d) AWNet-D (e) MW-ISPNet (f) LiteISPNet (g) FourierISP (h) Ours

Figure 4: Qualitative comparison of our method and competitive methods on the ZRR dataset (Ignatov, Van Gool, and Timofte 2020b) (row 1) and MAI dataset (Ignatov et al. 2021b) (row 2). The error maps represent the content difference between the generated sRGB images and the GT images, the darker the better. Best viewed by zooming in.

sharpness. MW-ISPNet, LiteISPNet, and FourierISP exhibit enhanced performance, yet they continue to encounter limitations in preserving texture richness and ensuring consistent color reproduction. These shortcomings become especially apparent in areas with complex illumination or intricate details, where these methods often produce unexpected artifacts or lose essential information. In contrast, our method can generate results that exhibit high visual fidelity to the ground truth. By effectively capturing and reproducing fine textures, such as intricate patterns and subtle gradients, while maintaining accurate and vibrant color representation, our approach achieves more lifelike, detailed, and visually compelling reconstructions. The superiority of our method is further substantiated by error maps, which reveal minimal discrepancies between our outputs and GT images, underscoring the effectiveness of our approach.

User Study

We further conduct a user study to compare the proposed method with four competitive methods including AWNet-D, MW-ISPNet, LiteISPNet, and FourierISP. We randomly select 20 images from the ZRR and MAI test sets and invite 26 participants to measure the subjective preference of the above methods. For each case, the input RAW data and the generated sRGB results of the five methods are shown at the same time. The participants are required to rank the results from 1 (best) to 5 (worst) based on the perceptual quality including local details and global color. Fig. 5 illustrates the rating distributions of different methods, where our method receives more 'best' ratings, proving our results are more preferred by human subjects.

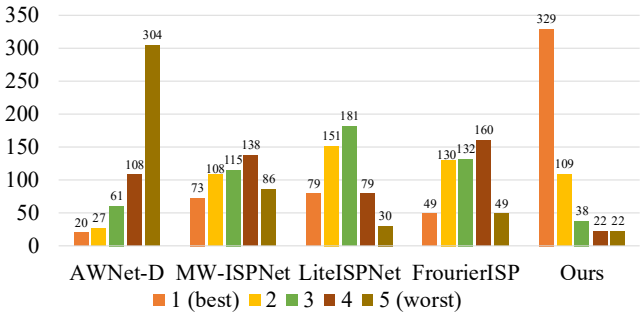


Figure 5: Score distributions of user study, where the ordinate axis records the rating frequency received from the 26 participants. Our method receives more "best" ratings.

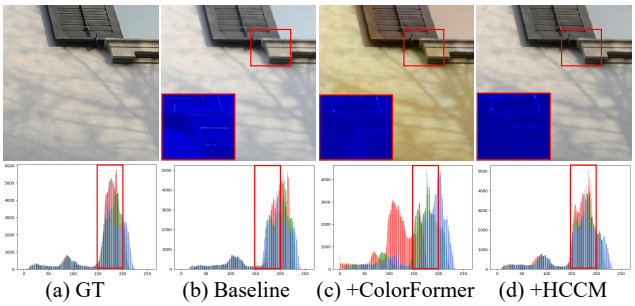
Ablation Study

In this section, we conduct a series of ablation studies to validate the effectiveness of newly proposed components in our method. The quantitative results on the ZRR (Ignatov, Van Gool, and Timofte 2020b) test set under the setting of "Original GT" are reported in Table 3 and Table 4.

Our Framework. To validate the effectiveness of our decoupled framework, we employ the TADM only to form a baseline that takes the sRGB image as input and the RAW image as the condition to directly perform RAW-to-sRGB mappings in the image space, i.e., $k = 0$. We can observe that the baseline without decoupled suffers from poor sharpness with incorrect exposure and color distortion as shown in Fig. 6(b), caused by the simultaneous handling of detail

Methods	Decouple	PSNR ↑	SSIM ↑	LPIPS ↓
Baseline		20.12	0.731	0.208
+DDColor	✓	18.83	0.711	0.293
+ColorFormer	✓	19.24	0.716	0.278
+HCCM	✓	20.93	0.740	0.204

Table 3: Ablation studies of our proposed decoupled framework and histogram-guided color consistency module.



(a) GT (b) Baseline (c) +ColorFormer (d) +HCCM

Figure 6: Visual results of the ablation study about our proposed framework and HCCM module. The second row showcases the color histogram of the image.

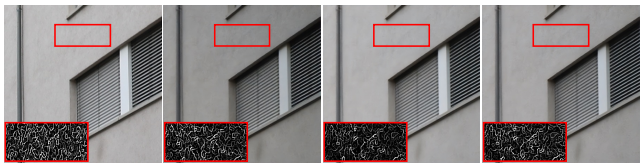
reconstruction and color mapping. In contrast, our decoupled framework that separates grayscale texture reconstruction from colorization delivers sharper details and presents vivid color as shown in Fig. 6(d), presenting overall performance improvements as reported in rows 1 and 4 of Table 3.

HCCM Module. To validate the effectiveness of our proposed histogram-guided color consistency module, we replace it with two established automatic image colorization methods ColorFormer (Ji et al. 2022) and DDColor (Kang et al. 2023), which achieve state-of-the-art performance in transforming grayscale images to sRGB images. As reported in rows 2-4 of Table 3, our method with the HCCM module outperforms these colorization methods in terms of all metrics. As illustrated in Fig. 6(c) and (d), the sRGB image generated by ColorFormer encounters overall color discrepancies, whereas our HCCM is capable of generating results with more accurate and consistent color.

Loss Functions. To validate the effectiveness of our proposed texture enrichment loss $\mathcal{L}_{tel}$ and color consistency loss $\mathcal{L}_{ccl}$, we conduct experiments by individually removing each component from the default setting, where the quantitative results are reported in Table 4. As shown in row 1, the removal of the above two objective functions results in overall performance degradation. The $\mathcal{L}_{ccl}$ is designed to facilitate the HCCM transforming the grayscale feature into the sRGB feature with vivid color, which is beneficial in enhancing the visual fidelity of the restored images, thereby leading to the overall performance improvement. As illustrated in Fig. 7(b) and (d), the inclusion of $\mathcal{L}_{ccl}$ for optimization is capable of correcting global color distortion, leading to more satisfactory results. The incorporation of $\mathcal{L}_{tel}$ is helpful to generate results with shaper details thus achieving noticeable improvements in terms of the distortion metrics. As illustrated in Fig. 7(c) and (d), the $\mathcal{L}_{tel}$ is helpful in generating images with enriched texture information.

$\mathcal{L}_{tel}$	$\mathcal{L}_{ccl}$	PSNR ↑	SSIM ↑	LPIPS ↓
		21.30	0.736	0.161
✓		21.33	0.751	0.156
	✓	21.62	0.750	0.158
✓	✓	21.77	0.754	0.157

Table 4: Ablation studies of our proposed texture enrichment loss $\mathcal{L}_{tel}$ and color consistency loss $\mathcal{L}_{ccl}$.



(a) GT (b) w/o $\mathcal{L}_{ccl}$ (c) w/o $\mathcal{L}_{tel}$ (d) Default

Figure 7: Visual results of the ablation study about our proposed texture enrichment loss $\mathcal{L}_{tel}$ and color consistency loss $\mathcal{L}_{ccl}$. 'w/o' denotes without.

Litimations

Although our method effectively transforms RAW images into DSLR-quality sRGB images, the generalizability to diverse weather, lighting, and devices is limited by the training data from specific cameras. Moreover, since diffusion-based methods rely on iterative denoising of Gaussian noise, our method shows inferior efficiency compared to some lightweight methods which can be applied to the camera to replace the traditional ISP pipeline, as reported in Table 1. In the future, we will explore more effective sampling strategies, such as DPM-Solver (Lu et al. 2022) and consistency model (Song et al. 2023), to improve inference efficiency and investigate the effectiveness of our method.

Conclusion

We have presented ISPDiffuser, a diffusion-based decoupled framework that separates the RAW-to-sRGB mapping into detail reconstruction in grayscale space and color consistency mapping from grayscale to sRGB. Technically, we propose a texture-aware diffusion model that leverages the generative ability of diffusion models to perform grayscale detail reconstruction without concern for color information, where a texture enrichment loss is further proposed to promote the diffusion model to generate more intricate texture details. Subsequently, we construct a histogram-guided color consistency module that utilizes the traditional color histogram as guidance to learn accurate color information for grayscale to sRGB color consistency mapping, with a color consistency loss designed to constrain the learned color information being close to standard DSLR color distribution. Extensive experiments demonstrate that the proposed ISPDiffuser outperforms existing state-of-the-art competitors both quantitatively and qualitatively.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant Nos. 62271334, 62372091, 62071097, and in part by the Sichuan Science and Technology Program under Grant Nos. 2023NSFSC0462, 2023NSFSC0458, 2023NSFSC1972.

References

Chen, C.; Mo, J.; Hou, J.; Wu, H.; Liao, L.; Sun, W.; Yan, Q.; and Lin, W. 2024. TOPIQ: A Top-Down Approach From Semantics to Distortions for Image Quality Assessment. *IEEE TIP*, 33: 2404–2418.

Cheng, D.; Price, B.; Cohen, S.; and Brown, M. S. 2015. Beyond white: Ground truth colors for color constancy correction. In *ICCV*, 298–306.

Chung, H.; Sim, B.; and Ye, J. C. 2022. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *CVPR*, 12413–12422.

Dabov, K.; Foi, A.; Katkovnik, V.; and Egiazarian, K. 2007. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE TIP*, 16(8): 2080–2095.

Dai, L.; Liu, X.; Li, C.; and Chen, J. 2020. Awnet: Attentive wavelet network for image isp. In *ECCVW*, 185–201.

Gharbi, M.; Chaurasia, G.; Paris, S.; and Durand, F. 2016. Deep joint demosaicking and denoising. *ACM TOG*, 35(6): 1–12.

Han, J.; and Ma, K.-K. 2002. Fuzzy color histogram and its use in color image retrieval. *IEEE TIP*, 11(8): 944–952.

He, C.; Shen, Y.; Fang, C.; Xiao, F.; Tang, L.; Zhang, Y.; Zuo, W.; Guo, Z.; and Li, X. 2024a. Diffusion Models in Low-Level Vision: A Survey. *arXiv preprint arXiv:2406.11138*.

He, X.; Hu, T.; Wang, G.; Wang, Z.; Wang, R.; Zhang, Q.; Yan, K.; Chen, Z.; Li, R.; Xie, C.; et al. 2024b. Enhancing RAW-to-sRGB with Decoupled Style Structure in Fourier Domain. In *AAAI*, volume 38, 2130–2138.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, 6840–6851.

Ignatov, A.; Chiang, C.-M.; Kuo, H.-K.; Sycheva, A.; and Timofte, R. 2021a. Learned smartphone isp on mobile npus with deep learning, mobile ai 2021 challenge: Report. In *CVPR*, 2503–2514.

Ignatov, A.; Chiang, C.-M.; Kuo, H.-K.; Sycheva, A.; and Timofte, R. 2021b. Learned smartphone isp on mobile npus with deep learning, mobile ai 2021 challenge: Report. In *CVPRW*, 2503–2514.

Ignatov, A.; Timofte, R.; Zhang, Z.; Liu, M.; Wang, H.; Zuo, W.; Zhang, J.; Zhang, R.; Peng, Z.; Ren, S.; et al. 2020. Aim 2020 challenge on learned image signal processing pipeline. In *ECCVW*, 152–170.

Ignatov, A.; Van Gool, L.; and Timofte, R. 2020a. Replacing mobile camera isp with a single deep learning model. In *CVPRW*, 536–537.

Ignatov, A.; Van Gool, L.; and Timofte, R. 2020b. Replacing mobile camera isp with a single deep learning model. In *CVPRW*, 536–537.

Ji, X.; Jiang, B.; Luo, D.; Tao, G.; Chu, W.; Xie, Z.; Wang, C.; and Tai, Y. 2022. Colorformer: Image colorization via color memory assisted hybrid-attention transformer. In *ECCV*, 20–36.

Jiang, H.; Luo, A.; Han, S.; Fan, H.; and Liu, S. 2023. Low-Light Image Enhancement with Wavelet-based Diffusion Models. *ACM TOG*, 42(6): 1–14.

Jiang, H.; Luo, A.; Liu, X.; Han, S.; and Liu, S. 2025. Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. In *ECCV*, 161–179.

Kang, X.; Yang, T.; Ouyang, W.; Ren, P.; Li, L.; and Xie, X. 2023. DDColor: Towards Photo-Realistic Image Colorization via Dual Decoders. In *ICCV*, 328–338.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising Diffusion Restoration Models. In *NeurIPS*, volume 35, 23593–23606.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 6007–6017.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *ICCV*, 5148–5157.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kwok, N. M.; Shi, H.; Ha, Q. P.; Fang, G.; Chen, S.; and Jia, X. 2013. Simultaneous image color correction and enhancement using particle swarm optimization. *EAAI*, 26(10): 2356–2371.

Li, H.; Jiang, H.; Luo, A.; Tan, P.; Fan, H.; Zeng, B.; and Liu, S. 2024. DMHomo: Learning Homography with Diffusion Models. *ACM TOG*, 43(3): 1–16.

Liu, Z.; Lin, W.; Li, X.; Rao, Q.; Jiang, T.; Han, M.; Fan, H.; Sun, J.; and Liu, S. 2021. ADNet: Attention-guided deformable convolutional network for high dynamic range imaging. In *CVPRW*, 463–470.

Liu, Z.; Wang, Y.; Zeng, B.; and Liu, S. 2022. Ghost-free high dynamic range imaging with context-aware transformer. In *ECCV*, 344–360.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *NeurIPS*, 35: 5775–5787.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 11461–11471.

Luo, A.; Li, X.; Yang, F.; Liu, J.; Fan, H.; and Liu, S. 2024. FlowDiffuser: Advancing Optical Flow Estimation with Diffusion Models. In *CVPR*, 19167–19176.

Ning, M.; Li, M.; Su, J.; Salah, A. A.; and Ertugrul, I. O. 2024. Elucidating the Exposure Bias in Diffusion Models. In *ICLR*.

Ramanath, R.; Snyder, W. E.; Yoo, Y.; and Drew, M. S. 2005. Color image processing pipeline. *IEEE SPM*, 22(1): 34–43.

Rana, A.; Singh, P.; Valenzise, G.; Dufaux, F.; Komodakis, N.; and Smolic, A. 2019. Deep tone mapping operator for high dynamic range images. *IEEE TIP*, 29: 1285–1298.

Rizzi, A.; Gatta, C.; and Marini, D. 2003. A new algorithm for unsupervised global and local color correction. *PRL*, 24(11): 1663–1677.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241.

Schwartz, E.; Giryes, R.; and Bronstein, A. M. 2018. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE TIP*, 28(2): 912–923.

Shekhar Tripathi, A.; Danelljan, M.; Shukla, S.; Timofte, R.; and Van Gool, L. 2022. Transform your smartphone into a dslr camera: Learning the isp in the wild. In *ECCV*, 625–641.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2256–2265.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023. Consistency Models. In *ICML*, 32211–32252.

Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 8934–8943.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4): 600–612.

Wang, Z.; She, Q.; and Ward, T. E. 2021. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys*, 54(2): 1–38.

Xing, Y.; Qian, Z.; and Chen, Q. 2021. Invertible image signal processing. In *CVPR*, 6287–6296.

Yi, M.; Zhang, K.; Liu, P.; Zuo, T.; and Tian, J. 2024. DiffRAW: Leveraging Diffusion Model to Generate DSLR-Comparable Perceptual Quality sRGB from Smartphone RAW Images. In *AAAI*, volume 38, 6711–6719.

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

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE TIP*, 26(7): 3142–3155.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 586–595.

Zhang, Z.; Han, L.; Ghosh, A.; Metaxas, D. N.; and Ren, J. 2023. Sine: Single image editing with text-to-image diffusion models. In *CVPR*, 6027–6037.

Zhang, Z.; Wang, H.; Liu, M.; Wang, R.; Zhang, J.; and Zuo, W. 2021. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *ICCV*, 4348–4358.

Zhou, T.; Li, H.; Wang, Z.; Luo, A.; Zhang, C.-L.; Li, J.; Zeng, B.; and Liu, S. 2024. RecDiffusion: Rectangling for Image Stitching with Diffusion Models. In *CVPR*, 2692–2701.