ERD: Extended RAW-Diffusion Framework for De-rendering sRGB Images

Yifan Qu, Jiaqi Shang

Abstract—Recovering RAW sensor measurements from sRGB images is a central problem in computational photography, as RAW data preserves the true scene radiance prior to the nonlinear transformations introduced by a camera's Image Signal Processing (ISP) pipeline. However, ISPs vary across different camera brands and models, making inverse ISP reconstruction particularly challenging when the sensor characteristics are unknown. Existing approaches often rely on metadata, modifiable ISP, or camera-specific training, which limits their ability to generalize across unseen devices. In this project, we investigate a diffusion-based inverse ISP framework designed for cross-sensor RAW reconstruction. Building upon the RAW-Diffusion model, we incorporate a ControlNet-guided architecture that provides structured conditioning to improve generalization without requiring metadata at inference time. Using the MIT-Adobe FiveK dataset, we evaluate seven camera models, which is sufficient to test cross-sensor robustness. Our results demonstrate that the proposed ControlNet-enhanced model enhances reconstruction accuracy on unseen sensors, outperforming the baseline RAW-Diffusion model on the Nikon dataset and achieving competitive performance on Canon, Leica, and Sony. These findings highlight the potential of guidance-based diffusion models for practical, camera-agnostic inverse ISP.

Index Terms—Inverse ISP, RAW reconstruction, diffusion models, ControlNet, cross-sensor generalization, computational photography, image restoration.

1 Introduction

Traditional cameras capture a raw image with the sensor and then immediately pass it to an Image Signal Processor (ISP), which converts the raw data into an RGB image that is easier for humans to read and store. However, some computer vision and remote sensing tasks rely on RAW images because they preserve the true scene-radiance captured by the camera sensor before any in-camera processing occurs. RAW images are essential in tasks such as low-light enhancement, deblurring, HDR reconstruction, demosaicing research, and other physics-based image formation studies. In consumer and professional cameras, they rarely store RAW data by default because RGB has a lower dynamic range and is easier to store. Hence, they pass sensor measurements through a complex ISP pipeline that performs operations such as white balancing, demosaicing, tone mapping, color correction, gamma compression, and noise reduction. These steps produce visually pleasing but heavily transformed sRGB images, resulting in a significant loss of radiometric information.

Recovering RAW images from sRGB is therefore a fundamental problem. The core difficulty lies in the fact that the ISP used in most cameras is nonlinear, camera-specific, and often proprietary. A typical ISP pipeline is shown in Figure 1. It reduces the dynamic range of the original RAW image; the task of recovering the actual RAW image becomes similar to a generative work to generate the loss of pixel values. This problem is further complicated in real-world scenarios where images come from arbitrary camera brands and sensor models, each with a distinct forward ISP.

In this paper, we propose a method that can infer RAW sensor measurements directly from sRGB inputs, with a particular focus on achieving cross-sensor generalization.

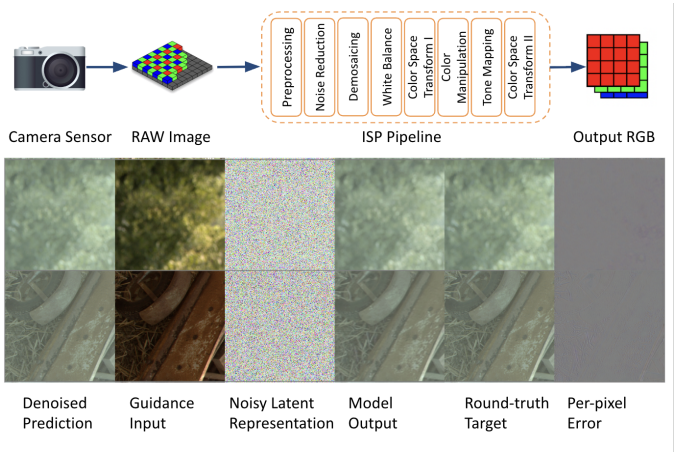

Fig. 1: The top row illustrates a typical camera ISP pipeline. The bottom row shows an example visualization during training.

Specifically, we aim to develop a method that can undo the ISP for images captured by multiple consumer camera brands, including Nikon, Canon, Leica, and Sony, while relying on metadata only during training. By learning a sensor-agnostic inverse mapping, our approach estimates the true irradiance hitting the sensor regardless of the camera used, without requiring sensor-specific models or detailed metadata at inference time. However, given the limited time and computational resources, we are unable to run experiments on all camera models. Instead, we report comprehensive experiments on seven representative models, including both quantitative and qualitative results.

Code available at https://github.com/jqshang/Extended-RAW-Diffusion.

2 Related Work

2.1 Raw Image Restoration

Previous work on inverse ISP reconstruction uses embedded camera metadata. Those methods assume access to parameters such as digital gains, color correction matrices (CCMs), or white balance coefficients contained in the RAW or JPEG images. By having this metadata, the reconstruction of RAW is achieved by reversing ISP operations step by step. For instance, Brooks et al. [1] propose a pipeline that undoes the ISP using a supervised trained neural network. Similarly, several works follow the inverse order of the ISP using learning-based architectures to synthesize RAW data from sRGB images [2], [3], [4]. Those methods are effective and accurate when camera information is known, but have difficulties with cross-sensor generalization because ISPs vary significantly across camera manufacturers and even between models from the same brand.

Alternative approaches attempt to redesign the forward ISP for invertibility rather than attempting to reverse complex proprietary pipelines. InvISP [5] introduces a mathematically invertible forward ISP, which enables reconstruction of RAW data by applying the inverse transform. Although this method works theoretically, its formulation requires full control and redesign of the ISP which is limited to specific cameras that can modify its internal design of ISP.

Recently, Nam et al. [6] proposed a superpixel-based sampling network that separates reconstruction and sampling phases to better model RAW-to-sRGB mappings. However, their model still assumes access to the underlying ISP and is limited when applied to unseen sensors.

2.2 Diffusion Models in Raw Image Restoration

Diffusion models have recently become a powerful class of generative models for image restoration. The foundation of these models was introduced by Ho et al. [7], who proposed Denoising Diffusion Probabilistic Models (DDPMs). DDPMs gradually corrupt images with noise and learn to reverse this noising process through a parameterized denoiser. This method provides a principled probabilistic framework for learning complex data distributions.

Subsequent work introduces Denoising Diffusion Implicit Models (DDIMs) [8], which reduces the sampling steps through a non-Markovian, deterministic reverse process while preserving fidelity and controllability. These improvements have made diffusion models increasingly practical for real-time or high-resolution image restoration tasks.

Diffusion models have also been applied to tasks other than generation, including data augmentation and structured image transformation. Trabucco et al. [9] demonstrate that diffusion-based augmentations can improve model generalization by creating high-fidelity, semantically consistent variations of training data. Fang et al. [10] propose a controllable diffusion framework for object detection augmentation, further highlighting the adaptability of diffusion processes to structured, domain-specific image transformations.

Building on these developments of the Diffusion model, Reinders et al. [11] use the concept of the diffusion model to restore the RAW image. This method utilizes a diffusion model with a UNet backbone and integrates an sRGB-guided module at multiple resolutions to generate RAW images. RAW-Diffusion achieves high-quality reconstructions but requires training a separate model for each sensor, making it unsuitable for broad cross-sensor deployment. Moreover, because the method does not incorporate metadata during training, metadata must later be used for bias evaluation, adding another layer of complexity.

Across these methods, the common limitation is that they all rely on known ISPs or sensor-specific training. Therefore, the performance of these methods degrades significantly when applied to sRGB images from unseen cameras. This motivates our project objective of building a truly metadata-agnostic and sensor-generalizable inverse ISP model.

3 Proposed Method

We propose the Extended RAW-Diffusion framework to generate high-fidelity RAW images conditioned on camera-specific priors. Although the naive RGB-guided RAW-Diffusion model has demonstrated success in reconstructing RAW data from RGB inputs, it is unable to explicitly model sensor-specific characteristics such as color filter arrays without retraining the entire backbone [11].

Our method consists of three primary components:

- **RGB-guided RAW-Diffusion backbone.**
  A pre-trained UNet–like diffusion model that directly predicts denoised RAW images guided by RGB features. This backbone is the naive RGB-guided RAW-Diffusion model.
- **Camera-conditioned ControlNet.**
  A camera-aware ControlNet branch that reuses the trainable copies of the encoder and bottleneck from the naive backbone and learns residual feature maps that encode camera-specific corrections.
- **Efficient residual injection.**
  A lightweight design that injects ControlNet residuals via zero convolutions only at the bottleneck and decoder layers of the frozen backbone to ensure stable, low-overhead conditioning.

We also propose a variant of residual blocks with global conditioning via Feature-wise Linear Modulation (FiLM) [12] that can inject sensor-specific conditions. Although our best model uses only ControlNet-based conditioning, we present FiLM to support our ablation studies.

3.1 Preliminaries: RGB-guided RAW-Diffusion

We build upon the RAW-Diffusion architecture, which is a Denoising Diffusion Probabilistic Model (DDPM) [7] specialized for the RAW domain and guided by an accompanying sRGB image. Given a data sample, $x_0 \in \mathbb{R}^{C\times H\times W}$, from the distribution of RAW images. The forward diffusion process corrupts x_0 with Gaussian noise according to

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}, x_0, (1-\bar{\alpha}t)\mathbf{I}\big), \quad t = 1, \ldots, T,$$

where $\beta_t t = 1^T$ is a pre-defined variance schedule, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}t = \prod s = 1^t \alpha_s$. The noisy state at timestep t is

$$x_t = \sqrt{\bar{\alpha}_t}, x_0 + \sqrt{1-\bar{\alpha}_t}, \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}).$$

The reverse process learns to recover x_0 from a noisy sample x_t by a UNet-style backbone $\mathcal{F}_\theta$. The backbone

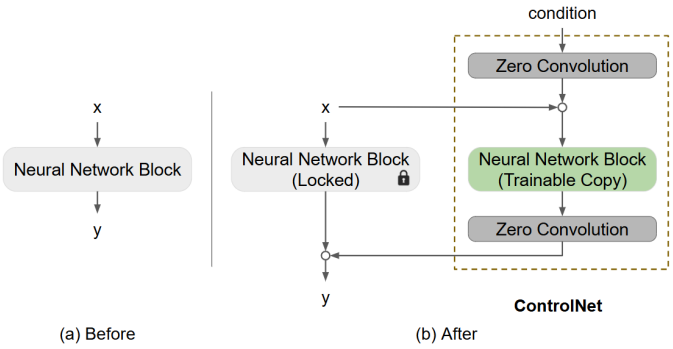

Fig. 2: ControlNet: (a) A neural network block maps an input feature map x to an output feature map y. (b) The parameters of the original neural network block are frozen. We create a trainable copy and link x and y using zero-convolution layers, i.e., 1×1 convolutions whose weights and biases are initialized to zero.

consists of an encoder, a middle block, and a decoder with residual connections. To enable color-awareness denoising, RAW-Diffusion incorporates an RGB-guidance module, which extracts a set of multi-scale guidance features g from a given processed sRGB image $I_{\text{rgb}} \in \mathbb{R}^{3\times H\times W}$.

The backbone is trained to directly predict the denoised RAW image, and the network prediction at timestep t is denoted by

$$x_{\text{pred}} = \mathcal{F}_\theta(x_t, t, g),$$

which is supervised to approximate x_0 under the diffusion training objective.

In our approach, we use a pre-trained instance of this RGB-guided RAW-Diffusion model on images captured by a particular camera as a backbone. Its parameters θ are frozen in all subsequent experiments when we introduce additional camera-conditioning modules, which preserves the learned RAW prior and the semantic RGB-to-RAW feature mapping.

3.2 Camera Representation

To explicitly model sensor-specific characteristics, we condition the generation process on the specific camera model. Let $\mathcal{K} = \{1, \dots, K\}$ denote the set of supported camera identifiers. For a training sample associated with camera index $k \in \mathcal{K}$, we define the condition vector $\mathbf{v}_c \in \{0,1\}^K$ as a one-hot encoding:

$$(\mathbf{v}_c)_i = \mathbb{I}(i = k), \quad \forall i \in \{1, \dots, K\},$$

where $\mathbb{I}(\cdot)$ is the indicator function.

3.3 ControlNet

Our ControlNet branch follows the design of ControlNet for text-to-image diffusion [13] and is adapted to RAW-Diffusion. The ControlNet architecture consists of a trainable copy of the backbone's encoder and bottleneck blocks, augmented with conditioning injection as shown in Fig. 2.

Formally, let $\mathcal{B}_\ell^{\text{enc}}$ denote the ℓ-th encoder or bottleneck block of RAW-Diffusion with parameters θ, and let $\mathcal{B}_\ell^c$ denote its trainable ControlNet counterpart. We freeze θ to preserve the pre-trained generative prior. The input to ControlNet is the same noisy state x_t, augmented with a camera-specific condition map. We represent the camera model indexed at $k \in \mathcal{K}$ by a one-hot vector $\mathbf{v}_c \in \{0,1\}^K$, and broadcast it over the spatial dimensions to obtain

$$H_{\text{hint}} \in \mathbb{R}^{H\times W\times K}, \qquad H_{\text{hint}}(i,j,\cdot) = \mathbf{v}_c \;\; \forall i,j.$$

For a sequence of L encoder and bottleneck blocks, the ControlNet branch computes

$$h_0 = x_t, \qquad h_{\ell+1} = \mathcal{B}_\ell^c\big(h_\ell, e_t, g, H_{\text{hint}}\big), \quad \ell = 0, \dots, L-1,$$

where e_t is the timestep embedding and g is the guidance features for the RGB-guidance module. In the classical ControlNet design, the output of every trainable block $\mathcal{B}_\ell^c$ is passed through a zero-initialized 1×1 convolutional layer $\mathcal{Z}(\cdot;\theta_z)$ initialized with both weights and biases set to zero, which is then injected to the corresponding feature map of the frozen backbone's decoder by

$$\tilde{y}_\ell = y_\ell + \mathcal{Z}(h_{\ell+1};\theta_z),$$

where y_ℓ is the feature map from the frozen backbone. Because the zero convolutions are initialized to zero, the combined model is identical to the original RAW-Diffusion backbone at the beginning of training, and camera conditioning is learned purely as a residual refinement without disrupting the pre-trained distribution.

3.4 Efficient residual injection

While the classical ControlNet copies all encoder and bottleneck blocks, applying residual connections at all L blocks introduces significant computational redundancy and memory overhead. To address this problem, we adopt a more efficient coupling for RAW-Diffusion as shown in Fig. 3. A lightweight design produces residual feature maps that are injected only at the bottleneck and decoder residual connections of the frozen backbone, with all ControlNet outputs routed through zero-initialized 1×1 convolutions. Formally, we modify the injection mechanism such that

$$\tilde{y}_\ell = \begin{cases} y_\ell + \mathcal{Z}(h_{\ell+1};\theta_z) & \text{if } \ell = 0, \dots, L-1, \\ y_\ell & \text{otherwise.} \end{cases}$$

This guarantees stable training and low-overhead conditioning without changing the original backbone behaviour at initialization.

3.5 Optional Global Conditioning via FiLM

Although our best model only uses ControlNet conditioning, the architecture also supports Feature-wise Linear Modulation (FiLM) [12] for injecting global conditioning in the backbone, shown in Fig. 4. The FiLM residual block introduces a global metadata vector $z \in \mathbb{R}^d$, and we use a Multi-Layer Perceptron (MLP) as a condition generator to map z to a set of scale (γ) and shift (β) parameters for each channel c in a target feature map. Given an input feature map h, timestep embedding e_t, and metadata vector z, a FiLM residual block modifies the standard normalization process by

$$\begin{aligned} h' &= \text{BatchNorm}(h), \\ (\gamma_1, \beta_1) &= \text{MLP}_1(z), \\ h' &= h' \odot (1 + r_1) + \beta_1, \end{aligned}$$

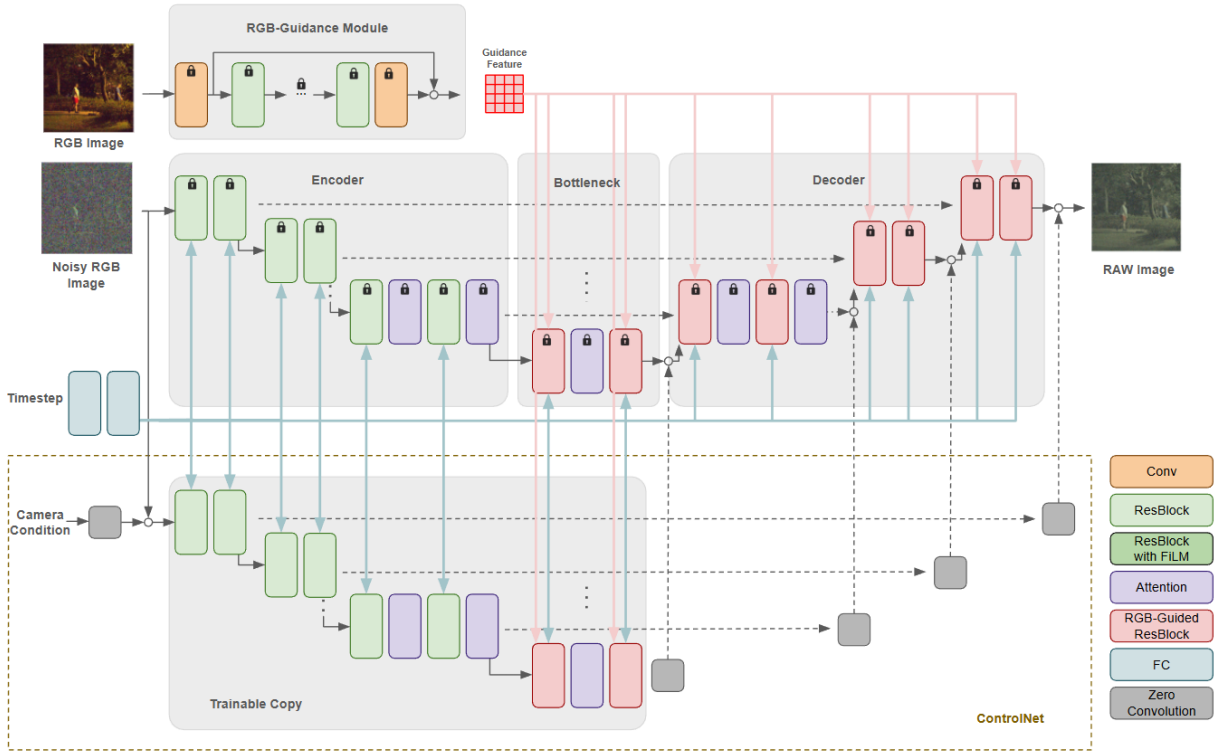

Fig. 3: Raw-Diffusion with ControlNet Injection. The upper pathway is the frozen RGB-guided RAW-Diffusion backbone; the lower pathway is the trainable ControlNet, which produces multi-scale residuals that are passed through zero convolutions.

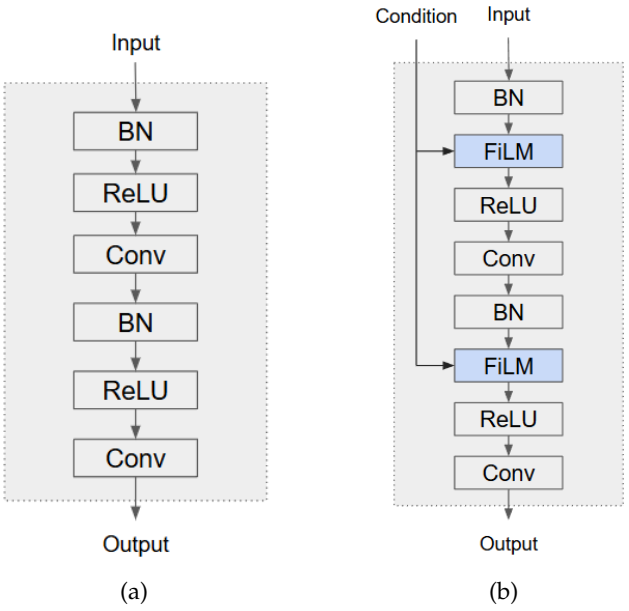

Fig. 4: (a) Original Residual Block; (b) Residual Block with Feature-wise Linear Modulation (FiLM).

where $\odot$ denotes element-wise multiplication. The modulated feature map h' is then passed through the activation function and subsequent convolutional layers. In our framework, the FiLM layers can be optionally inserted into the residual blocks of the naive RGB-guided RAW-Diffusion backbone. For a direct comparison with ControlNet-based camera conditioning, we set $z = \mathbf{v}_c$, i.e., the one-hot encoding of the camera model.

4 Experiments

4.1 Dataset

The experiments are conducted on the MIT-Adobe FiveK dataset [14]. The dataset contains 5,000 photographs captured by a diverse range of DSLR cameras, including Canon, Nikon, and Sony models. Therefore, it is particularly suitable for evaluating the cross-sensor generalization of our proposed method.

Each image in FiveK is provided in both the RAW format and its corresponding sRGB format. This helps us obtain the ground truth RAW image for training and testing. Due to the limited space of this report and the constraints on both computational resources and time, we only present our evaluation on seven datasets (see additional experiments in Appendix A), including the Canon EOS 5D and Nikon D700 datasets. However, the proposed framework is designed to be sensor-agnostic, and the generalization capability can be readily extended to additional cameras.

4.2 Experimental Setup

Training Details: We normalize all data to the range $[-1, 1]$. RAW images are first scaled using their corresponding black and white levels. During training, we extract patches using random cropping, rotation, and flipping as data augmentation. Both the naive RAW-Diffusion model and the RAW-Diffusion model with FiLM are trained for

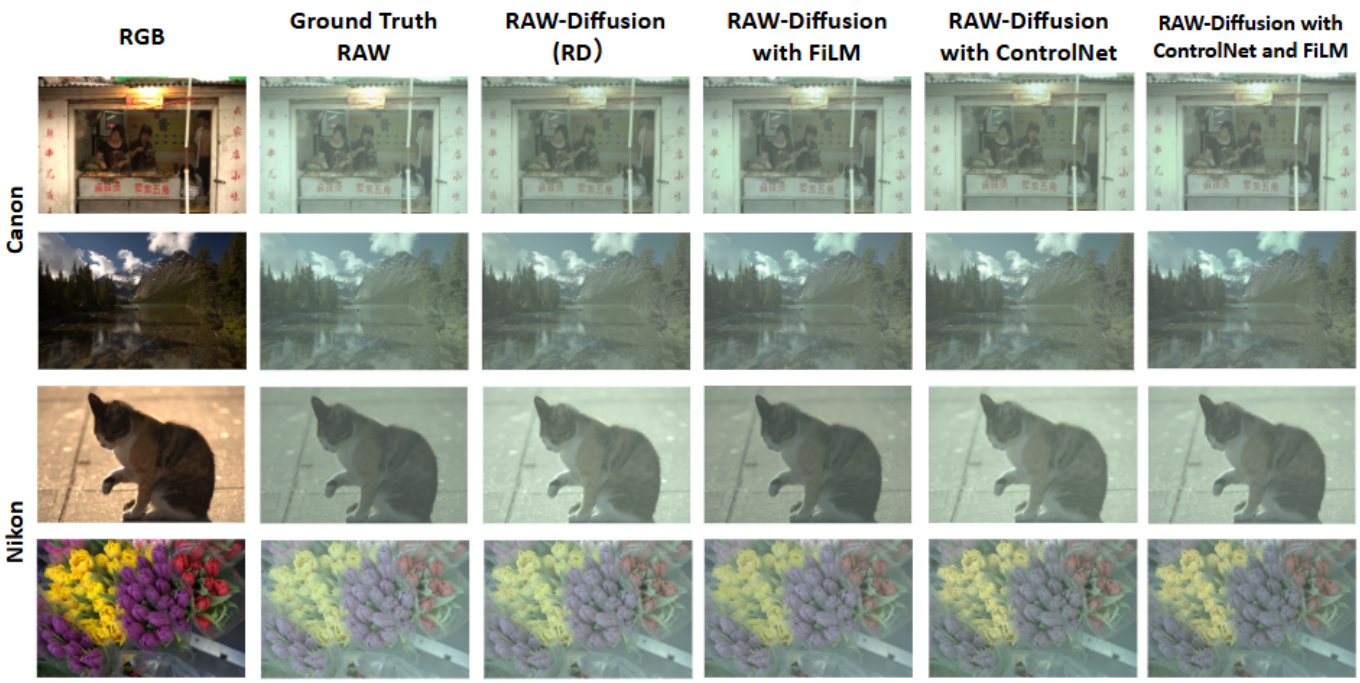

Fig. 5: Qualitative comparison of RAW reconstruction across different model variants. From left to right: input sRGB image, ground-truth RAW, RAW-Diffusion (RD), RAW-Diffusion with FiLM, RAW-Diffusion with ControlNet (ours), and RAW-Diffusion with both ControlNet and FiLM.

10k steps using AdamW with an initial learning rate of 0.0001, no weight decay, and a batch size of 4. The RAW-Diffusion model with ControlNet and the variant with both ControlNet and FiLM are trained for 70k steps under the same optimization hyperparameters. The longer training schedule is used because these models are trained on seven datasets, whereas the baseline is trained on a single dataset. The learning rate is decreased linearly to zero over the course of training. The number of diffusion steps, T, is set to 1000, with a linear variance schedule ranging from $\beta_1 = 0.0001$ to $\beta_T = 0.02$. All experiments are conducted with Colab A100 GPU.

Evaluation: We evaluate the PSNR and SSIM [15] metrics to assess the quality of the generated RAW images. All images are normalized to the range $[0, 1]$ prior to computing these metrics.

4.3 Evaluation and Results

Experiment	Canon EOS 5D		NIKON D700	
	PSNR	SSIM	PSNR	SSIM
RAW	**33.45**	**0.8981**	28.86	0.8173
RAW + FiLM	31.73	0.8628	29.62	0.8357
RAW + ControlNet (ours)	<u>33.06</u>	<u>0.8914</u>	**30.87**	**0.8753**
RAW + FiLM + ControlNet	31.78	0.8548	<u>30.37</u>	<u>0.8679</u>

TABLE 1: Results of training RAW-Diffusion on the two datasets. Our model achieves performance comparable to the naive RAW-Diffusion while additionally demonstrating cross-camera generalization.

We evaluate our proposed model against the original RAW-Diffusion baseline using both quantitative and qualitative metrics. We chose quantitative results for two models for analysis, which are illustrated in Table 1, the full results are presented in Table 2. For the Canon EOS 5D, our ControlNet-augmented model performs comparably to the original RAW-Diffusion model, achieving the second-best results with only a minor drop in PSNR (33.06 dB) and SSIM (0.8914) with a difference of approximately 1.3%. This suggests that ControlNet successfully guides reconstruction without heavily relying on Canon-specific characteristics. In contrast, introducing FiLM layers reduces performance for both RAW-Diffusion and the combined FiLM+ControlNet configuration, indicating that FiLM-based modulation may overfit to training distributions or conflict with the diffusion denoising process in the Canon setting.

In the Nikon D700 experiments, our ControlNet-enhanced model clearly outperforms all baselines, achieving the highest PSNR (30.87 dB) and SSIM (0.8753). Notably, our model also made an improvement over the original RAW-Diffusion model, which attains only 28.86 dB and 0.8173. The combined FiLM and ControlNet model also performs competitively, ranking second, but still falls short of the ControlNet-only variant. These findings indicate that ControlNet effectively provides structural and radiometric guidance that generalizes better to unseen or less-constrained sensor distributions, especially in cross-sensor scenarios.

To further evaluate reconstruction quality, we visualize RAW outputs produced by all model variants across several challenging scenes, the chosen images for analysis are shown in Figure 5, and there are more images at the appendix Figure 6 and Figure 7. The examples are chosen to represent different ISP challenges such as color distortion, tonal compression, and illumination variability.

Across all cases, the baseline RAW-Diffusion model captures the overall structure but often introduces noticeable color casts and reduced contrast. For instance, in the in-

door example, the baseline reconstruction appears overly desaturated and lacks the subtle tonal transitions present in the ground-truth RAW. Similar issues arise in the landscape and cat images, where shadows are washed out, and fine gradients are lost.

In contrast, our RAW-Diffusion with ControlNet substantially improves the realism of the reconstructions. ControlNet guidance better preserves local structure and global illumination cues.

These visual improvements align with the quantitative gains observed in our Nikon experiments. Although the FiLM + ControlNet combination offers some benefits, the effects of FiLM may interfere with ControlNet's hint, leading to less stable tonal rendering.

Overall, the qualitative and quantitative results demonstrate that ControlNet provides the most robust and visually faithful RAW reconstructions, especially in scenes with complex lighting or color distributions, supporting its effectiveness for cross-sensor generalization.

5 Conclusion

In this project, we explored the problem of converting to RAW images from sRGB images and addressed the limitations of traditional inverse ISP methods. Our proposed method builds on the RAW-Diffusion framework. We introduced both ControlNet-guided and Film-guided variants designed to achieve a generalization across different camera ISPs while maintaining PSNR comparable to the baseline model.

Our experimental results on the MIT-Adobe FiveK dataset showed that the proposed method is able to achieve cross-sensor performance. Particularly on the Nikon D700, where it outperforms the original RAW-Diffusion model in both PSNR and SSIM. Compared to FiLM-based modulation, ControlNet provides more stable results and enables the model to better adapt to sensor variability without overfitting. These findings show the effectiveness of structured guidance mechanisms for inverse ISP tasks and highlight the potential for camera-independent RAW reconstruction.

Due to the limited time and computational resources, our model is trained and tested on a limited number of cameras, and has not had a chance to test on our own datasets. Future work may expand the training set to a wider range of cameras, and create an RGB to RAW demonstration by taking photos using any camera and sending them to our model, and outputting its raw image. In addition, we employed a naive one-hot camera encoding as a hint for ControlNet, and we will explore richer structural hints for future work. Overall, this study takes a step toward developing a practical and generalizable inverse ISP solution capable of recovering accurate RAW images from any consumer device.

Acknowledgments

We would like to thank Professor David Lindell for his valuable insights, and Teaching Assistant Shayan Shekarforoush for his continuous support throughout this project.

References

[1] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, and J. T. Barron, "Unprocessing images for learned raw denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 036–11 045.

[2] Y.-L. Liu, W.-S. Lai, Y.-S. Chen, Y.-L. Kao, M.-H. Yang, Y.-Y. Chuang, and J.-B. Huang, "Single-image hdr reconstruction by learning to reverse the camera pipeline," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1651–1660.

[3] A. Punnappurath and M. S. Brown, "Learning raw image reconstruction-aware deep image compressors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 1013–1019, 2020.

[4] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Cycleisp: Real image restoration via improved data synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2696–2705.

[5] Y. Xing, Z. Qian, and Q. Chen, "Invertible image signal processing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6287–6296.

[6] S. Nam, A. Punnappurath, M. A. Brubaker, and M. 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*, 2022, pp. 17 704–17 713.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[8] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[9] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov, "Effective data augmentation with diffusion models," *arXiv preprint arXiv:2302.07944*, 2023.

[10] H. Fang, B. Han, S. Zhang, S. Zhou, C. Hu, and W.-M. Ye, "Data augmentation for object detection via controllable diffusion models," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 1257–1266.

[11] C. Reinders, R. Berdan, B. Besbinar, J. Otsuka, and D. 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)*. IEEE, 2025, pp. 8431–8443.

[12] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[13] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.

[14] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input / output image pairs," in *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

APPENDIX

In this appendix, we report experimental results on additional camera models.

		Raw-Diffusion	Raw-Diffusion + FiLM	Raw-Diffusion + ControlNet	Raw-Diffusion + FiLM + ControlNet
Canon EOS 5D	PSNR	**33.45**	31.73	33.13	32.56
	SSIM	**0.8981**	0.8628	0.8863	0.8661
Canon EOS 450D	PSNR	**33.10**	31.39	32.93	32.16
	SSIM	0.8646	0.8581	**0.8761**	0.8415
Canon EOS 20D	PSNR	29.43	**30.56**	29.86	29.03
	SSIM	0.8047	0.7923	0.8024	**0.8073**
NIKON D700	PSNR	**28.86**	26.62	28.07	27.27
	SSIM	**0.8173**	0.7857	0.8114	0.7905
Nikon D70s	PSNR	**29.24**	27.59	**29.24**	29.13
	SSIM	**0.7914**	0.7830	0.7895	0.7841
Leica M8	PSNR	**32.15**	30.32	32.09	31.04
	SSIM	**0.8481**	0.8072	0.8283	0.8146
Sony DSLR A900	PSNR	**31.57**	28.63	31.14	29.90
	SSIM	**0.8393**	0.7830	0.8367	0.8111

TABLE 2: Results of training RAW-Diffusion on the seven datasets. Our model achieves performance comparable to the naive RAW-Diffusion while additionally demonstrating cross-camera generalization.

Fig. 6: Qualitative comparison of RAW reconstruction across different model variants.

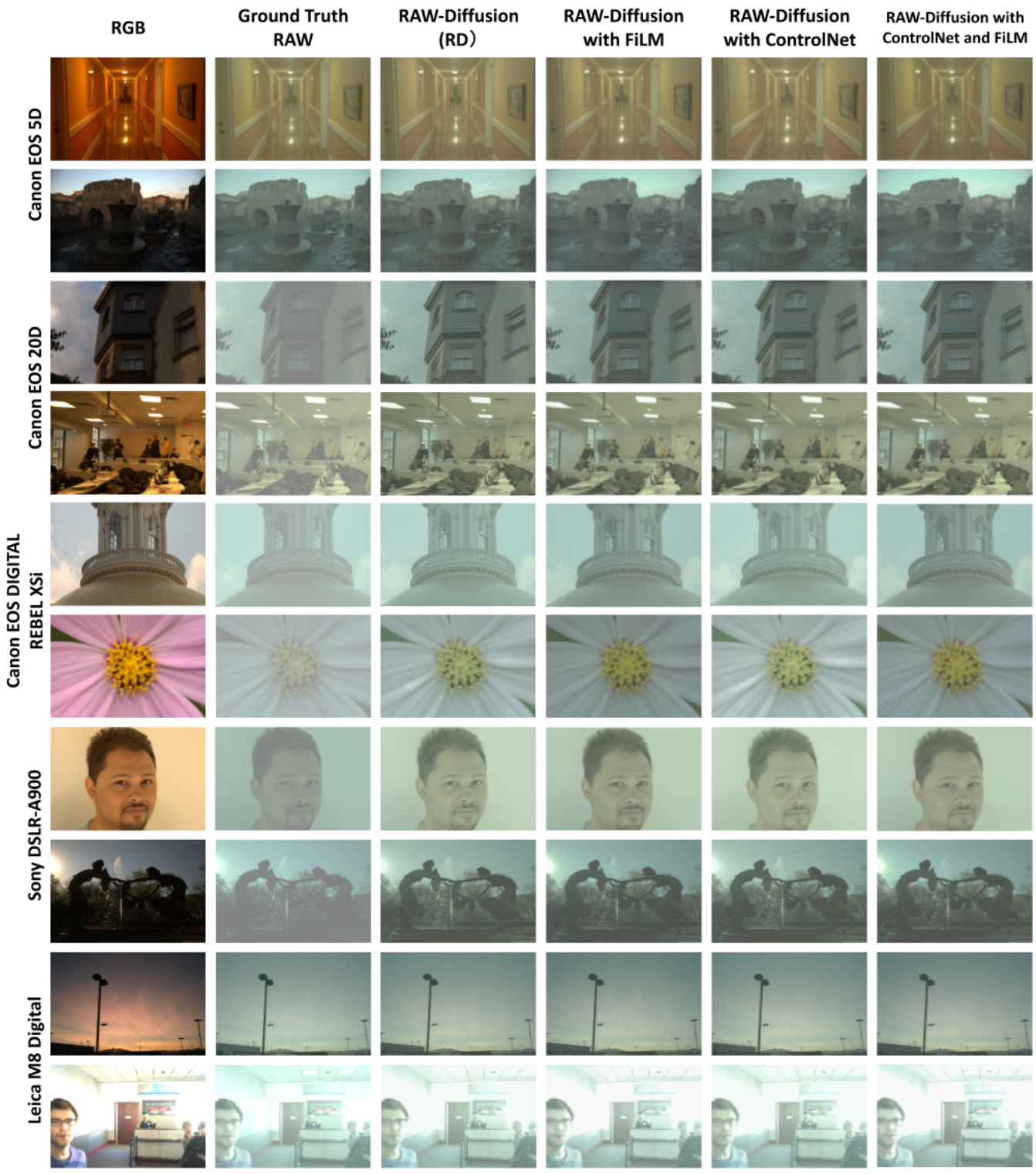

Fig. 7: Qualitative comparison of RAW reconstruction across different model variants.