MERIT: Multi-domain Efficient RAW Image Translation

Supplementary Material

6. Dataset Details

This section provides detailed descriptions of the datasets used in our study. We first introduce the **RAW-to-RAW Mapping Dataset** [1], which serves as a benchmark for cross-sensor RAW color mapping. Then, we present our newly collected **Multi-Domain RAW (MDRAW)** Dataset, designed for large-scale evaluation of multi-sensor RAW domain adaptation and alignment.

6.1. RAW-to-RAW Mapping Dataset

The RAW-to-RAW Mapping Dataset introduced by Afifi and Abuolaim [1] provides a benchmark for studying cross-sensor color-space mappings between different camera devices. It contains RAW images captured by two smartphone cameras—*Samsung Galaxy S9* and *Apple iPhone X*—and includes both unpaired and paired subsets.

The unpaired set consists of **392 RAW images** (196 per camera) captured under various scenes and illumination conditions. The dataset also provides **115 paired testing images** and a small **anchor set** of 22 paired images, which were used for semi-supervised training and evaluation.

Paired samples were generated using a **color-chart-based polynomial calibration** between the two camera RAW spaces. For each scene, corresponding patches were manually selected from chart regions and homogeneous areas to compute the mapping matrix, ensuring more robust color alignment.

All images are provided in **DNG format** with black-level subtraction and normalization applied. Metadata such as black/white levels and color matrices are included. The dataset supports supervised, unsupervised, and semi-supervised RAW-domain learning.

6.2. Multi-Domain RAW (MDRAW) Dataset

As one of our main contributions, we introduce **Multi-Domain RAW (MDRAW)**, a new dataset of RAW images captured by multiple cameras with distinct sensor characteristics. It is designed to facilitate research on cross-sensor RAW image processing, domain adaptation, and universal RAW representation learning.

MDRAW contains RAW images captured by **five cameras** with different sensor designs and optics: *Samsung Galaxy S23 Ultra*, *Huawei P30*, *iPhone 13 Pro*, *Nikon Z5*, and *Canon EOS Rebel T6*. The dataset includes both **unpaired** and **paired** RAW sets under diverse scenes and illumination conditions, along with their corresponding **sRGB references**.

For each camera, unpaired images were collected independently across indoor and outdoor environments, while paired data were captured under controlled conditions for cross-domain alignment. To construct a **pixel-level evaluation benchmark**, we further extracted spatially aligned patch pairs between images of the same scene taken by different devices using an extended **LoFTR-based correspondence pipeline** that combines dense matching, geometric verification, and patch-level synchronization.

All RAW files are stored in **DNG format** with corresponding metadata. Tab. 1 summarizes the camera sensors and data statistics, and Fig. 5 visualizes representative samples across devices.

7. MDRAW Construction via Cross-Domain LoFTR Matching

To quantitatively evaluate cross-domain RAW2RAW translation at the pixel level, we construct a high-quality benchmark dataset of paired RAW patches using real images captured by five distinct cameras: Huawei, Nikon, iPhone, Samsung, and Canon. Our goal is to extract spatially aligned patch pairs from images of the same scene taken by different devices, even in the presence of significant domain gaps in resolution, noise, exposure, and white balance.

7.1. Cross-Domain Patch Pairing Pipeline

Since no ground truth pixel-level alignment exists between images from different sensors, we adopt a feature-based correspondence approach. Specifically, we extend the LoFTR framework [29] to operate across domains, using a multi-stage pipeline that combines dense matching, geometric verification, spatial filtering, and patch-level synchronization. We briefly outline the core stages below:

Preprocessing and Normalization. Each RAW image is linearized by subtracting the camera-specific black level and normalizing by its dynamic range (white level − black level). The Bayer data is then packed into a $4-$channel RGGB format and spatially normalized via center-crop and downsampling to a fixed resolution. All images are rotated or flipped to ensure consistent spatial orientation across cameras.

LoFTR-based Feature Matching. We convert normalized RGGB images to pseudo-RGB representations and extract grayscale versions for matching. We use LoFTR [29], a detector-free dense matching method, to establish correspondences between grayscale image pairs. LoFTR excels in handling texture-poor regions and illumination inconsistencies, which are common issues in RAW domains.

Geometric Verification. Candidate matches are filtered via RANSAC [10] using homography estimation. Matches that are geometrically consistent form a reliable set of inliers for cropping.

Spatial Non-Maximum Suppression (NMS). To ensure spatial diversity and avoid redundancy, we apply NMS on the match locations using a minimum center-to-center distance threshold. This encourages the patch pairs to cover various regions and scene content.

Synchronized Cropping and Patch Extraction. For each retained match, we extract a 256×256 RGGB patch. Cropping is performed such that the patches are geometrically aligned at the matched keypoints, with optional correction for small shifts due to boundary constraints.

8. Evaluation Metrics

In this section, we provide detailed formulations of the four quantitative metrics used in our evaluation: Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and symmetric histogram-based KL divergence (KL Divergence). All metrics are computed on the denormalized RAW images in the linear intensity domain.

8.1. Mean Absolute Error (MAE)

MAE quantifies the average absolute difference between the predicted RAW image I_p and the reference RAW image I_r:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |I_p(i) - I_r(i)|, \tag{10}$$

where N is the total number of pixels. As RAW data reside in the linear intensity domain, MAE directly reflects radiometric consistency and low-level reconstruction accuracy. A lower MAE indicates better preservation of RAW signal characteristics.

8.2. Peak Signal-to-Noise Ratio (PSNR)

PSNR measures the ratio between the maximum possible signal and the distortion noise, derived from the mean squared error (MSE):

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N} (I_p(i) - I_r(i))^2, \tag{11}$$

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{L^2}{\mathrm{MSE}}\right), \tag{12}$$

where L denotes the maximum possible pixel value (set to 1.0 after normalization). PSNR values range from 0 to ∞, with ∞ indicating perfect similarity, implying no discernible difference between the compared images.

8.3. Structural Similarity Index Measure (SSIM)

SSIM assesses perceptual image quality by comparing local luminance, contrast, and structural components between I_p and I_r:

$$\mathrm{SSIM}(I_p, I_r) = \frac{(2\mu_p\mu_r + C_1)(2\sigma_{pr} + C_2)}{(\mu_p^2 + \mu_r^2 + C_1)(\sigma_p^2 + \sigma_r^2 + C_2)}, \tag{13}$$

where μ_p, μ_r are local means, σ_p^2, σ_r^2 are variances, and σ_{pr} is the covariance. Constants C_1 and C_2 stabilize the division. SSIM varies from 0 to 1, where a value of 0 indicates no structural similarity between the images, and 1 denotes identical local structures.

8.4. Symmetric KL Divergence (KL Divergence)

We compute a symmetric, histogram-based Kullback–Leibler divergence to evaluate distributional similarity between I_p and I_r:

$$D_{\mathrm{KL}}(p\|q) = \sum_i p_i \log\frac{p_i}{q_i}, \tag{14}$$

$$D_{\mathrm{sym}}(p,q) = \tfrac{1}{2}\big(D_{\mathrm{KL}}(p\|q) + D_{\mathrm{KL}}(q\|p)\big). \tag{15}$$

The probability distributions p and q are estimated from per-channel intensity histograms (256 bins, range [0,1]), with small-value regularization ε added to ensure numerical stability. This metric captures global statistical differences between RAW image distributions across sensors.

Together, these four metrics provide complementary perspectives on both pixel-level reconstruction accuracy and overall distributional alignment in the RAW domain.

9. More Results

Fig. 6 illustrates the diversity and cross-domain characteristics of the proposed MDRAWdataset. Each row corresponds to a specific camera sensor, including Huawei P30, Nikon Z5, iPhone 13 Pro Max, Samsung S23 Ultra, and Canon EOS Rebel T6, showcasing both RAW and their corresponding RGB images. The RAW samples reveal substantial variations in tone, exposure, and noise characteristics across devices, even under similar scene conditions. These differences stem from the distinct spectral response functions, sensor noise patterns, and in-camera processing pipelines of each sensor. In contrast, the RGB samples appear more visually consistent due to the influence of proprietary ISP modules. This highlights the challenge and necessity of direct RAW2RAW translation to enable fair cross-domain vision tasks without relying on device-specific ISP outputs. Fig. 6 also underscores the importance of building datasets like MDRAWto support learning models capable of generalizing across multiple RAW domains.

Fig. 6 provides a visual comparison of RAW-to-RAW translation results across five challenging real-world scenes,

Figure 5. **Sample images from our MDRAW benchmark.** (a). RAW images captured under various lighting and scenes from five different camera sensors. (b). Example of aligned multi-domain RAW captures of the same scene, used for evaluation in cross-domain RAW2RAW translation.

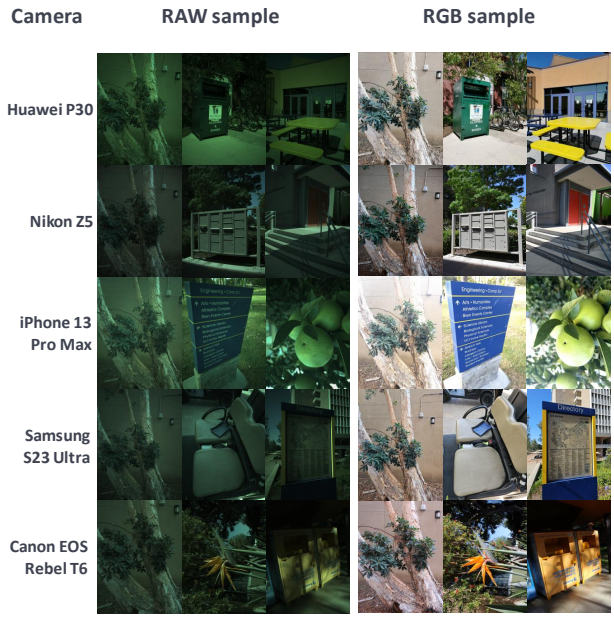

Figure 6. **Visualization of more captured images in MDRAW.**

highlighting the performance of different methods. Each row corresponds to a different scene, with the source image on the left, followed by translated outputs, and the ground truth on the right. For each method, the translated patch and its corresponding absolute error map (visualized via MAE) are shown, along with PSNR and SSIM metrics.

Across all examples, MERIT consistently produces translated outputs that are both visually faithful and quantitatively superior. In the first scene (world map), MERIT accurately reconstructs the fine textural patterns and subtle gradients with the highest PSNR (40.16) and SSIM (0.96), significantly outperforming other methods that either blur the details or introduce artifacts. In the second scene (indoor poster), MERIT achieves a near-perfect reconstruction (PSNR: 39.03, SSIM: 0.99), preserving text clarity and illumination distribution that others fail to retain. In the third and fourth scenes, characterized by repetitive textures and sharp edges, MERIT demonstrates superior structure preservation. Competing methods such as CycleGAN and Rawformer often produce distorted patterns or overly smoothed outputs, as evident in the high error maps. Conversely, MERIT maintains edge consistency and semantic accuracy, with notably lower MAE. Finally, in the fifth example (indoor hallway), MERIT again achieves the best perceptual quality and the most faithful reconstruction of lighting and geometry. The overall results demonstrate MERIT's strong generalizability across diverse domains and scenes, as well as its ability to balance perceptual quality and pixel-level fidelity more effectively than prior methods.

No. of Domains	Method	Parameters (M)	Training Iteration (K) / Training Time (Minutes)	Avg MAE↓	Avg PSNR↑	Avg SSIM↑	Avg KL Div↓
3	UVCGAN [30]	186	228 / 1955	0.039	28.03	0.72	2.72
	Xie *et al.* [33]	**33.6**	266 / 1338	0.038	27.82	0.73	2.33
	MERIT (Ours)	58.7	120 / 1400	**0.031**	**29.60**	**0.76**	**1.76**
4	UVCGAN [30]	372	476 / 3956	0.038	27.85	0.72	2.66
	Xie *et al.* [33]	67.2	507 / 2055	0.037	28.31	0.72	2.35
	MERIT (Ours)	**58.7**	165 / 1925	**0.033**	**29.08**	**0.74**	**2.07**
5	UVCGAN [30]	620	748 / 6581	0.038	28.10	0.72	2.48
	Xie *et al.* [33]	112	901 / 3428	0.037	28.23	0.73	**2.28**
	MERIT (Ours)	**58.9**	180 / 2100	**0.036**	**28.42**	**0.74**	2.30

Table 5. **Comparison of different models under increasing numbers of domains.** MERIT demonstrates both competitive performance and efficiency.

Source \ Target	Samsung	Huawei	iPhone	Nikon
Samsung	–	0.025 / 31.23 / 0.77 / 1.33	0.036 / 29.02 / 0.76 / 1.78	0.035 / 28.45 / 0.72 / 2.11
Huawei	0.029 / 30.11 / 0.74 / 1.70	–	0.033 / 29.12 / 0.77 / 2.34	0.032 / 29.42 / 0.74 / 1.95
iPhone	0.035 / 28.88 / 0.73 / 1.56	0.030 / 29.69 / 0.76 / 1.95	–	0.039 / 27.52 / 0.71 / 2.38
Nikon	0.040 / 27.22 / 0.68 / 3.00	0.034 / 28.71 / 0.73 / 2.69	0.043 / 26.51 / 0.71 / 3.34	–

Table 6. **4×4 Cross-domain RAW2RAW translation results on MDRAW.** Each cell reports results for translation from a source domain (column) to a target domain (row). Each line contains four metrics in the order of **MAE**(↓)/ **PSNR**(↑)/ **SSIM** (↑)/ **KL Divergence**(↓).

To evaluate the generalization and scalability of our proposed MERIT framework, we compare its cross-domain translation performance when trained on 3 domains (Samsung, Huawei, iPhone) versus 4 domains (adding Nikon). As shown in Tab. 7 and Tab. 6, MERIT consistently maintains strong translation performance across both settings.

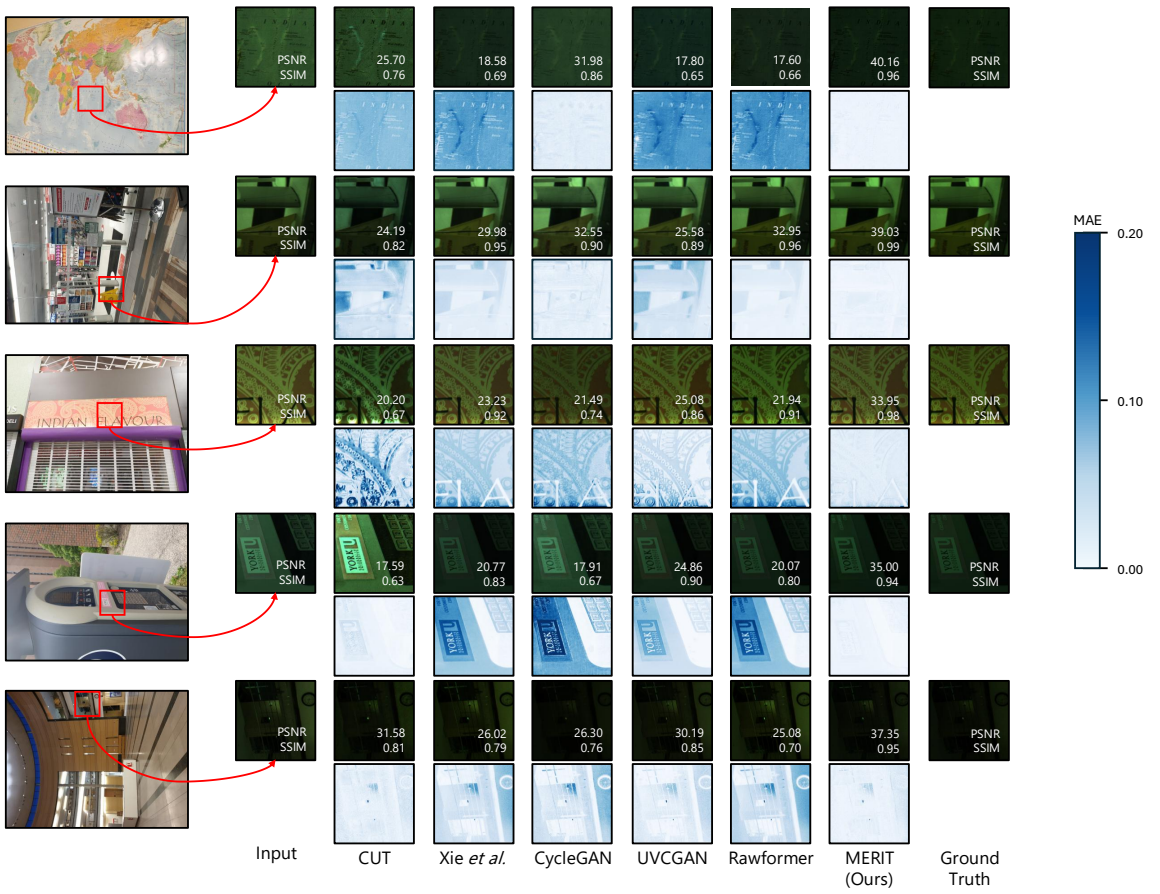

Figure 7. More qualitative results on the RAW-to-RAW mapping dataset.

Source \ Target	Samsung	Huawei	iPhone
Samsung	–	0.025 / 31.24 / 0.77 / 1.36	0.036 / 29.02 / 0.76 / 1.63
Huawei	0.029 / 30.04 / 0.75 / 1.48	–	0.034 / 28.96 / 0.77 / 2.27
iPhone	0.035 / 28.53 / 0.72 / 1.65	0.036 / 29.02 / 0.76 / 1.63	–

Table 7. **3×3 Cross-domain RAW2RAW translation results on MDRAW.** Each cell reports results for translation from a source domain (column) to a target domain (row). Each line contains four metrics in the order of **MAE**(↓)/ **PSNR**(↑)/ **SSIM** (↑)/ **KL Divergence**(↓).

Notably, the average PSNR across all 3×3 transfers remains high (e.g., 30.04→31.24 for Samsung→Huawei), with marginal variation from the corresponding 4×4 results. For instance, the translation from Samsung to Huawei achieves 31.24 dB PSNR in both settings, while Huawei to iPhone reaches 29.12 dB in the 4-domain setup and 28.96 dB in the 3-domain setup. Similarly, SSIM values remain stable across settings. These results indicate that MERIT generalizes well to varying numbers of domains without significant degradation. Furthermore, the stable performance with the added domain (Nikon) demonstrates MERIT 's scalability: the unified model scales gracefully to more diverse camera domains while maintaining high-quality translation, supporting its deployment in real-world multi-sensor environments.

We conduct a thorough ablation study to examine the sensitivity of MERIT to the weighting of its two proposed loss terms, $\mathcal{L}_{\text{noise}}$ and $\mathcal{L}_{\text{cycle-SSIM}}$, as shown in Tab. 8. For $\mathcal{L}_{\text{noise}}$, performance improves steadily as the weight λ increases from 0.25 to 1.0, reaching optimal results at $\lambda = 1$ with a PSNR of 33.60, SSIM of 0.8806, and MAE of 0.0185. Beyond this point, further increasing the weight degrades performance, suggesting that overly emphasizing noise alignment may compromise semantic fidelity or overall reconstruction quality. A similar trend is observed for $\mathcal{L}_{\text{cycle-SSIM}}$. The optimal setting is again $\lambda = 0.1$, producing identical peak performance across all three metrics. Lower weights underutilize the semantic supervision provided by SSIM, while larger values likely over-constrain the reconstruction, leading to degraded perceptual quality. Overall, both proposed loss terms contribute positively to

Loss Term	Setting	MAE ↓	PSNR ↑	SSIM ↑
$\mathcal{L}_{noise}$	$\lambda = 0.25$	0.0203	32.81	0.8668
	$\lambda = 0.75$	0.0199	33.27	0.8707
	$\lambda = 1$	**0.0185**	**33.60**	**0.8806**
	$\lambda = 1.5$	0.0195	33.24	0.8769
	$\lambda = 5$	0.0212	32.36	0.8576
$\mathcal{L}_{cycle-SSIM}$	$\lambda = 0.02$	0.0222	32.21	0.8560
	$\lambda = 0.05$	0.0200	33.18	0.8735
	$\lambda = 0.08$	0.0211	32.54	0.8683
	$\lambda = 0.1$	**0.0185**	**33.60**	**0.8806**
	$\lambda = 0.2$	0.0192	33.18	0.8695

Table 8. **Ablation study on loss weight sensitivity.** We evaluate the impact of loss weighting coefficients λ for the proposed noise modeling loss $\mathcal{L}_{\text{noise}}$ and the cycle consistency SSIM loss $\mathcal{L}_{\text{cycle-SSIM}}$. Each cell reports average performance on the RAW-to-RAW mapping dataset. Best results for each loss term are highlighted in bold.

performance when appropriately weighted, and the consistent peak at $\lambda = 1$ and $\lambda = 0.1$, respectively, highlights the effectiveness and stability of these components in the MERIT.