MERIT: Multi-domain Efficient RAW Image Translation

Wenjun Huang[1] Shenghao Fu[2] Yian Jin[1] Yang Ni[3] Ziteng Cui[4]
Hanning Chen[1] Yirui He[1] Yezi Liu[1] Sanggeon Yun[1] SungHeon Jeong[1]
Ryozo Masukawa[1] William Youngwoo Chung[1] Mohsen Imani[1]

[1]University of California, Irvine [2]University of Pennsylvania
[3]Purdue University Northwest [4]The University of Tokyo

m.imani@uci.edu

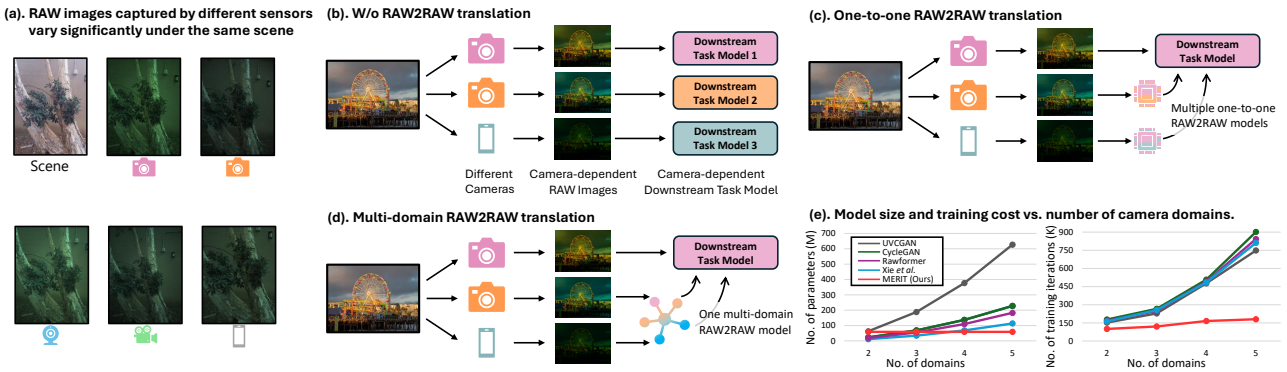

Figure 1. **Overview of the challenges and paradigms in multi-domain RAW-to-RAW (RAW2RAW) translation.** (a). RAW images captured under the same scene vary across sensors due to differing spectral sensitivity and noise characteristics. (b). Without RAW2RAW translation, domain-specific task models must be independently trained for each camera. (c). One-to-one translation enables shared task models, but requires a separate translation model for every source-target pair, leading to scalability issues. (d). Our proposed method, MERIT, introduces a unified multi-domain RAW2RAW translator that supports all domains with a single model. (e). MERIT achieves superior scalability in both parameter count and training iterations as the number of camera domains increases, outperforming prior methods.

Abstract

RAW images captured by different camera sensors exhibit substantial domain shifts due to varying spectral responses, noise characteristics, and tone behaviors, complicating their direct use in downstream computer vision tasks. Prior methods address this problem by training domain-specific RAW-to-RAW translators for each source–target pair, but such approaches do not scale to real-world scenarios involving multiple types of commercial cameras. In this work, we introduce MERIT, ***the first unified framework for multi-domain RAW image translation****, which leverages a single model to perform translations across arbitrary camera domains. To address domain-specific noise discrepancies, we propose a sensor-aware noise modeling loss that explicitly aligns the signal-dependent noise statistics of the generated images with those of the target domain. We further enhance the generator with a conditional multi-scale large kernel attention module for improved context and sensor-aware feature modeling. To facilitate standardized evaluation, we introduce MDRAW,* ***the first dataset tailored for multi-domain RAW image translation****, comprising both paired and unpaired RAW captures from five diverse camera sensors across a wide range of scenes. Extensive experiments demonstrate that MERIT outperforms prior models in both quality (+5.56 dB) and scalability (80% reduction in training iterations). Our code is available here.*

1. Introduction

In recent years, the potential of camera RAW data has been increasingly explored and leveraged across a wide range of downstream vision tasks, including low-level restoration (e.g., super-resolution [7, 34], low-light imaging [3, 15], and reflection removal [17]), high-level perception (e.g., object detection [8, 31], instance segmentation [4, 13], and pose estimation [19]), and 3D reconstruction [9, 16, 23]. Alongside the growing research interest in RAW-based vision tasks, commercial cameras and smartphones have also

been rapidly evolving. For instance, flagship smartphones introduce diverse RAW formats, such as Apple ProRAW™, Samsung Expert RAW™, and standard DNG options on Android devices [28, 36], reflecting the growing heterogeneity of RAW data across mobile imaging systems. The diversity in RAW formats also leads to significant inter-device variability in captured data, even when capturing the same scene; this variability can significantly affect color, dynamic range, and noise structure (Fig. 1 (a)). Therefore, multiple camera-specific downstream task models are needed (Fig. 1 (b)). This domain shift roots in the physics of image formation [24]. An ideal image formation is rooted in:

$$I(x) = \int_{\omega} R_c(\lambda) S(x, \lambda) L(\lambda) \mathbf{d}\lambda, \tag{1}$$

where λ represents the wavelength, ω denotes the visible spectrum. The term $S(x, \lambda)$ represents the scene's spectral response at pixel x, and $L(\lambda)$ is the lighting in the scene. R_c is the camera's spectral response, and c is the color channel, which are the variables that change in this case.

Recent work [1, 25, 27, 33] addressed this challenge through domain-specific RAW-to-RAW (RAW2RAW) translation, which seeks to map RAW images from one camera's domain to another, enabling the reuse of a fixed downstream task model trained on a particular camera (Fig. 1 (c)). These methods rely on supervised or unsupervised learning to perform domain alignment, but are inherently limited to one-to-one mappings between two camera domains. As a result, they do not scale well in scenarios involving multiple camera domains, where training a separate model for each camera pair becomes costly and impractical.

To overcome this scalability bottleneck, we propose the first **unified framework for multi-domain efficient RAW image translation** (MERIT). MERIT is capable of translating a RAW image from any source domain to any desired target domain, using a single model conditioned on domain embeddings (Fig. 1 (d)). MERIT introduces a unified multi-domain RAW2RAW translation architecture that enables one-to-many and many-to-many transfers using a single model. It maintains a superior translation performance compared with prior work while improving the scalability as the number of domains increases (Fig. 1 (e)).

While existing methods attempt to learn RAW domain characteristics implicitly through adversarial training, we observe that a significant portion of domain-specific variation arises from camera-dependent noise patterns. Therefore, motivated by the physical Poisson-Gaussian nature of RAW noise [11], we introduce a noise-aware loss that guides the model to match the statistical noise properties of the target domain. This shift from implicit to explicit modeling improves both the realism and fidelity of the translated RAW images, especially in noise-sensitive regions.

In addition, to enable effective context aggregation for RAW2RAW translation, we introduce a novel Multi-Scale Large Kernel Attention (MS-LKA) module that enhances spatial perception and enables domain-adaptive modulation, allowing the model to capture fine-grained local patterns and long-range dependencies simultaneously. This is particularly well-suited for RAW images, where signal distributions and noise characteristics vary across spatial scales and sensor types. This design not only improves the model's capacity to learn sensor-specific mappings but also maintains architectural efficiency with minimal parameter increase.

To facilitate training and evaluation in this new setting, we also introduce **the first benchmark dataset for multi-domain RAW translation** (named MDRAW), composed of 519 unpaired RAW images captured across a diverse set of camera sensors, lighting conditions, and scenes. MDRAW also includes multiple aligned image groups (285 images / 57 groups) across camera domains, enabling quantitative evaluation. The key contributions of the paper are:

- We propose MERIT, the first multi-domain RAW2RAW translation framework that generalizes RAW image translation across multiple camera domains using a single model. Our architecture can flexibly convert RAW images between arbitrary domain pairs during inference.
- We introduce a novel explicit noise modeling that enforces statistical consistency of signal-dependent noise across domains, improving translation fidelity under challenging conditions.
- We employ a multi-scale large kernel attention to enhance the spatial perception, enabling domain-adaptive modulation given target domain information.
- We construct and release MDRAW, the first benchmark dataset of multi-domain RAW images for training and evaluating RAW2RAW translation at scale.
- Extensive experiments on existing and MDRAW demonstrate that MERIT significantly outperforms prior models in both quality (+5.56 dB) and scalability (2× smaller).

2. Method

2.1. Overall Framework

Let $\mathcal{X}$ and $\mathcal{Y}$ be the sets of RAW images and RAW domains, respectively. Given an image $I^a \in \mathcal{X}$ from the source domain $a \in \mathcal{Y}$, our goal is to train a **single** generator $\mathcal{G}$ that can generate the corresponding RAW images $\hat{I}^b$ of an arbitrary target domain except the source domain $b \in \mathcal{Y} \setminus \{a\}$. We use a learnable style encoder $\mathcal{E}$ to generate domain-specific style embedding for each domain and train $\mathcal{G}$ to reflect the style embeddings. Fig. 2 (a) illustrates an overview of MERIT, which consists of three key modules.

Generator (Fig. 2 (c))**.** The generator $\mathcal{G}$ translates an input image I^a into an output image $\hat{I}^b = G(I^a, s_b)$ reflecting the domain-specific style embedding s_b, which is provided by $\mathcal{E}$. s_b is designed to represent the style of a specific domain

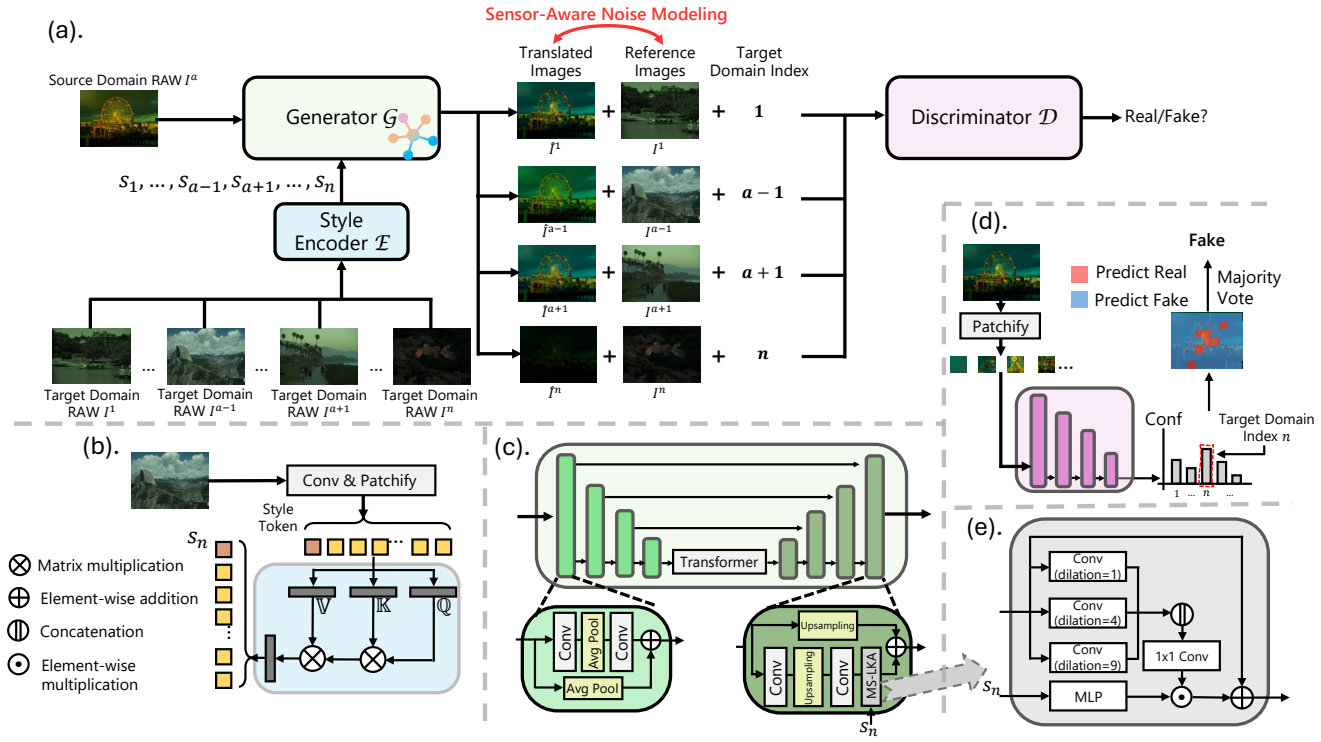

Figure 2. **MERIT Framework Overview.** (a). Overall architecture of MERIT. (b). The style encoder $\mathcal{E}$ extracts domain-specific embeddings from target RAW exemplars using a transformer-based architecture. (c). The generator $\mathcal{G}$ synthesizes target-domain RAW images. (d). The discriminator $\mathcal{D}$ operates on image patches and predicts the realism of each patch, using majority voting to make final decisions. (e). The proposed MS-KLA module captures multi-scale features and modulates them via style-aware channel attention.

b, which removes the necessity of providing a reference image to $\mathcal{G}$ and allows $\mathcal{G}$ to translate images of all domains.

Style encoder (Fig. 2 (b))**.** Given an image I^b from domain b, our encoder $\mathcal{E}$ extracts the style embedding $s_b = \mathcal{E}(I^b)$ of I^b. The extracted s_b is domain-specific and content-independent. This means any images from the same domain should result in a similar embedding, while it may still depend on some characters (e.g., brightness). We discuss this in Sec. 3.

Discriminator (Fig. 2 (d))**.** We also incorporate a discriminator $\mathcal{D}$ to supervise the training of $\mathcal{G}$. $\mathcal{D}$ receives as input the translated image, reference images from the target domain, and the target domain index, and determines whether the input is a real sample from the target domain or a synthesized output. Following the patch-based adversarial strategy of [14], each image is partitioned into smaller patches, which are independently evaluated by $\mathcal{D}$ to classify each as real or fake. The image-level prediction is obtained via majority voting over the patch-level decisions, encouraging $\mathcal{G}$ to produce globally coherent and locally realistic outputs.

2.2. Sensor-Aware Noise Modeling (SANM)

Unlike sRGB image-to-image translation, only generating visually realistic images is not sufficient for RAW2RAW translation [37]; capturing the sensor-dependent noise statistics is essential to preserve **physical plausibility**, especially under high ISO or low-light conditions. Traditional GAN-based translation frameworks attempt to learn such characteristics implicitly, but often fail to replicate domain-specific noise patterns accurately.

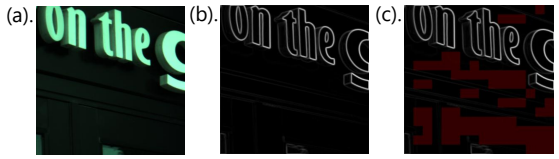

Figure 3. **Visualization of the sensor-aware noise modeling process.** (a). Input RAW image. (b). Gradient magnitude map computed via the Sobel operator. (c). Red-shaded regions indicate low-texture patches selected for sensor noise profile estimation.

Noise Modeling in RAW. In RAW data, which remains in the linear light domain, the sensor noise can be well-approximated by a Poisson-Gaussian noise model [11]:

$$\mathrm{Var}(x) = \alpha \cdot z + \beta, \tag{2}$$

where x is the noisy observed signal, z is the true scene intensity, α captures the signal-dependent shot noise, and β models the signal-independent read noise. Our goal is to ensure that the translated RAW images exhibit the same noise

profile (i.e., the functional dependence between intensity and variance) as the real RAW images from the target sensor domain. To achieve this, we introduce a differentiable histogram-based noise loss, $\mathcal{L}_{noise}$, which aligns the noise characteristics of the generated image with those of the real target-domain RAW images. Given an input RAW image (Fig. 3 (a)), we extract small, non-overlapping patches and compute the mean intensity and robust variance estimate via median absolute deviation for each patch. To ensure that the variance reflects sensor noise rather than texture, we use a Sobel gradient magnitude filter to identify flat regions, i.e., the patches with minimal structural content. (Fig. 3 (b) visualizes the calculated gradient magnitude map.) Patches are retained if their average gradient falls below a percentile threshold, as demonstrated in Fig. 3 (c), making the influence from local texture minimal, and the variance is dominated by sensor noise.

Let μ_i and σ_i^2 denote the mean intensity and variance of the i-th flat patch. We bin the patches based on μ_i into fixed-width intensity bins (e.g., 100 bins in $[0, 1]$), and compute the average variance per bin. This yields a noise histogram $\mathcal{H}_{fake} \in \mathbb{R}^{C\times B}$ for the generated image, where C is the number of channels and B the number of bins. The target noise profile $\mathcal{H}_{real}$ is precomputed from real images in the target domain using the same procedure and stored as a lookup table. $\mathcal{L}_{noise}$ compares the two histograms:

$$\mathcal{L}_{noise} = \frac{1}{BC}\sum_{c=1}^{C}\sum_{b=1}^{B} |\mathcal{H}_{fake}[c,b] - \mathcal{H}_{real}[c,b]| \cdot \mathbf{1}_{\text{valid}}[c,b], \tag{3}$$

where $\mathbf{1}_{\text{valid}}$ masks out empty bins to avoid degenerate gradient flow. $\mathcal{L}_{noise}$ is fully differentiable, robust to image content, and grounded in a physical noise model. It enables the generator to learn not only the appearance style of the target sensor domain but also its statistical noise behavior, resulting in more realistic and sensor-faithful outputs.

2.3. Multi-Scale Large Kernel Attention

RAW images exhibit unique characteristics compared to sRGB data, such as spatially correlated illumination patterns, sensor-specific tone responses, and signal-dependent noise distributions. Modeling these requires a network to perceive global spatial relationships while maintaining the precise local structure of the RAW signal. Conventional convolutions fail to capture such long-range dependencies, whereas transformer-based attention models are computationally expensive for high-resolution inputs. To bridge this gap, we extend the concept of large kernel attention [12], introducing a multi-scale large kernel attention (MS-LKA) module on the upsampling path of $\mathcal{G}$. MS-LKA aggregates features at multiple spatial scales without relying on expensive self-attention, as visualized in Fig. 2 (e).

Multi-Dilation Feature Extraction. Given an input feature map $F_{in} \in \mathbb{R}^{C\times H\times W}$, we apply three parallel branches of depthwise convolutions with large kernels and distinct dilation rates. Each branch captures context at a different receptive field scale, producing intermediate features $F_1, F_2, F_3 \in \mathbb{R}^{C\times H\times W}$, which are then concatenated channel-wise and compressed via a 1×1 convolution to maintain dimensionality. The resulting fused feature map $F_{concat} = \text{Conv}_{1\times1}([F_1; F_2; F_3])$ integrates spatial cues across a wide range of receptive fields.

Style-Modulated Channel Attention. To adaptively weight the multi-scale features according to the target domain, we introduce a style-conditioned attention mechanism. The style embedding s, extracted from the style encoder, is processed by a lightweight feed-forward network (FFN) to produce channel-wise attention weights $A_s \in \mathbb{R}^C$. These weights are applied to the fused feature map via element-wise multiplication: $F_{out} = A_s \odot F_{concat}$.

This modulation allows $\mathcal{G}$ to dynamically emphasize domain-relevant channels conditioned on the desired sensor style embedding. Compared to conventional convolution blocks or static attention layers, our MS-LKA module offers three distinct advantages for RAW2RAW translation: it expands the effective receptive field while maintaining convolutional inductive bias; it enables domain-aware adaptation through style-guided attention; and it integrates efficiently into our generator without significant parameter overhead.

2.4. Loss Functions.

Given an image I^a, we train our framework using the following objectives.

Adversarial objective. During training, we sample a target domain $b \in \mathcal{Y} \setminus \{a\}$, and randomly select an RAW image I^b from this domain. We compute its style embedding $s_b = \mathcal{E}(I^b)$, and generate a translated image $\hat{I}^b = \mathcal{G}(I^a, s_b)$. The discriminator $\mathcal{D}$ is trained to correctly classify real RAW images and distinguish them from generated ones using the adversarial loss:

$$\begin{aligned}\mathcal{L}_{adv}^{\mathcal{D}} =& \mathbb{E}_{I^a}[\log \mathcal{D}(a|I^a)] + \\ & \mathbb{E}_{I^a, I^b}[\log(1 - \mathcal{D}(b|\mathcal{G}(I^a, \mathcal{E}(I^b))))],\end{aligned} \tag{4}$$

$$\mathcal{L}_{adv}^{\mathcal{G}} = \mathbb{E}_{I^a, I^b}[\log \mathcal{D}(b|\mathcal{G}(I^a, \mathcal{E}(I^b)))], \tag{5}$$

where $\mathcal{D}(a|\cdot)$ denotes the output of $\mathcal{D}$ corresponding to domain a. $\mathcal{G}$ learns to utilize s_b and generate an image $\mathcal{G}(I^a, s_b)$ that is indistinguishable from real RAW images of the domain b.

Style reconstruction. To enforce $\mathcal{G}$ to utilize the style embedding, we apply a style reconstruction loss:

$$\mathcal{L}_{style} = \mathbb{E}_{I^a, I^b}\left[\left\|\mathcal{E}(I^b) - \mathcal{E}(\mathcal{G}(I^a, \mathcal{E}(I^b)))\right\|_1\right] \tag{6}$$

Specifically, after generating the image $\hat{I}^b = \mathcal{G}(I^a, \mathcal{E}(I^b))$, we re-extract the style embedding from the generated image using the same encoder $\mathcal{E}$. The loss minimizes the L1 distance between the original and reconstructed style em-

beddings, encouraging style consistency.

Preserving source semantics and content. To ensure that $\mathcal{G}$ preserves domain-invariant characteristics, such as content and layout, we adopt a cycle consistency L1 loss:

$$\mathcal{L}_{cycle-L1} = \mathbb{E}_{I^a, I^b}\left[\left\| I^a - \mathcal{G}(\mathcal{G}(I^a, \mathcal{E}(I^b)), \mathcal{E}(I^a))\right\|_1\right] \tag{7}$$

Specifically, given a source image I^a and a target style embedding $\mathcal{E}(I^b)$, we first generate the translated image then translate it back to the source domain using the original style embedding $\mathcal{E}(I^a)$, and minimize the L1 distance between the reconstructed image and the original input.

Empirically, unlike sRGB image translation, we observe that solely relying on pixel-wise L1 loss is insufficient for RAW2RAW translation to preserve structural and semantic content. The L1 term enforces intensity alignment but fails to maintain consistent texture and fine-grained sensor details. **We introduce a cycle-consistency SSIM loss that complements the L1 reconstruction** by encouraging higher perceptual and structural fidelity between the original and cycle-reconstructed RAW images:

$$\mathcal{L}_{\text{cycle-SSIM}} = \mathbb{E}_{I^a, I^b}\left[1 - \text{SSIM}\left(I^a, \mathcal{G}(\mathcal{G}(I^a, \mathcal{E}(I^b)), \mathcal{E}(I^a))\right)\right], \tag{8}$$

The full loss function is constructed as:

$$\begin{aligned}\mathcal{L}_{total} =& \lambda_1 \mathcal{L}_{noise} + \lambda_2 \mathcal{L}^{\mathcal{D}}_{adv} + \lambda_3 \mathcal{L}^{\mathcal{G}}_{adv} \\ &+ \lambda_4 \mathcal{L}_{cycle-L1} + \lambda_5 \mathcal{L}_{cycle-SSIM},\end{aligned} \tag{9}$$

where $\lambda_1, \cdots, \lambda_5$ are emperical hyper-parameters.

3. Experiments

Datasets. In this section, we provide a description of the datasets used in this work. Particularly, we use two datasets: **(i).** RAW-to-RAW mapping dataset [1]. It contains a total of 392 unpaired RAW images (196 from each of the Samsung Galaxy S9 and iPhone X). Additionally, the dataset includes 115 paired testing RAW images from each camera for evaluation. **(ii).** our collected dataset, Multi-Domain RAW (MDRAW). As one of our contributions, we propose MDRAW, a new dataset of RAW images captured by five cameras with different sensors: Samsung Galaxy S23 Ultra, Huawei P30, iPhone 13 Pro, Nikon Z5, and Canon EOS Rebel T6. Fig. 5 (a) shows example RAW images from each camera. To the best of our knowledge, there is no publicly available dataset of RAW images captured by multiple cameras that meets our setup requirements (i.e., containing both unpaired and paired RAW image sets under diverse illuminations and scenes). MDRAW consists of unpaired and paired RAW images for each camera with their corresponding sRGB images. Tab. 1 summarizes the sensor and statistics of MDRAW. More details about the datasets are provided in the supplementary material.

To construct a high-quality, pixel-level evaluation benchmark across multiple RAW domains, we also extract spatially aligned patch pairs from images of the same scene taken by different devices (as shown in Fig. 5 (b)). Since no ground truth pixel-level alignment exists between images from different sensors, we adopt a feature-based correspondence approach to identify regions suitable for evaluation. Specifically, we extend the LoFTR [29] to operate across domains, using a multi-stage pipeline that combines dense matching, geometric verification, spatial filtering, and patch-level synchronization. The whole construction process is discussed in detail in the supplementary material.

Camera	Sensor	No. of unpaired images
Samsung Galaxy S23 Ultra	ISOCELL HP2	81
Huawei P30	Sony IMX650 Exmor RS	119
iPhone 13 Pro Max	Sony IMX703	98
Nikon Z5	Sony IMX128	114
Canon EOS Rebel T6	DIGIC 4+	107

Table 1. **Summarization of our collected dataset (MDRAW).**

Metrics. We assess the models utilizing three commonly accepted metrics: mean absolute error (MAE), peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [32], and KL divergence. Details about the adopted metrics are discussed in the supplementary material.

Training Details. We train our model with a batch size of 8 using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.99$), on a NVIDIA H200 GPU. The learning rate is fixed at 1×10^{-4} for all network components, including the generator, discriminator, and style encoder. The training process is conducted for 200,000 iterations. We apply random horizontal flipping ($p = 0.5$) as the data augmentation strategy. All input RAW images are pre-processed into 256×256 patches. All models are trained on unpaired images and evaluated on paired image sets to assess cross-domain translation performance. During evaluation, we adopt a brightness-based reference image selection strategy. Specifically, for each input, we select the target-domain image from the trainset with the most similar brightness as the reference for style embedding extraction. Brightness similarity is quantified using channel-wise means: we compute the spatial mean of each of the four Bayer channels to form a 4-dimensional channel-mean vector, which is then used for comparison. In addition to these settings, we use the following loss weights in our full objective: $\lambda_{\text{reg}} = 1$, $\lambda_{\text{cyc}} = 10$, $\lambda_{\text{sty}} = 1$, $\lambda_{\text{ds}} = 5$, $\lambda_{\text{id}} = 1.0$, $\lambda_{\text{noise}} = 1.0$, and $\lambda^{\text{SSIM}}_{\text{cyc}} = 0.1$.

3.1. Results

Fig. 4 presents a visual comparison of different methods on RAW-to-RAW mapping dataset. Odd-numbered rows depict the RAW outputs, while even-numbered rows display the corresponding absolute error maps with respect to the ground truth. MERIT consistently produces translated images with fewer perceptual artifacts and lower reconstruction errors. Compared to prior methods, MERIT generates outputs with more faithful scene illumination and finer structural alignment, especially in regions with sharp edges or strong color contrast. The error maps show that MERIT leads to smaller residuals, in terms of magnitude and spatial extent, indicating superior pixel-level accuracy.

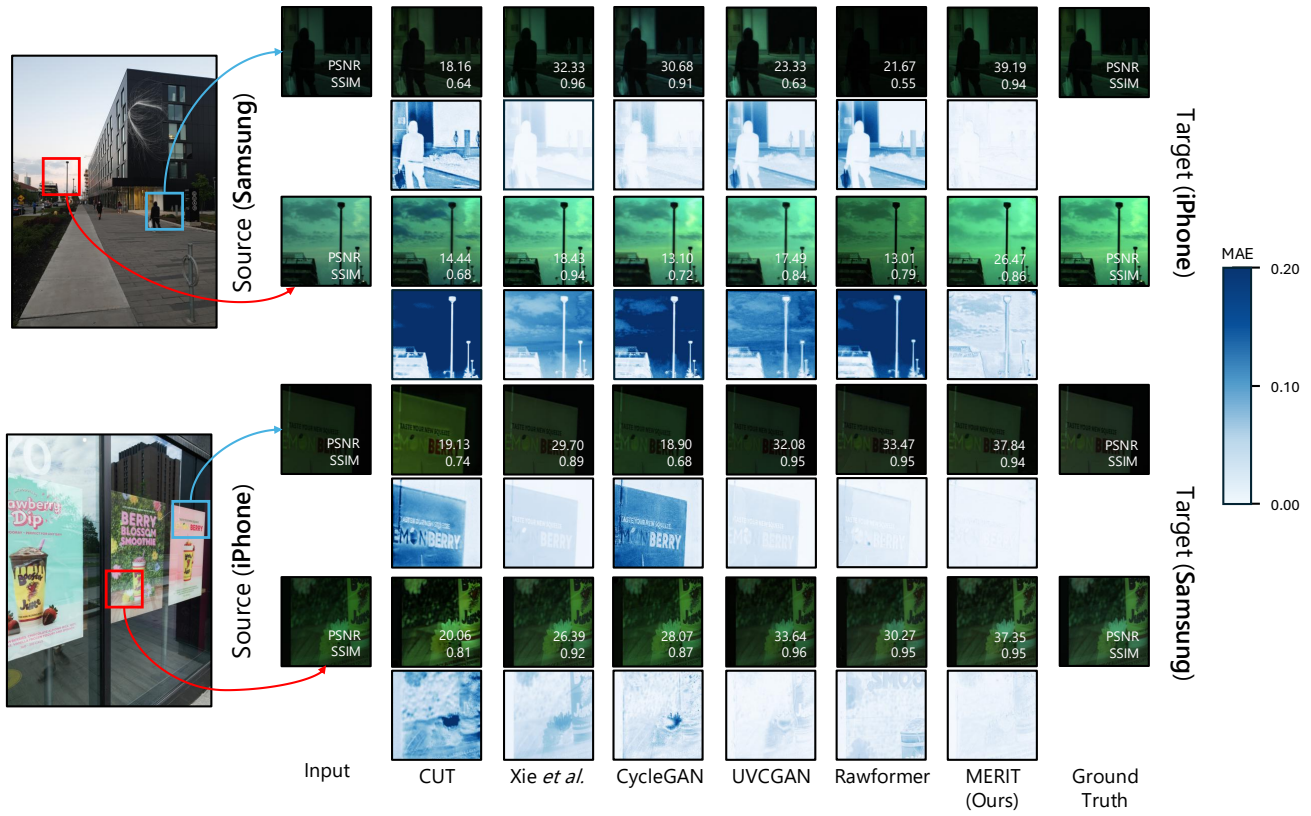

Figure 4. **Qualitative results on the RAW-to-RAW mapping dataset.** Each group of images (from left to right) shows: the input source RAW image, predictions from CUT [26], Xie *et al*. [33], CycleGAN [38], UVCGAN [30], Rawformer [27], our method MERIT, and the target ground truth. Odd-numbered rows display the translated RAW outputs, while even-numbered rows visualize the corresponding absolute error maps with respect to the ground truth.

Training Paradigm	Model	Samsung-S9→iPhone-X			iPhone-X→Samsung-S9		
		PSNR↑	SSIM↑	MAE↓	PSNR↑	SSIM↑	MAE↓
Non-learning	Global calibration (3×3) [24]	24.52	0.71	0.049	17.03	0.51	0.160
	Global calibration (poly) [24]	24.88	0.72	0.048	16.88	0.50	0.160
	FDA [35]	20.95	0.48	0.060	19.18	0.47	0.090
Semi-supervised	Afifi *et al.* [1]	29.65	0.89	0.027	28.58	0.90	0.033
Unsupervised	CycleGAN [38]	24.55	0.76	0.046	25.21	0.76	0.042
	Cut [26]	23.51	0.71	0.050	22.44	0.71	0.053
	Swin-UNIT [20]	23.92	0.72	0.057	23.77	0.75	0.051
	Chai *et al.* [2]	29.35	0.86	0.028	27.78	0.86	0.037
	UVCGAN [30]	27.22	0.82	0.031	26.10	0.79	0.037
	Xie *et al.* [33]	<u>29.73</u>	<u>0.90</u>	0.025	28.09	**0.89**	0.037
	Rawformer [27]*	29.32	<u>0.90</u>	<u>0.023</u>	<u>28.45</u>	<u>0.88</u>	<u>0.034</u>
	MERIT (Ours)	**35.29**	**0.91**	**0.015**	**31.90**	<u>0.85</u>	**0.021**

* We retrain the model under the same settings as our MERIT and other baselines, since the results reported in the original paper were trained on the images demosaiced using the Menon algorithm [22].

Table 2. **Quantitative results on the RAW-to-RAW mapping dataset.** Best results are **bolded** and second-best are <u>underlined</u>.

Tab. 2 presents a comprehensive quantitative comparison across various training paradigms. MERIT consistently outperforms all baselines, achieving state-of-the-art (SOTA) results under the unsupervised setting. Traditional non-learning approaches perform poorly, especially in the direction iPhone→Samsung, with MAE values exceeding 0.16. This demonstrates the inherent limitations of linear mappings in modeling complex inter-sensor differences. Afifi *et al*. [1] performs competitively but requires paired supervision. In contrast, MERIT achieves outstanding results without any paired supervision, validating the strength of our unsupervised multi-domain training strategy. Among unsupervised baselines, prior GAN-based and contrastive methods yield moderate performance but lag behind recent domain-specific designs, such as Rawformer [27] and Xie *et al*. [33]. Even so, **MERIT consistently outperforms these tailored approaches in 5 out of 6 evaluation metrics across both translation directions.** MERIT achieves a PSNR gain of $+5.56$ dB and a MAE reduction of 0.008 on Samsung→iPhone, while also improving iPhone→Samsung by $+3.45$ dB in PSNR and reducing MAE by 0.013. These results emphasize the importance of our innovations in effectively bridging the spectral diversity of real-world cameras, confirming the effectiveness and robustness of MERIT in producing high-quality and semantically consistent RAW image translations.

Tab. 3 presents a comprehensive evaluation of our MERIT against two strong baselines, UVCGAN [30] and Xie *et al*. [33], across all pairwise source-to-target camera domain combinations in our proposed MDRAW dataset. Each cell summarizes the performance for a specific domain pair using four metrics: MAE, PSNR, SSIM, and KL divergence. Across the 20 non-diagonal translation pairs,

Source \ Target	Samsung	Huawei	iPhone	Nikon	Canon
Samsung	–	0.026 / 30.77 / 0.77 / 1.53 0.026 / 31.15 / **0.78** / 1.46 **0.025 / 31.27 / 0.78 / 1.24**	0.036 / **29.11** / 0.75 / **1.64** 0.036 / 28.88 / 0.76 / 1.89 **0.035 / 29.11 / 0.77** / 1.73	0.038 / 27.99 / 0.72 / 2.36 0.037 / 28.21 / 0.72 / 2.28 **0.034 / 28.71 / 0.73 / 2.06**	0.047 / 26.71 / 0.72 / 3.18 0.045 / 27.16 / 0.73 / 2.77 **0.043 / 27.59 / 0.75 / 2.65**
Huawei	**0.028 / 30.26** / 0.74 / 1.73 0.032 / 29.15 / 0.73 / 1.65 0.029 / **30.26 / 0.76 / 1.51**	–	**0.033** / 29.20 / 0.77 / 2.38 0.035 / 28.48 / 0.76 / **2.03** **0.033 / 29.23 / 0.79** / 2.30	0.033 / 29.43 / **0.75** / 2.05 **0.032 / 29.52** / 0.74 / 1.93 **0.032** / 29.39 / **0.75 / 1.89**	0.045 / 26.98 / 0.72 / 3.63 0.043 / 27.46 / 0.73 / **2.59** **0.042 / 27.60 / 0.75** / 3.07
iPhone	0.038 / 28.71 / 0.73 / 1.55 0.035 / 29.07 / 0.72 / 1.67 **0.034 / 29.12 / 0.74 / 1.42**	0.033 / 29.07 / 0.74 / 2.17 0.030 / **30.19 / 0.76** / 1.91 **0.029** / 29.77 / **0.76 / 1.87**	–	0.044 / 26.67 / 0.69 / 2.79 **0.039 / 27.68** / 0.71 / 2.58 0.038 / **27.68 / 0.72 / 2.50**	0.046 / 26.01 / 0.70 / 2.85 **0.038 / 28.17** / 0.73 / **2.30** 0.041 / 27.09 / **0.74** / 2.47
Nikon	**0.039 / 27.64** / 0.69 / 2.65 0.041 / 26.79 / 0.67 / 2.69 **0.039** / 27.37 / **0.70 / 2.60**	0.034 / 28.58 / 0.70 / 2.54 **0.033** / 28.83 / **0.74** / 2.65 **0.033 / 28.92 / 0.74 / 2.47**	0.044 / 26.54 / 0.71 / 3.15 **0.043** / 26.32 / 0.71 / **2.65** **0.043 / 26.61 / 0.73** / 3.16	–	0.033 / 29.31 / 0.76 / **2.12** 0.033 / 29.38 / 0.77 / 2.20 **0.032 / 29.51 / 0.78** / 2.32
Canon	0.047 / **26.61 / 0.70** / 3.06 0.049 / 25.78 / 0.66 / 2.82 **0.045** / 26.49 / **0.70 / 2.79**	0.044 / 26.88 / 0.71 / **3.20** 0.042 / 27.08 / 0.70 / 3.28 **0.041 / 27.15 / 0.72** / 3.24	0.044 / 26.40 / **0.73** / 2.94 0.043 / 26.07 / 0.71 / **2.31** **0.042 / 26.42 / 0.73** / 2.63	0.033 / 29.13 / 0.75 / 2.11 0.032 / **29.31** / 0.75 / **1.99** **0.031** / 29.27 / **0.76** / 2.27	–

Table 3. **Cross-domain RAW2RAW translation results on MDRAW.** Each cell reports results for translation from a source domain (column) to a target domain (row). Within each cell, three lines correspond to different methods: UVCGAN [30], Xie *et al.* [33] , and MERIT (Ours) . Each line contains four metrics in the order of **MAE**(↓)/ **PSNR**(↑)/ **SSIM** (↑)/ **KL Divergence**(↓).

Setting	Source→Target	MAE↓	PSNR↑	SSIM↑
Baseline				
	iPhone-X → Samsung-S9	0.0256	30.99	0.8357
	Samsung-S9 → iPhone-X	0.0186	33.75	0.8841
	Avg all	0.0221	32.37	0.8599
+ SANM				
	iPhone-X → Samsung-S9	0.0250	31.24	0.8426
	Samsung-S9 → iPhone-X	0.0169	34.37	0.8966
	Avg all	0.0210	32.71	0.8682
+ MS-LKA + SANM				
	iPhone-X → Samsung-S9	0.0226	31.74	0.8482
	Samsung-S9 → iPhone-X	0.0164	34.66	0.9006
	Avg all	0.0195	33.20	0.8744
+ MS-LKA + SANM + $\mathcal{L}_{cycle-SSIM}$				
	iPhone-X → Samsung-S9	0.0219	31.90	0.8508
	Samsung-S9 → iPhone-X	0.0151	35.29	0.9104
	Avg all	**0.0185**	**33.60**	**0.8806**

Table 4. **Ablation study on RAW-to-RAW mapping dataset.**

MERIT consistently achieves the best or second-best performance in nearly every metric and direction. Specifically, MERIT records the **lowest MAE in 17 out of 20 cases**, the **highest PSNR in 14 out of 20**, the **highest SSIM in 20 out of 20**, and the lowest KL divergence in 11 out of 20, demonstrating superior accuracy and perceptual quality in modeling inter-domain RAW distributions. These results underscore the robustness and generalizability of MERIT in diverse cross-camera scenarios, outperforming prior RAW translation models.

3.2. Ablation Study

We conduct a detailed ablation study on the RAW-to-RAW mapping dataset [1] to assess the individual and cumulative impact of key components in MERIT. As shown in Tab. 4, our baseline model achieves reasonable performance, with an average MAE of 0.0221, PSNR of 32.37 dB, and SSIM of 0.8599 across both translation directions. Notably, performance is asymmetric: Samsung-S9→iPhone-X translation performs better across all metrics, indicating domain-specific variation in complexity. Adding SANM leads to a substantial performance boost, particularly in the Samsung-S9→iPhone-X direction. MAE drops to 0.0210, while PSNR improves to 32.71 dB, and SSIM rises to 0.8682. This suggests that explicit modeling of sensor-specific noise profiles contributes significantly to image fidelity and structural consistency. Incorporating MS-LKA alongside SANM further improves the overall translation quality. The average PSNR reaches 33.20 dB, and the MAE is reduced to 0.0195. These gains reflect MS-LKA's ability to capture long-range dependencies and contextual structures, which are particularly beneficial in sensor-specific color or tone transitions that occur in challenging RAW domains. Finally, adding an explicit consistency SSIM loss term yields the best overall performance. The average MAE drops to 0.0185, PSNR improves to 33.60 dB, and SSIM reaches 0.8806. This confirms that directly optimizing for perceptual similarity complements the pixel-level losses.

Tab. 5 presents a comparative study of MERIT against recent SOTA models, as the number of camera domains increases. Across all domain configurations, MERIT achieves superior accuracy with significantly fewer parameters and lower training cost than UVCGAN, and matches or exceeds the performance of Xie *et al*. despite requiring $2-5\times$ fewer training iterations. In the 3-domain case, MERIT achieves the best performance with only 58.7M parameters, 1/3 the size of UVCGAN. Notably, while Xie *et al*. has the smallest parameter count in this setting, it requires

longer training (266K iterations), and still underperforms MERIT in all metrics. As the number of domains increases, MERIT exhibits remarkable scalability. When moving from 3 to 5 domains, UVCGAN's parameter count scales linearly (186M→620M) and training time grows proportionally. In contrast, MERIT maintains a nearly constant model size (∼58.7M) and training iteration budget (180K), demonstrating strong efficiency and parameter-sharing capability. At 5 domains, MERIT still outperforms the baselines in PSNR, SSIM, and MAE, while achieving the second-best KL divergence. These results validate MERIT's effectiveness in maintaining high quality, while being significantly more scalable and training-efficient than prior models.

More results are discussed in the supplementary material.

4. Related Work

4.1. RAW-to-RAW Translation

The objective of the RAW2RAW translation is to establish a mapping function f capable of accurately translating RAW images from Camera A's RAW space to Camera B's RAW space, accommodating diverse scenes and lighting conditions. In brief, the mapping function can be expressed as $I^B = f(I^A)$, where I^A and I^B denote the packed RAW images $[r, g_r, g_b, g]$ captured by camera A and camera B, respectively. Work [18, 24] proposed that by employing a quadratic transformation, the mapping function can be approximated by $I^B \approx I^A_{qt} T_{qt}$, where $I^A_{qt} = [r^2, g_r^2, \cdots, r \times g_r, g_r \times b, \cdots, r, g_r, g_b, g]$, and $T_q t \in \mathbb{R}^{14\times4}$ is a transformation matrix. Work [1] was the first to introduce neural networks into the RAW task; they placed a standard color chart in the scene and captured RAW images using two cameras. Subsequently, they estimated T_{qt} by minimizing the color differences between the two color charts. They applied this transformation to the RAW images and obtained pixel-level paired data. They then utilized a neural network for semi-supervised training. Although the trained model can accomplish RAW2RAW translation without having to capture paired images again, creating a paired dataset for training is still inevitable. The following work [27, 33] released the requirement of using paired images for training. Rawformer [27] proposed a transformer-based encoder-decoder model to implement fully unsupervised learning. They introduced contextual-scale aware downsampler and upsampler blocks that efficiently summarize the local-global contextual details in mixed scale representations. Concurrent work [33] proposed a color space predictor to predict the space transformation parameters in a patch-wise manner, which accurately performs transformation and flexibly manages complex lighting conditions. Work [25] leveraged the spline capabilities of Kolmogorov-Arnold Networks [21] to model the color matching between source and target distributions. They developed a hypernetwork that generates spatially varying weights to control the nonlinear splines of a KAN, enabling accurate matching.

4.2. Unpaired Image Translation

Unpaired image translation aims to map images from the domain I^A to I^B without ground truth. CycleGAN [38] utilized a cycle-consistency loss and identity loss to bridge domain gaps. UVCGAN [30] introduced a pixel-wise transformer into the CycleGAN framework, while Swin-UNIT [20] addressed performance issues caused by high-resolution images by introducing swin-transformer block. StarGAN series [5, 6] extended the task from two-domain transformation to multi-domain. StarGAN [5] proposed a novel GAN that learns the mappings among multiple domains using only a single generator and a discriminator by adding a mask vector to the domain label. StarGANv2 [6] introduced a style encoder. The style encoder learns to extract the style code from a reference image from the target domain. Given the style codes, the generator learns to synthesize images over multiple domains.

5. Conclusion

We presented MERIT, a novel and unified framework for multi-domain RAW image translation, capable of performing one-to-many and many-to-many mappings across diverse camera sensor domains using a single model. By explicitly modeling signal-dependent noise and enhancing global context perception, MERIT significantly improves the fidelity and generalizability of RAW2RAW translations. To support standardized benchmarking, we also introduced MDRAW, the first dataset specifically curated for multi-domain RAW translation, including both paired and unpaired RAW captures from five diverse camera sensors under various real-world conditions. Extensive experiments on existing datasets and MDRAW validate the advantages of MERIT. MERIT achieves superior performance in terms of quality and scalability while maintaining a compact parameter footprint.

Acknowledgement

This work was supported in part by the DARPA Young Faculty Award, the National Science Foundation (NSF) under Grants #2431561, #2127780, #2319198, #2321840, #2312517, and #2235472, the Semiconductor Research Corporation (SRC), the Office of Naval Research through the Young Investigator Program Award and Grants #N00014-21-1-2225 and #N00014-22-1-2067, Army Research Office Grant #W911NF2410360, and DARPA under Support Agreement No. USMA 23004. Additionally, support was provided by the Air Force Office of Scientific Research under Award #FA9550-22-1-0253, along with generous gifts from Xilinx and Cisco.

References

[1] Mahmoud Afifi and Abdullah Abuolaim. Semi-supervised raw-to-raw mapping. *arXiv preprint arXiv:2106.13883*, 2021. 2, 5, 6, 7, 8, 11

[2] Yoav Chai, Raja Giryes, and Lior Wolf. Supervised and unsupervised learning of parameterized color enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 992–1000, 2020. 6

[3] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[4] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023. 1

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 8

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 8

[7] Marcos V. Conde, Florin Vasluianu, and Radu Timofte. Bsraw: Improving blind raw image super-resolution. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8485–8495, 2024. 1

[8] Ziteng Cui and Tatsuya Harada. Raw-adapter: Adapting pre-trained visual model to camera raw images. In *European Conference on Computer Vision*, pages 37–56. Springer, 2024. 1

[9] Ziteng Cui, Xuangeng Chu, and Tatsuya Harada. Luminance-gs: Adapting 3d gaussian splatting to challenging lighting conditions with view-adaptive curve adjustment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26472–26482, 2025. 1

[10] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 12

[11] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE transactions on image processing*, 17(10):1737–1754, 2008. 2, 3

[12] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational visual media*, 9(4):733–752, 2023. 4

[13] Wenjun Huang, Ziteng Cui, Yinqiang Zheng, Yirui He, Tatsuya Harada, and Mohsen Imani. Dr. raw: Towards general high-level vision from raw with efficient task conditioning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 1

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[15] Hai Jiang, Binhao Guan, Zhen Liu, Xiaohong Liu, Jian Yu, Zheng Liu, Songchen Han, and Shuaicheng Liu. Learning to see in the extremely dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7676–7685, 2025. 1

[16] Xin Jin, Pengyi Jiao, Zheng-Peng Duan, Xingchao Yang, Chong-Yi Li, Chun-Le Guo, and Bo Ren. Lighting every darkness with 3dgs: Fast training and real-time rendering for hdr view synthesis. In *NIPS*, 2024. 1

[17] Eric Kee, Adam Pikielny, Kevin Blackburn-Matzen, and Marc Levoy. Removing reflections from raw photos. *CVPR*, 2025. 1

[18] Seon Joo Kim, Hai Ting Lin, Zheng Lu, Sabine Süsstrunk, Stephen Lin, and Michael S Brown. A new in-camera imaging model for color computer vision and its application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2289–2302, 2012. 8

[19] Sohyun Lee, Jaesung Rim, Boseung Jeong, Geonu Kim, Byungju Woo, Haechan Lee, Sunghyun Cho, and Suha Kwak. Human pose estimation in extremely low-light conditions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 704–714, 2023. 1

[20] Yifan Li, Yaochen Li, Wenneng Tang, Zhifeng Zhu, Jinhuo Yang, and Yuehu Liu. Swin-unit: Transformer-based gan for high-resolution unpaired image translation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4657–4665, 2023. 6, 8

[21] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024. 8

[22] Daniele Menon, Stefano Andriani, and Giancarlo Calvagno. Demosaicing with directional filtering and a posteriori decision. *IEEE Transactions on Image Processing*, 16(1):132–141, 2006. 6

[23] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. NeRF in the dark: High dynamic range view synthesis from noisy raw images. *CVPR*, 2022. 1

[24] Rang Nguyen, Dilip K Prasad, and Michael S Brown. Raw-to-raw: Mapping between image sensor color responses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3398–3405, 2014. 2, 6, 8

[25] Artem Nikonorov, Georgy Perevozchikov, Andrei Korepanov, Nancy Mehta, Mahmoud Afifi, Egor Ershov, and Radu Timofte. Color matching using hypernetwork-based kolmogorov-arnold networks. *arXiv preprint arXiv:2503.11781*, 2025. 2, 8

[26] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020. 6

[27] Georgy Perevozchikov, Nancy Mehta, Mahmoud Afifi, and Radu Timofte. Rawformer: Unpaired raw-to-raw translation

for learnable camera isps. In *European Conference on Computer Vision*, pages 231–248. Springer, 2024. 2, 6, 8

[28] Iskra Petrova. Google working on bringing raw photos to more third-party android camera apps. https://www.phonearena.com/news/google-working-on-raw-photos-more-third-party-android-camera-apps_id164651, 2024. Accessed: 2025-11-13. 2

[29] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 5, 11

[30] Dmitrii Torbunov, Yi Huang, Haiwang Yu, Jin Huang, Shinjae Yoo, Meifeng Lin, Brett Viren, and Yihui Ren. Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 702–712, 2023. 6, 7, 8, 13

[31] Yujin Wang, Tianyi Xu, Fan Zhang, Tianfan Xue, and Jinwei Gu. Adaptiveisp: learning an adaptive image signal processor for object detection. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 1

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[33] Dongyu Xie, Chaofan Qiao, Lanyue Liang, Zhiwen Wang, Tianyu Li, Qiao Liu, Chongyi Li, Guoqing Wang, and Yang Yang. Generalizing isp model by unsupervised raw-to-raw mapping. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3809–3817, 2024. 2, 6, 7, 8, 13

[34] Xiangyu Xu, Yongrui Ma, and Wenxiu Sun. Towards real scene super-resolution with raw images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1723–1731, 2019. 1

[35] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095, 2020. 6

[36] Andy Zahn. Samsung expert raw vs. apple proraw. https://www.digitaltrends.com/phones/expert-raw-vs-apple-proraw/, 2022. Accessed: 2025-11-13. 2

[37] Yi Zhang, Hongwei Qin, Xiaogang Wang, and Hongsheng Li. Rethinking noise synthesis and modeling in raw denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4593–4601, 2021. 3

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 6, 8