RawGen: Learning Camera Raw Image Generation

Dongyoung Kim[1], Junyong Lee[1*], Abhijith Punnappurath[1*], Mahmoud Afifi[1*], Sangmin Han[2], Alex Levinshtein[1], and Michael S. Brown[1]

[1]AI Center - Toronto, Samsung Electronics [2]Yonsei University [*]Equal contribution

Project Page: `https://dy112.github.io/rawgen-page/`

Fig. 1: We present RawGen, a diffusion-based method for generating realistic camera raw images. RawGen produces a latent representation of linear CIE XYZ images, conditioned on an sRGB image or a text prompt. The latent is decoded to CIE XYZ and mapped to arbitrary target camera raw spaces.

Abstract. Cameras capture scene-referred linear raw images, which are processed by onboard image signal processors (ISPs) into display-referred 8-bit sRGB outputs. Although raw data is more faithful for low-level vision tasks, collecting large-scale raw datasets remains a major bottleneck, as existing datasets are limited and tied to specific camera hardware. Generative models offer a promising way to address this scarcity; however, existing diffusion frameworks are designed to synthesize photo-finished sRGB images rather than physically consistent linear representations. This paper presents RawGen, to our knowledge the first diffusion-based framework enabling text-to-raw generation for arbitrary target cameras, alongside sRGB-to-raw inversion. RawGen leverages the generative priors of large-scale sRGB diffusion models to synthesize physically meaningful linear outputs, such as CIE XYZ or camera-specific raw representations, via specialized processing in latent and pixel spaces. To handle unknown and diverse ISP pipelines and photo-finishing effects in diffusion-model training data, we build a many-to-one inverse-ISP dataset where multiple sRGB renditions of the same scene generated using diverse ISP parameters are anchored to a common scene-referred target. Fine-tuning a conditional denoiser and specialized decoder on this dataset allows RawGen to obtain camera-centric linear reconstructions that effectively invert the rendering pipeline. We demonstrate RawGen's superior performance over traditional inverse-ISP methods that assume a fixed ISP. Furthermore, we show that augmenting training pipelines with RawGen's scalable, text-driven synthetic data can benefit downstream low-level vision tasks.

1 Introduction

Camera sensors natively record scene-referred linear raw data. Raw images are subsequently processed by onboard image signal processors (ISPs) through a lossy rendering pipeline to produce display-referred outputs, typically in the 8-bit sRGB format. While raw captures preserve physically meaningful radiometric information, they are rarely available at scale and often tied to specific camera models. Collecting large raw datasets is costly and labor-intensive, and data must be re-captured whenever the sensor or the camera pipeline changes, limiting reproducibility and cross-device coverage. This scarcity remains a fundamental bottleneck for low-level vision and computational photography research that relies on linear, sensor-centric measurements.

Generative models offer a potential route to mitigate this limitation. Recent diffusion-based frameworks have achieved remarkable success in producing high-fidelity, semantically coherent sRGB images [20, 24, 25, 31, 40, 42, 43, 55–58]. In particular, text-to-image (T2I) systems [5, 8, 16, 21, 44, 46, 47, 50, 52, 65, 67] enable strong semantic controllability via textual prompts, facilitating scalable and diverse image synthesis. However, these models operate primarily in the 8-bit sRGB domain, which is optimized for display and storage rather than preserving physical linearity. This representation typically includes nonlinear photo-finishing operations (e.g., gamma correction, tone mapping, and color styling) that distort the underlying radiometric structure of the scene. Consequently, diffusion model outputs are not directly compatible with scene-referred linear signals, such as camera raw measurements or device-independent linear representations that can serve as canonical references.

A natural strategy is to reconstruct linear radiometric data from diffusion-generated sRGB images using inverse-ISP techniques [2, 7, 48, 62]. However, most existing inverse-ISP methods are trained on paired datasets generated under a fixed imaging pipeline, such as software ISPs (e.g., RawPy, DCRAW) or a specific in-camera ISP configuration. These methods therefore implicitly assume a single camera response and a known sequence of color and tone transformations. In contrast, diffusion models are trained on large-scale sRGB corpora that encompass diverse and heterogeneous photo-finishing styles originating from unknown ISPs and post-processing pipelines.

This discrepancy creates a fundamental domain gap. Conventional inverse-ISP models expect sRGB inputs produced by a fixed and known rendering pipeline, whereas diffusion-generated sRGB images entangle unknown, diverse, and often highly nonlinear photo-finishing effects. Directly applying standard inverse-ISP methods to diffusion model outputs fails to reliably recover a consistent scene-referred linear representation. Enabling physically meaningful linear signal synthesis from generative models therefore remains an open challenge.

In this paper, we introduce RawGen, a diffusion-based framework designed to bridge this gap. RawGen preserves the strong semantic priors and promptability of large-scale diffusion models while incorporating learned latent-level and pixel-level unprocessing to recover a canonical scene-referred linear representation from diverse sRGB inputs. Rather than assuming a fixed inverse mapping,

RawGen is trained to suppress uncontrolled and heterogeneous photo-finishing effects implicitly embedded in diffusion-generated images, thereby enabling consistent and physically grounded linear synthesis.

The central idea of RawGen is to factor out ISP-induced variability and recover a shared linear anchor that remains consistent across multiple sRGB versions of the same underlying scene. To achieve this, we adopt a many-to-one reconstruction objective in which multiple photo-finished sRGB images of a single scene serve as inputs and are mapped to a common linear reference. This formulation explicitly encourages the recovered linear representation to remain invariant to differences in post-processing operations—such as tone mapping, gamma correction, and color styling—that are embedded in the sRGB inputs, while preserving the underlying scene structure and semantic content.

With this learned invariance, RawGen supports two complementary modes of generation. First, it converts an sRGB image into a canonical scene-referred linear representation that is robust to diverse post-processing styles. Second, it maps nonlinear image representations produced from text prompts by a pre-trained diffusion prior into canonical linear representations. When a target camera raw space is required, the canonical linear output can be transformed into the camera-specific raw domain using standard color-space mappings and camera metadata, enabling camera-agnostic raw synthesis without retraining (Fig. 1).

We demonstrate that RawGen consistently outperforms conventional inverse-ISP methods that rely on fixed imaging assumptions, yielding a more stable and scene-referred linear representation under diverse and uncontrolled photo-finishing variations. Moreover, RawGen enables scalable text-driven raw generation as a practical source of training data for downstream tasks such as illuminant estimation, neural ISP learning, and denoising, providing broad scene diversity without additional capture effort or scene asset design.

Contribution

In summary, our main contributions are as follows:

- We propose RawGen, a diffusion-based camera-agnostic raw generation framework that preserves strong generative priors while unprocessing diverse sRGB inputs with unknown photo-finishing into a canonical scene-referred linear domain via latent- and pixel-level processing.
- We introduce a many-to-one linear reconstruction task and propose a method to construct a new type of dataset that maps multiple photo-finished sRGB observations to a single common linear reference, enabling robustness to heterogeneous and unknown ISP transformations.
- We develop, to our knowledge, the first unified framework enabling text-driven synthesis of physically meaningful linear and camera-specific raw data for arbitrary target cameras, and demonstrate that the resulting synthetic data alleviates data acquisition challenges in downstream low-level vision tasks.

2 Related Works

Forward & Inverse Camera ISP Pipelines. Traditional camera hardware ISPs consist of carefully engineered signal-processing stages [19], such as denoising, demosaicing, white balancing, color correction, tone mapping, and more, often tuned to produce a manufacturer-specific visual aesthetic. Because in-camera pipelines are proprietary and largely inaccessible, the rendered sRGB photographs available on the internet reflect diverse and unknown photo-finishing operations. Large-scale generative models trained on such data therefore inherit these uncontrolled rendering characteristics, producing outputs in the nonlinear sRGB domain with implicit and heterogeneous ISP effects.

Recent work has explored learning-based ISPs that replace handcrafted modules with end-to-end neural mappings (e.g., [23, 27, 28, 45, 66]). These networks are trained to convert raw images, typically from a specific camera, to their display-referred outputs assuming a one-to-one relationship between the sRGB images and their originating ISPs.

Recovering linear sensor data from rendered images has long been studied for low-level vision. Early approaches to raw reconstruction relied on calibration-based techniques [11, 12, 18, 22, 33, 38]. Later work, such as UPI [9], performed sRGB-to-raw reconstruction by sequentially inverting the ISP with non-learnable parametric operations. Nam et al. [39] proposed one of the earliest deep learning frameworks to jointly model the forward and inverse ISP mappings. Building on this idea, subsequent methods such as CIE-XYZ-Net [2], CycleISP [64], InvISP [62], ParamISP [34], and model-based ISP approaches [14] further advanced the field through convolutional networks and differentiable, model-driven architectures trained on paired raw-sRGB data.

Despite their differences, existing data-driven inverse ISP approaches largely assume a one-to-one relationship between rendered images and their underlying ISPs. This assumption is reasonable when training data is captured or synthesized under controlled pipelines but breaks down for diffusion-generated imagery, where rendering parameters are unknown and highly variable. Consequently, applying conventional inverse-ISP strategies to generative outputs cannot reliably recover a consistent scene-referred representation.

Diffusion Models for Low-Level Vision Tasks. Denoising diffusion models [24,50,55] are powerful generative models that synthesize images by reversing a gradual noising process. Their state-of-the-art performance has been demonstrated across diverse low-level vision tasks [35,51], including super-resolution [53], image restoration [32, 60], and low-light enhancement [29, 61]. Most methods approach these tasks as conditional generation problems, guiding the diffusion process with auxiliary inputs like degraded images or semantic maps.

Recently, diffusion-based approaches have been explored for ISP-related tasks. Some works perform inverse-ISP by reconstructing raw from sRGB inputs [48], while others learn forward ISP mappings (raw-to-sRGB) [13,49]. These methods, however, typically rely on synthetic training pairs generated via fixed one-to-one ISP logic (e.g., using RawPy). More recently, Yuan et al. [63] proposed controlling camera parameters (e.g., white balance, exposure) by guiding the diffusion

process to maintain scene consistency. While effective for guided synthesis, this approach embeds parameter adjustment within the iterative generation, potentially requiring a new generative pass for each modification. For HDR reconstruction, diffusion-based works [6, 59] focus on fusing multiple exposures, generally outputting tone-mapped sRGB or display-oriented HDR formats.

Consequently, these approaches do not address the challenge of reconstructing camera-centric, physically linear representations (e.g., CIE XYZ or raw) from in-the-wild sRGB inputs. Furthermore, they do not explore text-driven synthesis of raw data. Such an approach that generates raw data would decouple the initial generation from subsequent adjustments, thereby enabling flexible and efficient post-processing in standard dedicated software.

To handle unknown ISP pipelines, we adopt a many-to-one training paradigm that anchors diverse photo-finished observations to a shared linear target, enabling physically consistent raw generation at scale. RawGen is the first framework capable of synthesizing physically meaningful raw images directly from open-world text prompts without assuming a fixed imaging pipeline.

3 Method

RawGen synthesizes physically meaningful linear data by repurposing pretrained sRGB diffusion models. As discussed in Sec. 2, the same raw capture can yield substantially different sRGB renderings under different ISP and photo-finishing settings, and large-scale diffusion training corpora entangle these unknown, heterogeneous effects. Rather than assuming a fixed inverse mapping as in conventional inverse-ISP methods, RawGen learns to suppress this uncontrolled variability and recover a canonical, scene-referred linear representation.

We adopt CIE XYZ as the canonical space. It is linear and device-independent, and it relates to camera-specific raw spaces through standard color transforms, enabling camera-agnostic synthesis without retraining.

As shown in Fig. 2, our method builds on a pretrained rectified-flow DiT [42] with native image conditioning capability. We first describe the many-to-one training data construction (Sec. 3.1) that underlies our training stages, then present the two fine-tuning stages: denoiser fine-tuning (Sec. 3.2) and decoder fine-tuning (Sec. 3.3), corresponding to panels (A) and (B) of Fig. 2, and finally detail inference for image-to-raw and text-to-raw generation (Sec. 3.4), corresponding to Fig. 2(C).

3.1 Many-to-One Training Data Construction

To invert diverse, unknown photo-finishing effects into a single linear target, we first construct paired data that links multiple sRGB renditions of the same scene to one common anchor. Starting from a raw image I_{raw}, we derive a linear CIE XYZ image I_{XYZ} by applying per-scene white balancing and a camera-to-XYZ color conversion matrix, yielding a scene-referred representation that

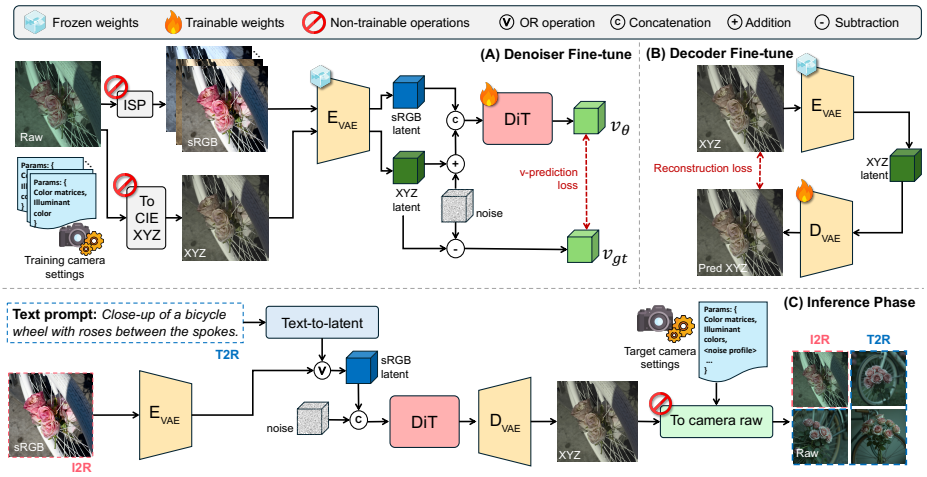

Fig. 2: Overview of the RawGen framework. During training, raw images are converted to CIE XYZ and sRGB representations to (A) fine-tune the DiT to denoise CIE XYZ latents conditioned on an sRGB image, and (B) fine-tune the VAE decoder to reconstruct CIE XYZ images. During inference (C), either an sRGB image (image-to-raw, I2R) or a text prompt (text-to-raw, T2R) is used to condition the DiT to generate a CIE XYZ latent, which is decoded to obtain a CIE XYZ image and subsequently mapped to the target camera's raw space using its calibration metadata, which can be easily acquired from a single DNG file of target camera.

precedes any photo-finishing. We treat I_{XYZ} as the *anchor*, shared across all sRGB variants of the same scene.

Using this anchor, we generate N sRGB variants $\{I^{(n)}_{\mathrm{sRGB}}\}_{n=1}^{N}$ by sampling diverse photo-finishing parameter vectors $P^{(n)}$, which control global color tone, contrast, and tone mapping. Concretely, we randomize three ISP parameter groups: (i) white balance, by perturbing red and blue gains around the DNG `AsShotNeutral` initialization; (ii) tone mapping, by varying the per-channel tone-curve shape; and (iii) contrast, via a global gain about a mid-gray pivot in sRGB. Together, these perturbations yield diverse photo-finishing styles while preserving the same scene content. By pairing multiple sRGB variants with a single XYZ anchor, this strategy approximates the heterogeneous ISP behavior found in real-world sRGB images, encouraging invariance to photo-finishing while preserving scene content. Each image is encoded by the frozen VAE encoder E_{VAE}:

$$z^{(n)}_{\mathrm{sRGB}} = E_{\mathrm{VAE}}\left(I^{(n)}_{\mathrm{sRGB}}\right), \quad z_{\mathrm{XYZ}} = E_{\mathrm{VAE}}(I_{\mathrm{XYZ}})\,. \tag{1}$$

The resulting latent pairs $\{(z^{(n)}_{\mathrm{sRGB}}, z_{\mathrm{XYZ}})\}_{n=1}^{N}$ are cached for both training stages. More details of the ISP pipeline and photo-finishing parameter sampling are provided in the supplementary material.

3.2 Denoiser Fine-Tuning

The first training stage, shown in Fig. 2(A), fine-tunes the pretrained DiT to map sRGB-conditioned inputs to the linear XYZ anchor latent.

Objective. Given the anchor latent z_{XYZ} from Eq. (1), we randomly sample one sRGB variant latent $z^{(n)}_{\mathrm{sRGB}}$ under the many-to-one setting. We corrupt the anchor latent by interpolating with Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$:

$$z_t = (1-t)\, z_{\mathrm{XYZ}} + t\, \epsilon, \quad t \in [0,1], \tag{2}$$

and tune the DiT parameterized by θ to predict the rectified-flow velocity target (also referred to as a v-prediction target) $v_{\mathrm{gt}} = \epsilon - z_{\mathrm{XYZ}}$:

$$\mathcal{L}_{\mathrm{denoise}} = \mathbb{E}_{n,\,t,\,\epsilon} \left\| v_{\mathrm{gt}} - v_\theta\Big(z_t,\, t;\, z^{(n)}_{\mathrm{sRGB}}\Big) \right\|_2^2. \tag{3}$$

By tuning with randomly sampled variants, the denoiser learns to suppress ISP-induced variability in the sRGB condition and consistently recover the shared latent representation of the linear anchor XYZ image.

Conditioning. We leverage the image conditioning mechanism of the pretrained DiT. The sRGB context latent $z^{(n)}_{\mathrm{sRGB}}$ and the noisy target latent z_t are patchified into token sequences and concatenated along the sequence dimension, and the combined sequence is processed jointly through the DiT transformer blocks. Only the target portion of the output contributes to the loss. We fine-tune the model using LoRA adapters [26] on attention projection layers, while keeping the pretrained backbone weights frozen.

3.3 Decoder Fine-Tuning

The second training stage, shown in Fig. 2(B), fine-tunes the VAE decoder to reconstruct linear XYZ images from XYZ-domain latents. Since the pretrained decoder is optimized for sRGB, directly decoding XYZ latents may yield degraded outputs. We fine-tune D_{VAE} using the ground-truth anchor latent z_{XYZ} from Eq. (1), minimizing an ℓ_1 loss:

$$\widehat{I}_{\mathrm{XYZ}} = D_{\mathrm{VAE}}(z_{\mathrm{XYZ}}),\ \mathcal{L}_{\mathrm{recon}} = \left\| \widehat{I}_{\mathrm{XYZ}} - I_{\mathrm{XYZ}} \right\|_1. \tag{4}$$

This retargets the VAE decoder from sRGB to linear XYZ while preserving the spatial representations learned during pretraining.

3.4 Inference: Image-to-Raw and Text-to-Raw

As illustrated in Fig. 2(C), RawGen uses a unified inference pipeline that first synthesizes a scene-referred CIE XYZ image and then renders it into a target camera's linear raw space. The only difference between image-to-raw (I2R) and text-to-raw (T2R) lies in how the sRGB conditioning latent is obtained

within the same pretrained backbone that shares the core DiT and the VAE encoder/decoder. Specifically, I2R encodes an input sRGB image using the frozen VAE encoder, whereas T2R follows the conventional diffusion model's standard text-conditioned generation route and directly takes the intermediate latent produced from the text prompt prior to VAE decoding (i.e., the latent that would otherwise be decoded to an sRGB image). In both cases, this sRGB latent conditions the same DiT to generate an XYZ latent, which is decoded with our fine-tuned VAE decoder and finally mapped deterministically to the target camera's raw space using its calibration metadata under a chosen illuminant. In our implementation, we instantiate the backbone with FLUX.1 and reuse its native text-to-latent pathway for T2R conditioning.

Image-to-Raw (I2R). Given an sRGB image I_{sRGB}, we obtain the conditioning latent $z_{\mathrm{sRGB}} = E_{\mathrm{VAE}}(I_{\mathrm{sRGB}})$. We initialize the target tokens with noise and solve the reverse ordinary differential equation (ODE) conditioned on z_{sRGB}:

$$\hat{z}_{\mathrm{XYZ}} = z_1 + \int_1^0 v_\theta(z_t,\, t;\, z_{\mathrm{sRGB}})\, \mathrm{d}t, \tag{5}$$

approximated using Euler steps. We retain only the target tokens, reshape them into the spatial latent $\hat{z}_{\mathrm{XYZ}}$, and decode $\widehat{I}_{\mathrm{XYZ}} = D_{\mathrm{VAE}}(\hat{z}_{\mathrm{XYZ}})$. The decoded XYZ image is then converted to the target camera raw space using the mapping described below.

Text-to-Raw (T2R). Given a text prompt, we first obtain an sRGB conditioning latent z_{sRGB} using the pretrained text-to-latent pathway of the base model. We then run the same conditional generation and decoding steps as in I2R to produce $\hat{z}_{\mathrm{XYZ}}$ and $\widehat{I}_{\mathrm{XYZ}}$, and apply the same XYZ-to-camera mapping to obtain camera-specific linear raw outputs.

CIE XYZ to Camera Raw. To convert $\widehat{I}_{\mathrm{XYZ}}$ into a target camera's linear raw-RGB space, we apply calibration transforms derived from the target camera's DNG metadata together with a chosen illuminant parameterized by correlated color temperature (CCT). In practice, we (i) interpolate camera calibration matrices based on the selected CCT, (ii) map $\widehat{I}_{\mathrm{XYZ}}$ to a white-balanced camera RGB using the interpolated forward model, and (iii) apply an illuminant gain in camera RGB space to obtain the illuminated camera RGB, which we treat as the target camera's linear raw-RGB representation. This mapping is deterministic and non-trainable, enabling re-rendering of the same scene-referred output across cameras without retraining. Full details of matrix interpolation and illuminant sampling are provided in the supplementary material.

Optional Noise and CFA. If required, we synthesize sensor noise using a heteroscedastic noise model to better match the noise characteristics of real raw camera images; we refer the reader to the supplementary material for details. While more advanced noise synthesis models (e.g., [1]) can produce more realistic noise characteristics, detailed noise modeling is beyond the scope of this paper. Finally, if required by the downstream task, the resulting linear RGB raw image can be re-mosaiced according to a specified color filter array (CFA) pattern to produce a single-channel mosaiced raw image.

4 Experiments

4.1 Training Details and Experiment Overview

Training Details. We train RawGen using the MIT-Adobe FiveK [10] and RAISE [17] datasets. Both datasets provide large-scale raw DNG images covering diverse scenes and subjects, including urban and natural environments, as well as animals and human portraits. To obtain CIE XYZ anchor targets, we parse the `AsShotNeutral` and `ForwardMatrix` tags from each DNG file. We then apply the software ISP [54] with randomized photo-finishing parameters to render diverse sRGB variants from the same anchor (Sec. 3.1). For the DiT backbone, we adopt FLUX.1-Kontext [5], which natively supports image-context conditioning. We fine-tune the DiT model via LoRA with rank r=64 and scaling α=64.

Experiment Overview. As described in Sec. 3, once trained, RawGen synthesizes scene-referred linear representations under two conditioning modes: (i) an input sRGB image (I2R) and (ii) a text prompt (T2R). We evaluate RawGen from three perspectives. First, we assess its many-to-one reconstruction capability by recovering a canonical linear XYZ representation from sRGB inputs with diverse and unknown photo-finishing (Sec. 4.2). Second, we evaluate device-specific raw synthesis by mapping canonical outputs to target camera domains and examining their distribution alignment with real raw data (Sec. 4.3). Third, we study practical applications enabled by RawGen, including scalable T2R-based data generation for downstream low-level vision tasks and raw-domain image editing. (Sec. 4.4). Together, these experiments demonstrate that RawGen enables physically grounded linear synthesis while supporting camera-aware and scalable raw generation.

4.2 Evaluation of Many-to-One sRGB-to-XYZ Invertibility

RawGen projects sRGB inputs containing unknown and diverse photo-finishing effects into a shared canonical scene-referred XYZ domain, suppressing photo-finishing variability while preserving scene content. To evaluate this many-to-one inverse-ISP capability, we conduct two complementary analyses using the Image-to-Raw (I2R) inference pathway (Fig. 2C): (1) expert-retouched real variations and (2) text-guided synthetic variations. We compare RawGen against CIE XYZ Net [2], InvISP [62], and Raw-Diffusion [48]. Since these baselines typically assume a one-to-one correspondence between an sRGB input and its reconstruction target (e.g., raw or XYZ), we train all methods to invert the ISP induced by the *default* configuration of our software ISP, using the corresponding (default sRGB, anchor XYZ) pair as supervision.

Expert-Retouched Real Variations. For this experiment, we use the MIT-Adobe FiveK dataset [10], where each scene is edited by five experts (A–E) with distinct aesthetic preferences. For each scene, the five expert-retouched sRGB images are fed into RawGen, and reconstruction accuracy is measured against the reference scene-referred XYZ target using PSNR and SSIM. All evaluations

Table 1: Quantitative comparison of sRGB-to-XYZ reconstruction across five photo-finishing styles on the MIT-Adobe FiveK dataset [10]. We report results for ablated variant of RawGen. Best results are highlighted in **yellow**.

Method	Expert A		Expert B		Expert C		Expert D		Expert E	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CIE XYZ Net [2]	19.60	0.7861	21.04	0.8431	19.49	0.7795	19.44	0.8058	18.64	0.7918
InvISP [62]	16.04	0.6854	16.04	0.6854	14.92	0.6323	14.24	0.6802	13.30	0.6473
Raw-Diffusion [48]	19.30	0.7832	20.66	0.8397	18.79	0.7705	18.69	0.7966	17.49	0.7751
$\text{RawGen}_{\text{one-to-one}}$	19.57	0.7782	21.21	0.8349	19.48	0.7741	18.74	0.7929	18.07	0.7839
$\text{RawGen}_{\text{tuned denoiser}}$	**23.36**	0.8407	24.35	0.8540	**23.40**	**0.8390**	23.50	0.8476	23.77	0.8448
RawGen (Ours)	**23.20**	**0.8432**	**24.35**	**0.8581**	**23.37**	**0.8387**	**23.51**	**0.8531**	**23.89**	**0.8500**

Table 2: Ablation study results on training protocol. The result shows the average sRGB-to-XYZ reconstruction over Experts A–E on MIT-Adobe FiveK [10]. Deterministic baselines (CIE XYZ Net, InvISP) are *retrained* under our many-to-one (N-to-1) protocol; Even when deterministic baselines are trained on the same many-to-one data, RawGen remains best in every metric.

Method	**PSNR** ↑	**SSIM** ↑	ΔE_{00} ↓
CIE XYZ Net [2] (N-to-1)	21.69	0.8070	9.66
InvISP [62] (N-to-1)	8.92	0.2713	44.10
RawGen (Ours)	**23.66**	**0.8486**	**8.39**

are conducted on test splits unseen during training for both RawGen and the compared methods. As a baseline, we evaluate $\text{RawGen}_{\text{one-to-one}}$, which uses the same architecture as RawGen but is trained with strict one-to-one supervision on (default sRGB, anchor XYZ) pairs rather than our many-to-one training scheme.

As shown in Tab. 1, RawGen consistently generalizes to unseen photo-finishing styles and more accurately recovers scene-referred linear XYZ representations than prior methods. The performance drop of $\text{RawGen}_{\text{one-to-one}}$ further highlights the importance of domain-level many-to-one supervision. Qualitative reconstructions and error maps in Fig. 3 further corroborate RawGen's robustness under diverse and previously unseen edits.

Many-to-One Retraining of Baselines Ablation. A natural question is whether deterministic inverse-ISP networks could match RawGen if granted the same many-to-one supervision. Tab. 2 reports average reconstruction over Experts A–E when CIE XYZ Net [2] and InvISP [62] are *retrained* under our many-to-one (N-to-1) protocol. Retraining substantially improves these baselines over their one-to-one counterparts (e.g., CIE XYZ Net: PSNR 19.64 → 21.69 dB), confirming that the many-to-one formulation, not merely our backbone, is responsible for the gains. Nonetheless, RawGen still achieves the highest overall performance, demonstrating that beyond the many-to-one protocol, our DiT-based generative backbone is highly effective at addressing the challenging many-to-one restoration task.

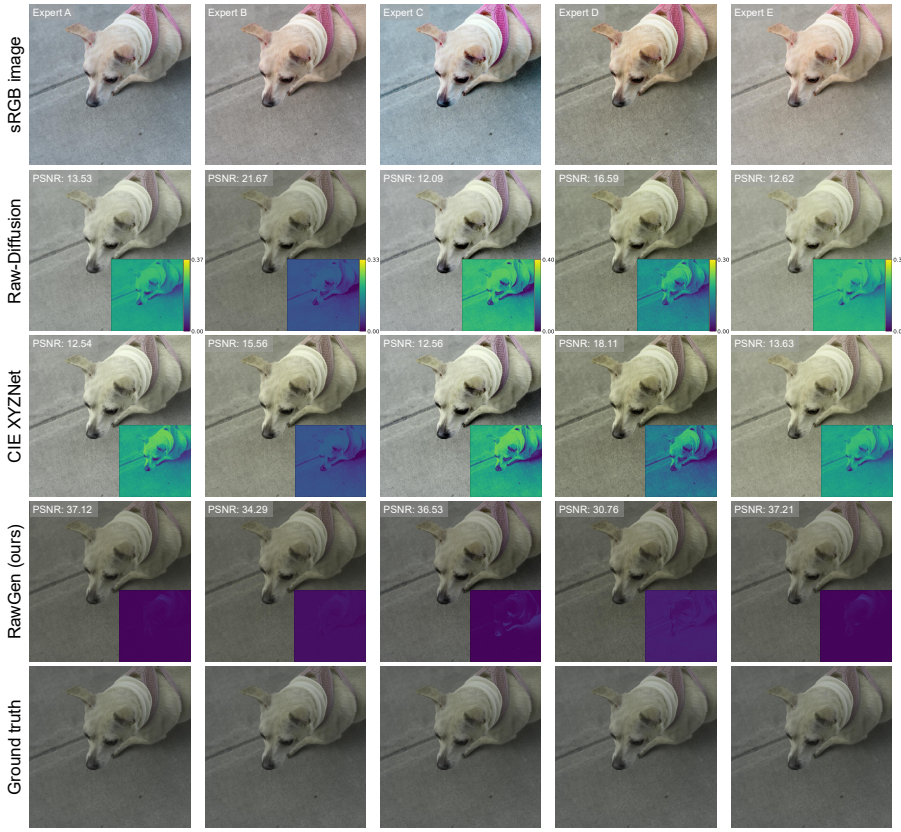

Fig. 3: CIE XYZ reconstruction results. Shown is an sRGB input image rendered with different rendering preferences (Expert A–E) and the corresponding CIE XYZ reconstructions produced by Raw-Diffusion [48], CIE XYZ Net [2], and our RawGen. The last row shows the ground-truth CIE XYZ image.

Text-Augmented Synthetic Variations. To further examine many-to-one inverse-ISP capability, we construct a larger and more diverse set of sRGB variants using text-guided generation. Unlike the real-variant experiment above, which relies on five expert retouchings per scene, this setting synthesizes controlled color variations for a broader evaluation of linear invertibility.

We use the pre-trained FLUX.1-Kontext [5] in a two-stage process. First, an anchor sRGB image is generated from a prompt (e.g., "`capture narrow alley after rainfall`"). The model is then conditioned on this anchor together with 100 color-grading prompts (e.g., "`warm cast:`", "`cool cast:`", "`teal-and-orange grade:`"), producing 100 variants. Leveraging the image-conditioning feature of FLUX.1-Kontext, scene structure and texture are preserved, while the text prompts induce diverse global color shifts. These variants are converted to XYZ by each evaluated method, with RawGen operating via the Image-to-Raw inference pathway (Fig. 2C).

Table 3: Latent space compactness of sRGB-to-XYZ conversion methods. 100 color-graded sRGB variants per prompt are converted to XYZ and encoded into the VAE latent space; lower mean L2 distance to the centroid indicates more compact clusters.

Method	Mean distance to centroid ↓		
	PCA	t-SNE	UMAP
sRGB	286.8	27.65	2.099
InvISP	265.4	24.33	1.913
XYZNet	256.1	25.52	2.014
RAW-Diffusion	318.2	19.58	2.557
RawGen	**160.0**	**10.69**	**1.067**

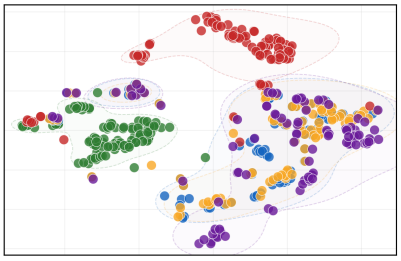

Fig. 4: t-SNE visualization of VAE latent spaces for each sRGB-to-XYZ conversion method. Dashed contours indicate the estimated density region per method. RawGen produces the most compact cluster. Marker colors follow Tab. 3.

To assess many-to-one consistency, the reconstructed outputs are encoded into the VAE latent space. The latent embeddings are projected using PCA [41], t-SNE [36], and UMAP [37]. PCA captures dominant linear variance, while t-SNE and UMAP preserve local neighborhood structure under nonlinear embeddings. Within each projected space, we compute the mean L2 distance to the centroid across variants. Lower distances indicate stronger suppression of photo-finishing variability and greater convergence toward a canonical representation. Tab. 3 reports the quantitative results, showing that RawGen produces more compact latent clusters than competing methods under PCA, t-SNE, and UMAP projections, whereas other approaches yield more dispersed embeddings, indicating weaker suppression of photo-finishing variability. The t-SNE visualization in Fig. 4 further confirms that RawGen forms a tight and well-separated cluster with reduced overlap. This result aligns with the many-to-one objective, which encourages diverse sRGB variants of the same scene to converge toward a canonical XYZ representation and promotes invariance to photo-finishing effects.

4.3 Device-Specific Raw Synthesis

RawGen produces canonical linear XYZ representations independent of any specific camera. When a target device is specified, these outputs are mapped to the corresponding raw domain using device-specific color correction matrices (e.g., `ForwardMatrix`) and metadata, with heteroscedastic Gaussian noise injected to approximate sensor measurements. This decoupled design enables camera-agnostic linear generation and device-specific raw synthesis without retraining. Figure 5 presents examples, including reconstruction from sRGB inputs with unknown edits (A–C) and multiple raw samples generated from text prompts via InstructBLIP-2 [15] (D), supporting scalable raw-domain augmentation.

We further assess distribution alignment between RawGen-synthesized and real-world raw images. Specifically, we evaluate neural ISP models trained on

Fig. 5: Raw reconstruction and generation results. Shown are the input sRGB images (A) and corresponding ground-truth raw images (B). Our reconstructed raw results are shown in (C). Additional generated raw images of the same semantic scene are shown in (D), where InstructBLIP-2 [15] is used to generate text descriptions from image (A).

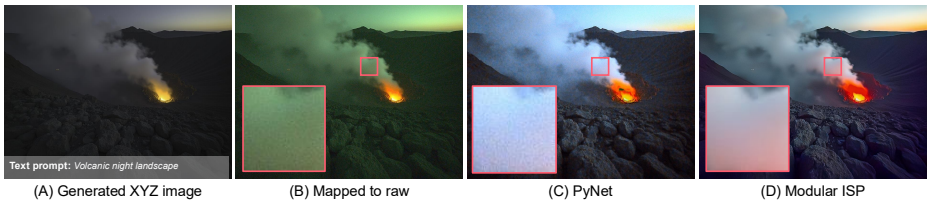

Fig. 6: Rendering results of pre-trained neural ISPs on real data from the Samsung S24 main camera (S24 dataset [4]). (A) CIE XYZ image generated by our RawGen. (B) Image mapped to the S24 main camera raw space with synthetic noise. (C) Result of PyNet [28]. (D) Result of Modular Neural ISP [3].

real raw data from the Samsung Galaxy S24 main camera (S24 dataset [4]) on RawGen-generated raw images mapped to the same device domain, without retraining. As shown in Figure 6, both ISP models produce plausible sRGB renderings. Notably, among the two evaluated ISPs, Modular Neural ISP includes an explicit denoising module and effectively suppresses the injected synthetic noise, indicating strong distribution alignment with real device-specific raw statistics.

4.4 Application of RawGen

Scalable Raw Synthesis for Downstream Tasks. Various low-level vision studies that rely on raw images have faced difficulties in data acquisition due to device-specific domain shifts, and collecting raw data at scale has required substantial human labor and extensive capture time. RawGen's Text-to-Raw (T2R) synthesis flow enables scalable generation of diverse device-specific raw images without physical capture, making large-scale augmentation practical for low-level vision tasks. Having established the realism of these synthesized raw samples (Sec. 4.3), we evaluate whether they can effectively support downstream tasks that require camera-specific raw data.

Table 4: Quantitative comparison of downstream tasks: illuminant estimation (reported as angular error statistics), neural ISP learning, and denoising at two ISO levels. Each row corresponds to a model trained using data generated by the method listed. Models trained on real data are shown in the bottom row as a reference. The best results of models trained on *non-real* raw data are highlighted in **yellow**.

Method	Illuminant estimation			Neural ISP			Denoising			
							ISO 1600		ISO 3200	
	Mean	Median	Worst 25%	PSNR	SSIM	ΔE	PSNR	SSIM	PSNR	SSIM
EnlightenGAN [30]	7.01	6.82	11.07	35.58	0.965	3.137	48.82	0.991	47.25	0.988
UPI [9]	6.26	5.89	10.33	36.43	0.966	2.907	49.05	0.990	47.51	0.988
Graphics2RAW [54]	4.21	3.38	8.57	38.10	**0.974**	2.301	49.37	0.991	48.16	0.989
RawGen	**3.14**	**2.11**	**7.37**	**38.42**	0.970	**2.183**	**50.63**	**0.994**	**48.57**	**0.992**
Real	3.02	2.17	6.77	38.32	0.974	2.133	49.80	0.993	48.25	0.990

We evaluate impact of the scalability on three tasks: illuminant estimation, neural ISP learning, and raw-domain denoising, all requiring device-specific raw data. For controlled comparison, we follow the same datasets, training procedures, and evaluation protocols as Graphics2RAW [54]. For each task, we synthesize 3K samples. Text prompts for T2R generation are obtained by captioning RAISE [17] images with InstructBLIP-2 [15] (using the instruction "`briefly describe this photo`") and prepending the prefix "`Realistic photography:`" to each caption before generation. These synthetic samples are used only for training, and performance is measured on the corresponding real-world test splits without modification, isolating the effect of the synthetic data source while enabling direct comparison with prior methods.

The scalability advantage is particularly salient for illuminant estimation, which requires training data for nine cameras. In the real NUS-8 dataset, each camera provides only 217 images on average, and prior graphics-based raw synthesis [54] is further constrained by limited assets, yielding 125 synthetic raw images for training. In contrast, RawGen can generate diverse, camera-specific raw images directly from text prompts at scale, substantially expanding training coverage without additional capture effort or scene asset design. Detailed experimental settings are provided in the supplementary material.

Tab. 4 shows the results. RawGen-based T2R improves performance across all three tasks compared to prior synthetic raw generation approaches. These results indicate that scalable and physically grounded raw synthesis can alleviate data acquisition bottlenecks in device-specific low-level vision tasks.

Generative Image Editing. Image editing is more reliable in a scene-referred linear space, where radiometric linearity and a wider dynamic range are preserved, enabling physically grounded operations such as white balance and exposure scaling [2]. Conventional inverse-ISP methods [2, 7, 48, 62] enable linear-domain editing but are often tuned to specific devices and fixed imaging assumptions, which can lead to failures when inverting heterogeneous sRGB inputs outside the training distribution (Sec. 4.2) and degrade editing quality.

Camera-controllable generation frameworks such as Generative Photography [63] support parameter-driven edits (e.g., white balance, shutter speed, and f-number) while preserving scene context, but they operate entirely in the sRGB

domain without access to an underlying raw image. Consequently, applying a new parameter setting typically requires rerunning an iterative diffusion process, rather than performing ISP operations on linear sensor data. Moreover, their controllability is limited to the camera parameters explicitly learned by the model, and supporting additional types of edits or new control dimensions generally requires retraining or additional fine-tuning.

In contrast, RawGen decouples content synthesis from rendering by producing a scene-referred linear image from either an input image or a text prompt, and then enabling edits via standard software-ISP operations. After generating the linear representation once, a wide range of edits, including white balance, exposure compensation, and tone mapping, can be applied directly in the linear domain using an arbitrary software ISP pipeline without re-invoking the generative model. This supports efficient exploration of diverse editing operations beyond a fixed set of learned camera parameters.

5 Conclusion and Discussion

We present RawGen, a diffusion-based framework for camera-agnostic raw generation that bridges display-referred generative models and scene-referred linear representations. Using a many-to-one objective, RawGen suppresses photofinishing effects and recovers a canonical XYZ representation consistent across diverse sRGB variants. We demonstrated many-to-one linear XYZ reconstruction and device-specific raw synthesis. Moreover, scalable text-to-raw data synthesis enables RawGen to improve downstream low-level vision tasks.

Limitations. While RawGen generates canonical scene-referred XYZ representations, accurate device-specific raw synthesis also depends on factors beyond color mapping, such as noise characteristics and PSFs. Future work may explore conditioning RawGen on device-specific priors to incorporate more physically grounded sensor modeling, including spatially varying noise, lens shading, and optical blur, as well as learning the camera-specific mapping itself to better capture spectral and dynamic-range differences, further improving device fidelity and physical consistency.

Acknowledgments

We thank Ran Zhang for assistance with rendering the generated raw images using neural ISP models.

References

1. Abdelhamed, A., Brubaker, M.A., Brown, M.S.: Noise flow: Noise modeling with conditional normalizing flows. In: ICCV (2019)
2. Afifi, M., Abdelhamed, A., Abuolaim, A., Punnappurath, A., Brown, M.S.: CIE XYZ Net: Unprocessing images for low-level computer vision tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 4688–4700 (2021)

3. Afifi, M., Wang, Z., Zhang, R., Brown, M.S.: Modular neural image signal processing. arXiv preprint arXiv:2512.08564 (2025)
4. Afifi, M., Zhao, L., Punnappurath, A., Abdelsalam, M.A., Zhang, R., Brown, M.S.: Time-aware auto white balance in mobile photography. In: ICCV (2025)
5. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., et al.: FLUX. 1 Kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
6. Bemana, M., Leimkühler, T., Myszkowski, K., Seidel, H.P., Ritschel, T.: Bracket diffusion: HDR image generation by consistent LDR denoising. Computer Graphics Forum **44**(2), e70086:1–13 (2025)
7. Berdan, R., Besbinar, B., Reinders, C., Otsuka, J., Iso, D.: ReRAW: RGB-to-raw image reconstruction via stratified sampling for efficient object detection on the edge. In: CVPR (2025)
8. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., Ramesh, A.: Improving image generation with better captions. Tech. rep., OpenAI (2023), `https://cdn.openai.com/papers/dall-e-3.pdf`, technical report
9. Brooks, T., Mildenhall, B., Xue, T., Chen, J., Sharlet, D., Barron, J.T.: Unprocessing images for learned raw denoising. In: CVPR (2019)
10. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input / output image pairs. In: CVPR (2011)
11. Chakrabarti, A., Scharstein, D., Zickler, T.E.: An empirical camera model for internet color vision. In: BMVC (2009)
12. Chakrabarti, A., Xiong, Y., Sun, B., Darrell, T., Scharstein, D., Zickler, T., Saenko, K.: Modeling radiometric uncertainty for vision with tone-mapped color images. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(11), 2185–2198 (2014)
13. Chen, Y., Wen, Y., Li, W., Liu, J., Guo, Y., Hu, J., Chen, X.: RDDM: Practicing raw domain diffusion model for real-world image restoration. arXiv preprint arXiv:2508.19154 (2025)
14. Conde, M.V., McDonagh, S., Maggioni, M., Leonardis, A., Pérez-Pellitero, E.: Model-based image signal processors via learnable dictionaries. In: AAAI (2022)
15. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. NeurIPS (2023)
16. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., et al.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)
17. Dang-Nguyen, D.T., Pasquini, C., Conotter, V., Boato, G.: Raise: A raw images dataset for digital image forensics. In: ACM Multimedia Systems (2015)
18. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: ACM SIGGRAPH (2008)
19. Delbracio, M., Kelly, D., Brown, M.S., Milanfar, P.: Mobile computational photography: A tour. Annual review of vision science **7**(1), 571–604 (2021)
20. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. NeurIPS (2021)
21. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)

22. Grossberg, M.D., Nayar, S.K.: Determining the camera response from images: What is knowable? IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(11), 1455–1467 (2003)
23. He, X., Hu, T., Wang, G., Wang, Z., Wang, R., Zhang, Q., Yan, K., Chen, Z., Li, R., Xie, C., et al.: Enhancing raw-to-sRGB with decoupled style structure in Fourier domain. In: AAAI (2024)
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
25. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
26. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
27. Ignatov, A., Sycheva, A., Timofte, R., Tseng, Y., Xu, Y.S., Yu, P.H., Chiang, C.M., Kuo, H.K., Chen, M.H., Cheng, C.M., et al.: MicroISP: Processing 32MP photos on mobile devices with deep learning. In: ECCV (2022)
28. Ignatov, A., Van Gool, L., Timofte, R.: Replacing mobile camera ISP with a single deep learning model. In: CVPRW (2020)
29. Jiang, H., Luo, A., Fan, H., Han, S., Liu, S.: Low-light image enhancement with wavelet-based diffusion models. ACM Transactions on Graphics (TOG) **42**(6), 1–14 (2023)
30. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: EnlightenGAN: Deep light enhancement without paired supervision. IEEE Transactions on Image Processing **30**, 2340–2349 (2021)
31. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. NeurIPS (2022)
32. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. NeurIPS **35**, 23593–23606 (2022)
33. Kim, S.J., Lin, H.T., Lu, Z., Süsstrunk, S., Lin, S., Brown, M.S.: 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)
34. Kim, W., Kim, G., Lee, J., Lee, S., Baek, S.H., Cho, S.: ParamISP: Learned forward and inverse ISPs using camera parameters. In: CVPR (2024)
35. Luo, A., Li, X., Yang, F., Liu, J., Fan, H., Liu, S.: FlowDiffuser: Advancing optical flow estimation with diffusion models. In: CVPR (2024)
36. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**, 2579–2605 (2008)
37. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
38. Mitsunaga, T., Nayar, S.K.: Radiometric self calibration. In: CVPR (1999)
39. Nam, S., Joo Kim, S.: Modelling the scene dependent imaging in cameras with a deep neural network. In: ICCV (2017)
40. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML (2021)
41. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philosophical Magazine **2**(11), 559–572 (1901)
42. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
43. Pernias, P., Rampas, D., Richter, M.L., Pal, C.J., Aubreville, M.: Würstchen: An efficient architecture for large-scale text-to-image diffusion models. arXiv preprint arXiv:2306.00637 (2023)
44. Podell, D., Weng, C.I., Onoe, Y., et al.: SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

45. Raimundo, D.W., Ignatov, A., Timofte, R.: LAN: Lightweight attention-based network for raw-to-RGB smartphone image processing. In: CVPRW (2022)
46. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 (2022)
47. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021)
48. Reinders, C., Berdan, R., Besbinar, B., Otsuka, J., Iso, D.: Raw-diffusion: RGB-guided diffusion models for high-fidelity raw image generation. In: WACV (2025)
49. Ren, Y., Jiang, H., Yang, M., Li, W., Liu, S.: ISPDiffuser: Learning RAW-to-sRGB mappings with texture-aware diffusion models and histogram-guided color consistency. In: AAAI (2025)
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
51. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: SIGGRAPH (2022)
52. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
53. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022)
54. Seo, D., Punnappurath, A., Zhao, L., Abdelhamed, A., Tedla, S.K., Park, S., Choe, J., Brown, M.S.: Graphics2RAW: Mapping computer graphics images to sensor raw images. In: ICCV (2023)
55. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
56. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. arXiv e-prints (2023)
57. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. NeurIPS (2019)
58. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
59. Wang, C., Xia, Z., Leimkuhler, T., Myszkowski, K., Zhang, X.: LEDiff: Latent exposure diffusion for HDR generation. In: CVPR (2025)
60. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. arXiv preprint arXiv:2212.00490 (2022)
61. Wang, Y., Yu, Y., Yang, W., Guo, L., Chau, L.P., Kot, A.C., Wen, B.: ExposureDiffusion: Learning to expose for low-light image enhancement. In: ICCV (2023)
62. Xing, Y., Qian, Z., Chen, Q.: Invertible image signal processing. In: CVPR (2021)
63. Yuan, Y., Wang, X., Sheng, Y., Chennuri, P., Zhang, X., Chan, S.: Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. In: CVPR (2025)
64. Zamir, S.W., Arora, A., Khan, S., et al.: CycleISP: Real image restoration via improved data synthesis. In: CVPR (2020)
65. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)
66. Zhang, Z., Wang, H., Liu, M., Wang, R., Zhang, J., Zuo, W.: Learning raw-to-sRGB mappings with inaccurately aligned supervision. In: ICCV (2021)

67. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024)