RAWild: Sensor-Agnostic RAW Object Detection via Physics-Guided Curve and Grid Modeling

Shuhong Liu[1,2,♠], **Gengjia Chang**[2,♠], **Jun Liu**[2], **Xuangeng Chu**[1,2]
Yinqiang Zheng[1], **Tatsuya Harada**[1,3], **Ziteng Cui**[1,◇]
[1]The University of Tokyo [2]I2WM [3]RIKEN



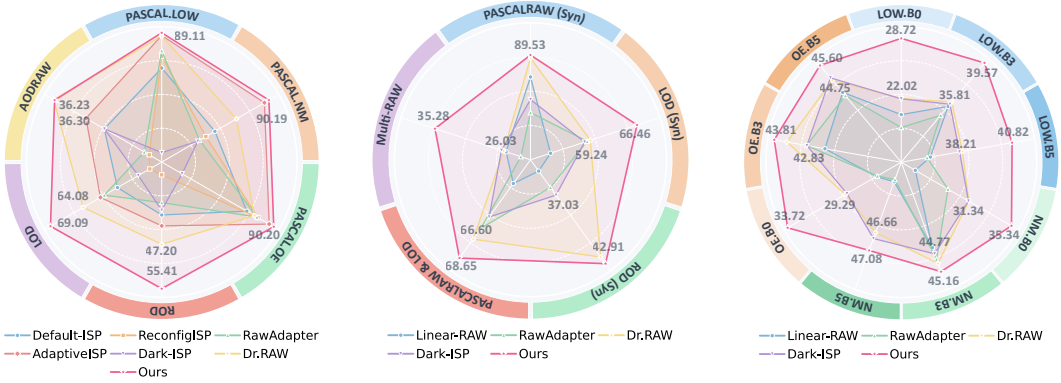
Figure 1: RAW object detection across sensors. Four benchmarks [35, 49, 60, 100] captured by distinct sensors under different environments, exposure settings, spectral sensitivities, and bit-depths exhibit substantially different pixel distributions, posing a fundamental challenge for sensor-agnostic generalization.

Abstract

Camera sensor RAW data offers intrinsic advantages for object detection, including deeper bit depth, preserved physical information, and freedom from image signal processor (ISP) distortions. However, varying exposure conditions, spectral sensitivities, and bit depths across devices introduce substantially larger domain gaps than in sRGB, making sensor-agnostic generalization a fundamental challenge. In this study, we present **RAWild**, a physics-guided global-local tone mapping framework for sensor-agnostic RAW object detection. By factoring sensor-induced variations into a global tonal correction and a spatially adaptive local color adjustment, both driven by RAW distribution priors, our framework enables a single network to train jointly across heterogeneous sensors. To further support cross-sensor generalization, we construct a physics-based RAW simulation pipeline that synthesizes realistic sensor outputs spanning diverse spectral sensitivities, illuminants, and sensor non-idealities. Extensive experiments across multiple RAW benchmarks covering bit depths from 10 to 24 demonstrate state-of-the-art (SOTA) performance under single-dataset, mixed-dataset, and challenging robustness settings. The code is available at https://i2wm.github.io/RAWild/.

♠These authors contributed equally to this work. ◇Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Detection (b) Multi-dataset detection (c) Segmentation

Figure 2: Radar comparison with representative baselines on (a) single-dataset RAW object detection, (b) mixed-sensor dataset RAW object detection, and (c) RAW image segmentation.

1 Introduction

Object detection stands as one of the most fundamental challenges in computer vision. In recent years, there has been a growing shift in focus from conventional sRGB-based detection to operating directly on camera sensor RAW data [18, 27, 92, 98], driven by several intrinsic advantages of camera RAW: deeper bit depth, richer physical information, and freedom from the nonlinear distortions introduced by the Image Signal Processor (ISP) pipeline [40]. These properties endow RAW-based object detection with superior robustness under extreme conditions, such as challenging illumination and adverse weather [49], thereby holding significant practical value for real-world applications, including embodied intelligence, surveillance, and autonomous driving [78].

Unlike sRGB images, which ISP pipelines normalize into a unified color space [83], RAW sensor data preserves device-specific characteristics such as spectral sensitivity, native bit depth, and CFA layout. These properties vary considerably across cameras. For instance, Figure 1 shows that the Sony IMX490 sensor in the ROD dataset [100] captures 24-bit RAW images with a standard Bayer CFA, whereas the Canon 5D-IV camera in the LOD dataset [35] records 16-bit RAW images with a distinct spectral response. Even for identical scenes, hardware discrepancies induce much larger domain gaps in RAW than in sRGB [66]. More fundamentally, RAW appearance is shaped by a chain of factors spanning illumination, capture settings, and sensor hardware. The illuminant spectrum and radiance determine the incident photon flux, while shutter time and ISO gain scale the accumulated signal. The color filter response, lens-shading profile, near-saturation non-linearity, and ADC bit depth then jointly govern how the signal is recorded. Each factor varies by orders of magnitude across cameras and capture conditions, and their effects compound nonlinearly in the RAW image.

To address these discrepancies, previous methods either learn the end-to-end ISP pipeline [27, 92, 102] or insert learnable ISP-style blocks by stages [18, 24]. Such designs remain tightly coupled to specific sensor configurations and rely on handcrafted operator chains. Consequently, they struggle to establish a unified pipeline across heterogeneous data [74], limiting their scalability across diverse environmental illuminations, camera responses, and bit-depths.

To achieve sensor-agnostic robustness, we propose **RAWild**, a histogram-guided global-local tone mapping framework. Lightweight ***differentiable Bézier curves*** first handle global tonal adjustment, producing physically plausible illumination correction. The ***Bilateral Grid*** then performs local, spatially-adaptive color adjustment in a compact 3D bilateral space [9], where learned affine transformations deliver efficient, edge-aware detail enhancement at negligible computational cost and preserve the high-frequency structures relevant for detection. Statistical priors extracted from the RAW pixel distribution jointly guide both the curve parameters and the grid coefficients, yielding a content-aware, scene-adaptive mapping that generalizes robustly across diverse lighting conditions. Extensive experiments demonstrate that our method achieves SOTA performance not only on multiple established RAW benchmark datasets but also under mixed-dataset training regimes spanning

heterogeneous sensors and bit-depths, as well as more challenging robustness scenarios involving severe degradation and domain shift, as illustrated in Figure 2.

Our key contributions could be summarized as follows:

- We propose **RAWild**, a histogram-guided global-local tone mapping framework for sensor-agnostic RAW object detection. It decomposes sensor variations into a global Bézier curve for tonal correction and a spatially-adaptive Bilateral Grid for local color refinement.
- We reparametrize both components with a delta-residual Bézier curve, and a gain–mixing decomposition of the Bilateral Grid that separates per-channel gain from color mixing, enabling stable training across heterogeneous RAW images.
- Extensive experiments on six per-sensor and mixed-sensor real-capture RAW benchmarks and three datasets synthesized by our physics-based RAW simulation pipeline, spanning 10- to 24-bit sensors, demonstrate SOTA performance in most settings.

2 Related Work

Camera RAW-based Object Detection. RAW data preserves linear photon measurements with high dynamic range and avoids the irreversible information loss of ISP processing [21, 40], benefiting perception in low light [7, 35, 37] and HDR scene reconstruction [39, 41, 48, 56]. Detection-oriented efforts either jointly learn task-aware ISPs with the detector [43, 46, 92, 102], adapt sRGB-pretrained models to RAW with adaptive modules [18, 36], or build RAW-native detectors spanning embedded [60], low-light [27], and diverse-condition settings [24, 45, 47, 49, 98, 100]. These detectors remain optimized per sensor and are brittle to cross-sensor radiometric shifts.

Image Signal Processing with Bilateral Grid. Bilateral Grid [8, 9] embeds pixels in a 3D spatial-intensity space for efficient edge-aware operations sliced back to full resolution; HDRnet [26] first paired it with deep learning to predict spatially-varying color transforms, joining other learned ISP components [25, 32]. Follow-ups cascade grids with neural 3D LUTs for retouching [17, 42, 104], embed per-view grids in radiance fields to disentangle ISP effects [87, 89], or combine them with learned tonal curves and color-name priors [44, 50, 80]. We are the first to apply Bilateral Grids to RAW detection and couple them with a global Bézier curve to achieve physics-aware tonal mapping.

Sensor-Agnostic Computer Vision. Sensor-agnostic learning unifies heterogeneous inputs so one model generalizes across modalities, e.g., depth completion across LiDAR/ToF [64], 3D reconstruction across fisheye and rolling-shutter cameras [95], and video generation across diverse optics [107]. In the RAW/ISP domain, well-documented cross-sensor variation in spectral sensitivity [38] and radiometric response [31] has motivated camera-prior approaches such as unpaired RAW-to-RAW translation [66], parameter-aware forward/inverse ISPs [43], and metadata/EXIF-conditioned models [1, 109]. Unlike these, we target sensor-agnostic RAW *detection*, addressing bit depth, spectral sensitivity, and exposure via a physics-guided global-local framework.

3 Method

3.1 Motivation

Many of the factors discussed in the introduction collapse along common axes. The illuminant spectrum and camera spectral response jointly control the per-channel color balance, i.e., color temperature and tint. Scene radiance, shutter time, and analog gain together scale the overall signal level, i.e., exposure. The sensor's non-linear response and ADC bit depth govern tonal compression and quantization at highlights and shadows. We train a linear-RAW-stem detector on three RAW benchmarks [35, 49, 100] and evaluate it under controlled shifts along each axis. As shown in Figure 3, exposure and bit-depth variation cause the strongest degradation, while color-temperature variation has a smaller effect over the evaluated range. These results motivate adaptive tone mapping across complementary scales.

The global Bézier curve g handles image-wide tonal and per-channel sensitivity variation, while the Bilateral Grid, which stores a spatially- and luminance-adaptive color matrix M, handles residual

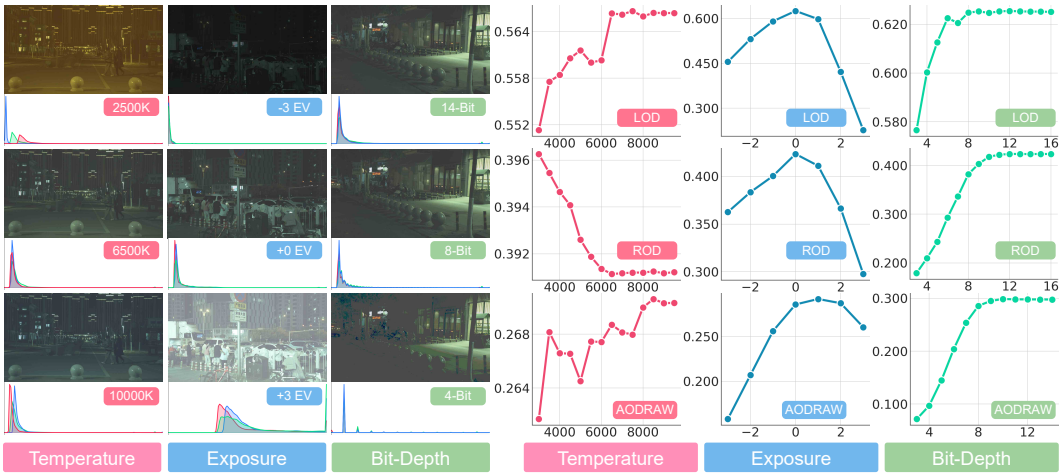


Figure 3: Sensitivity of RAW object detection to environmental and sensor-induced variations. Left: three AODRAW frames rendered under shifted illumination color temperature, exposure, and sensor bit depth, each paired with its pixel histogram. Right: mAP@50 of a pretrained ResNet-50 backbone with a linear RAW stem on LOD, ROD, and AODRAW as the three factors sweep across their ranges.

spatially and luminance-varying tonal and color corrections. Both operate on a channel-aligned linear RAW tensor: after black- and white-level normalization, each 2×2 Bayer unit of the mosaic X is packed into one three-channel linear sample, which for an RGGB pattern reads

$$I(u,v) \;=\; \Big[\, X(2u,2v),\;\; \tfrac{1}{2}\big(X(2u,2v{+}1)+X(2u{+}1,2v)\big),\;\; X(2u{+}1,2v{+}1)\,\Big].$$

Other CFA layouts reorder the coordinates according to the recorded Bayer pattern. On this tensor, the output y takes the form:

$$y(p) \;=\; M\big(p,\,\ell(p)\big)\,g\big(I(p)\big), \qquad \ell(p) \;=\; \frac{1}{3}\sum_{c} I_c(p), \tag{1}$$

where $I, y \in [0,1]^{3\times H\times W}$ are the packed linear RAW input and the backbone-ready output. Crucially, $\ell(p)$ is computed from the unmapped I rather than $g(I)$, so its luminance axis retains a fixed physical meaning across sensors and training stages. We parameterize g as a degree-n Bézier curve (Section 3.2) and store M as a low-resolution Bilateral Grid (Section 3.3); both are jointly predicted by a shared histogram-guided network (Section 3.4).

3.2 Global Differentiable Bézier Curve

As illustrated at the top of Figure 4, the global tone curve g maps the linear RAW signal through a smooth, bounded per-channel function, absorbing exposure, sensor non-linear response, and per-channel sensitivity into $3(n{-}1)$ scalar coefficients. We parameterize g as a degree-n Bézier curve per channel [80], which is pointwise bounded, monotone under ordered control points, and pinned to pure black and white at its endpoints. Let $\{p_{c,i}\}_{i=0}^{n} \subset [0,1]$ denote the $n{+}1$ control points for channel $c \in \{R,G,B\}$. Given $t = I_c(p) \in [0,1]$, the curve evaluates pointwise as:

$$g_c(t) \;=\; \sum_{i=0}^{n} B_{n,i}(t)\,p_{c,i}, \qquad B_{n,i}(t) \;=\; \binom{n}{i}(1-t)^{n-i}\,t^{i}, \tag{2}$$

where $\{B_{n,i}\}_{i=0}^{n}$ are the Bernstein basis polynomials. Since $B_{n,i}(t) \geq 0$ and $\sum_i B_{n,i}(t) = 1$, each $g_c(t)$ is a convex combination of its control points and stays within $[\min_i p_{c,i}, \max_i p_{c,i}]$. The derivative of g_c admits the closed form:

$$g_c'(t) \;=\; n\sum_{i=0}^{n-1} B_{n-1,i}(t)\,\big(p_{c,i+1}-p_{c,i}\big), \tag{3}$$

which is n times a convex combination of adjacent control-point differences. It stays non-negative under ordered control points $p_{c,0} \leq \cdots \leq p_{c,n}$ and renders g_c monotonically non-decreasing. To

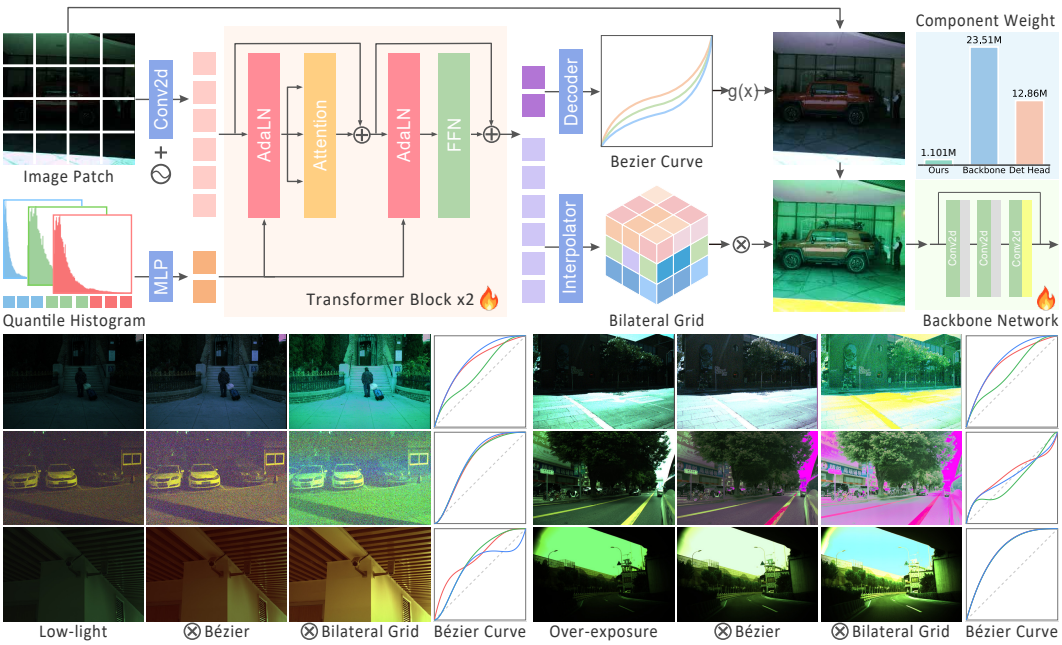


Figure 4: Overall architecture of the proposed method. A RAW image and its quantile histogram drive transformer blocks via AdaLN, jointly predicting a global Bézier curve and a Bilateral Grid for local color correction. Qualitative examples on low-light and over-exposed inputs show the Bézier-only output, the full Bilateral Grid output, and the learned curve. Linear and detector-facing images are scaled and clipped to $[0, 1]$ and quantized to 8 bits for display only.

regularize the predicted curve toward this ordered regime, we anchor the interior points on the uniform ramp $\bar{p}_{c,i} = i/n$ and apply a bounded residual that keeps each update a bounded perturbation of identity as:

$$p_{c,i} \;=\; \mathrm{clip}_{[0,1]}\Big(\tfrac{i}{n} \;+\; \tfrac{1}{2}\tanh(\delta_{c,i})\Big), \qquad i = 1, \ldots, n-1. \tag{4}$$

Our Bézier parameterization is both more expressive than a per-channel gamma [98], whose single scalar cannot fit the S-shaped response of modern sensors, and more constrained than free-form polynomial bases [27], which provide no endpoint or monotonicity guarantees and may crush or invert tones near black and white. Because $\delta_{c,i}$ is predicted per image from sensor-dependent statistics in Section 3.4, g specializes adaptively across sensors. A wide-dynamic-range Bayer sensor receives an aggressively compressive curve, whereas a narrow-range low-bit sensor receives a near-linear one. The $3(n-1)$ learned degrees of freedom thus absorb exposure, gamma, and per-channel sensitivity differences without any sensor-specific architecture.

3.3 Local Bilateral Grid Color Transform

The global curve leaves spatially-varying color shifts intact. A single RAW frame routinely contains mixed illumination, lens-shading falloff, and sensor-dependent crosstalk. Detection does not require a colorimetrically faithful image, but benefits from enhanced local contrast and balanced color that expose object boundaries and textures to the backbone. We therefore apply a spatially-adaptive 3×3 color transform, represented as a bilateral grid following [8, 26], as depicted at the bottom of Figure 4.

Grid Representation and Trilinear Slicing. The grid $\mathbf{G} \in \mathbb{R}^{9\times G_d\times G_h\times G_w}$ stores the nine coefficients of a 3×3 matrix at each cell, with spatial resolution $(H/16)\times(W/16)$ and $G_d{=}8$ luminance bins. Each full-resolution pixel p recovers its coefficients by trilinear interpolation into $\mathbf{G}$ at the continuous coordinate $(xG_w/W, yG_h/H, \ell(p)G_d)$,

$$m(p) \;=\; \sum_{(d,h,w)\in\mathcal{N}(p)} w_d(p)\, w_h(p)\, w_w(p)\, \mathbf{G}_{:,d,h,w} \;\in\; \mathbb{R}^9, \tag{5}$$

where $\mathcal{N}(p)$ is the eight enclosing grid vertices and (w_d, w_h, w_w) are the linear interpolation weights. The sampled vector reshapes into a 3×3 matrix $M(p)$ and applies pointwise as:

$$y(p) \;=\; \mathrm{clip}_{[0,1]}\big(M(p)\, g\big(I(p)\big)\big). \tag{6}$$

Prediction operates at grid resolution and rendering is $\mathcal{O}(HW)$, so the luminance axis lets pixels at the same location but different brightness sample different matrices without per-pixel prediction cost. Because the matrix also depends on pixel position, two pixels with the same RAW value can receive different transforms, which a global 3D LUT cannot represent. Appendix D.1 formalizes this difference and compares RAWild with conventional global and local tone-mapping operators.

Structured Matrix Parameterization. A free 3×3 prediction at every cell is fragile in practice. Sensor-to-sensor color differences are dominated by per-channel gain absorbing white balance, RAW-domain exposure, and spectral-sensitivity scaling, which can vary by a factor of three or more across illuminants and sensors. Channel mixing absorbs residual cross-spectral overlap and color-filter leakage, and is a small correction of at most 10 to 20% relative to the dominant channel [40]. An unconstrained 3×3 conflates these two effects at markedly different magnitudes. Gradients on the dominant gain contaminate the off-diagonals, and early-iteration matrices that are ill-conditioned or sign-flipped destroy the pretrained backbone's features. We therefore parameterize every cell of $\mathbf{G}$ as:

$$M \;=\; D\,(I + A), \qquad D = \mathrm{diag}(d),\; d = \exp\big(\tanh(\hat{d})\big),\; A_{ij|i\neq j} = k\tanh(\hat{a}_{ij}), \tag{7}$$

where $\hat{d} \in \mathbb{R}^3$ and $\hat{a} \in \mathbb{R}^6$ are the nine raw coefficients emitted by the prediction network, with fixed activation bound k=0.05. The matrix takes the explicit form:

$$M \;=\; \begin{bmatrix} d_1 & d_1A_{12} & d_1A_{13} \\ d_2A_{21} & d_2 & d_2A_{23} \\ d_3A_{31} & d_3A_{32} & d_3 \end{bmatrix}, \tag{8}$$

in which each row is a gain d_i multiplied by a unit-diagonal mixing vector. The diagonal entries lie in $[e^{-1}, e^{1}] \approx [0.37, 2.72]$, so d neither collapses to zero nor flips sign, and color channels retain their physical orientation throughout training. With $|A_{ij}| \leq k$, the matrix $I + A$ is strictly diagonally dominant and invertible, and so is $M = D(I + A)$, which prevents the detector backbone from observing degenerate inputs. Left-multiplication by D further ensures that A encodes mixing ratios relative to each row's gain rather than absolute channel deltas, so globally rescaling a channel adjusts only d_i and leaves A invariant. This decoupling eliminates the gradient interference between exposure and color-mixing corrections that plagues unconstrained 3×3 predictions.

Zero-initializing the grid head gives $\hat{d}$=$\hat{a}$=$\mathbf{0}$, hence d=$\mathbf{1}$, A=$\mathbf{0}$, and M=I. Combined with the identity-initialized Bézier curve of Equation (4), the full adapter Equation (1) is exactly the identity at initialization and thereafter learns a bounded, well-conditioned residual, allowing a single network to train jointly across heterogeneous sensors.

3.4 Histogram-Guided Curve and Grid Prediction

The Bézier residual δ and the Bilateral Grid $\mathbf{G}$ are predicted by a single shared network from the normalized linear RAW input I. Four stride-2 3×3 convolutions with channel progression 3→32→64→128→D produce a token map of shape $D \times G_h \times G_w$, which is augmented with fixed 2D sin–cos positional encodings [85] and refined by two transformer blocks.

To inject sensor-level priors, we compute a differentiable per-channel quantile descriptor $\mathbf{q} \in \mathbb{R}^{3\times Q}$ with Q=64 entries per channel as a differentiable surrogate of the channel histogram, where for each rank $r \in \{1, \ldots, Q\}$ the entry q_r is the r/Q-th empirical quantile of the pixel intensities, obtained via soft sorting for end-to-end differentiability, and project it through a small MLP to produce per-layer AdaLN [65] scale and shift parameters. The nominal bit depth used for input normalization enters this projection through a second small MLP. These modulate the attention and FFN activations of each block, so the same network adapts to sensor statistics such as bit depth, white-balance bias, exposure regime, and dynamic-range occupancy. The descriptor is equivariant to monotone rescalings of the signal, so relative spacings between successive q_r rather than absolute values encode distribution shape, yielding robustness across heterogeneous bit depths and gains.

A 1×1 convolution expands each spatial token into $G_d \cdot 9$ raw grid coefficients, reshaped and passed through Equation (7) to yield $\mathbf{G}$. A linear head maps the average-pooled token features to the Bézier

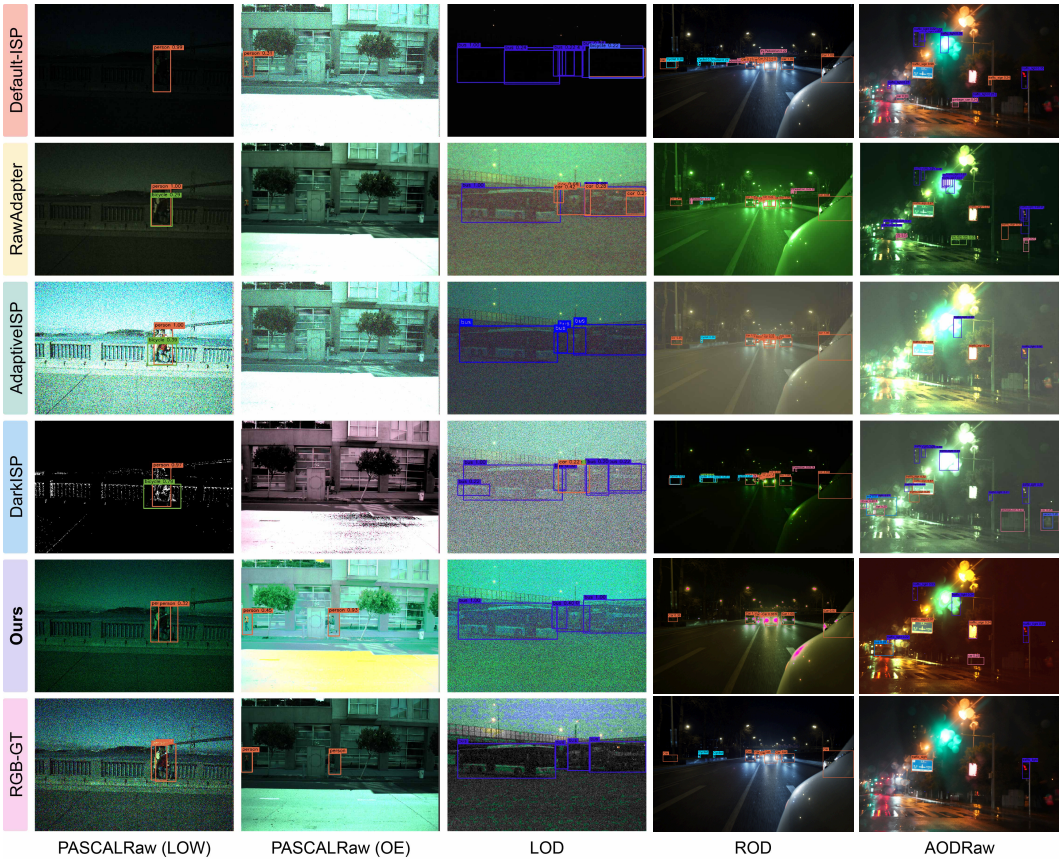


Figure 5: Qualitative comparison of object detection on PASCALRAW [60], LOD [35], ROD [100], and AODRAW [49]. Our method produces more robust and accurate detections than the baselines. Linear and detector-facing images are scaled and clipped to $[0, 1]$ and quantized to 8 bits for display only, and official sRGB images are shown directly. Zoom in for details.

residual δ, combined with the uniform anchor via Equation (4). The adapter's cost depends on HW but not on the detector, so it adds a near-constant per-image cost across detector architectures.

4 Experiment

Implementation Details. We implement our models in PyTorch with MMDetection [11]. All methods start from a pretrained RetinaNet or, with Swin Transformer, RepPointsV2 [13] and are fine-tuned with SGD (lr 1×10^{-3}, momentum 0.9, weight decay 5×10^{-4}) for 20 epochs at batch size 4, with a 500-step linear warmup. All RAW-input methods receive the same packed input. Augmentation consists of resizing and random horizontal flipping (p=0.5). Figure 4 (upper right) reports the parameter breakdown of the network components. Table 3 compares inference time and memory with representative baselines on an NVIDIA H100 GPU, and Appendix C.3 reports their parameters and FLOPs.

Baseline Methods. We compare our method with Default-ISP and recent task-oriented ISP approaches, including ReconfigISP [102], RAW-Adapter [18], AdaptiveISP [92], Dark-ISP [27], Dr.RAW [36], and RAM [24]. Appendix C.2 gives the Default-ISP and RAM protocols.

Benchmarks. For object detection on RAW images, we use the 12-bit PASCALRAW dataset [60] and its lowlight (*LOW*), normal (*NM*), and overexposure (*OE*) variants [18]. We further adopt the 14-bit AODRAW [49], 16-bit LOD [35], and 24-bit ROD [100] datasets, along with the MultiRAW dataset [47] that contains 10-, 12-, 16-, and 24-bit RAW images captured by four sensors. We use 1,991 unique training and 2,064 unique test RAW images, and the OnePlus 5T sensor appears only in

the test split, as detailed in Appendix C.1. We additionally construct a mixed-sensor dataset from PASCALRAW and LOD, and synthesize a multi-sensor dataset, with synthesis details provided in Appendix A. Additional experiments on RAW image segmentation are provided in Appendix D.4.

Table 1: Evaluation of object detection performance in mAP@50 and @75 on PASCALRAW Low/Normal/Overexposed, LOD, ROD, and AODRAW. Models use a ResNet-50 [30] or a Swin Transformer [54] backbone. For RAM [24], we report the best observed value of each metric over 20 epochs.

		PAS.LOW		PAS.NM		PAS.OE		LOD		ROD		AODRAW	
	Method	@50	@75	@50	@75	@50	@75	@50	@75	@50	@75	@50	@75
ResNet-50	Linear-RAW	0.7668	0.5793	0.8661	0.7273	0.8747	0.7298	0.5710	0.3713	0.3041	0.1963	0.2301	0.1485
	Default-ISP	0.8390	0.6676	0.8959	0.7328	0.8967	0.7333	0.5880	0.3549	0.4178	0.2709	0.3241	0.2063
	ReconfigISP	0.8562	0.6662	0.8949	0.7286	0.8989	0.7263	0.5383	0.2903	0.3436	0.2137	0.2896	0.1763
	RAW-Adapter	0.8647	0.6935	0.8942	0.7270	0.8987	0.7333	0.6082	0.3285	0.3950	0.2560	0.2934	0.1878
	AdaptiveISP	0.8871	0.7293	0.9014	0.7872	0.9011	0.7636	0.6142	0.4071	0.4379	0.2848	0.3382	0.2131
	Dark-ISP	0.7139	0.4663	0.8940	0.7550	0.8840	0.6890	0.5560	0.3630	0.4090	0.2710	0.3240	0.2060
	Dr.RAW	0.8865	0.7305	0.8983	0.7639	0.8982	0.7624	0.6408	0.4399	0.4720	0.3160	0.3630	0.2310
	RAM	0.8980	0.7310	0.9040	0.7580	0.9020	0.7500	0.6660	0.4470	0.4550	0.3040	0.3650	0.2360
	Ours	**0.8911**	**0.7332**	**0.9019**	**0.7690**	**0.9020**	**0.7690**	**0.6909**	**0.4363**	**0.5541**	**0.3752**	**0.3623**	**0.2311**
Swin Transformer	Linear-RAW	0.5517	0.3336	0.8707	0.6882	0.8733	0.6196	0.5279	0.2828	0.3251	0.2120	0.1241	0.0768
	Default-ISP	0.8581	0.7156	0.9023	0.7769	0.8966	0.7410	0.6978	0.5015	0.4885	0.3119	0.3428	0.2308
	ReconfigISP	0.8803	0.6999	0.8946	0.7165	0.9017	0.7618	0.5935	0.3746	0.4051	0.2527	0.2977	0.1966
	RAW-Adapter	0.8727	0.7121	0.9017	0.7869	0.8967	0.7546	0.6289	0.4149	0.4386	0.2869	0.3035	0.1985
	AdaptiveISP	0.7830	0.5516	0.8704	0.7239	0.8883	0.7308	0.5990	0.4581	0.4440	0.2711	0.3661	0.2471
	Dark-ISP	0.7233	0.4852	0.9010	0.7672	0.8980	0.7556	0.6303	0.4146	0.5039	0.3273	0.3398	0.2233
	Dr.RAW	0.8814	0.7183	0.8986	0.7591	0.8958	0.7190	0.6950	0.5266	0.4933	0.3192	0.3268	0.2160
	Ours	**0.9083**	**0.7406**	**0.9342**	**0.7792**	**0.9313**	**0.7921**	**0.6986**	**0.5057**	**0.5191**	**0.3467**	**0.4394**	**0.3263**

Table 2: Evaluation on synthetic and real mixed-sensor datasets.

	PAS (Syn)		LOD (Syn)		ROD (Syn)		PAS & LOD		MultiRAW	
Method	@50	@75	@50	@75	@50	@75	@50	@75	@50	@75
Linear-RAW	0.8921	0.7295	0.5310	0.3297	0.3497	0.2273	0.6057	0.4175	0.2539	0.1227
RAW-Adapter	0.8869	0.7233	0.5874	0.3945	0.3637	0.2397	0.6540	0.4370	0.2334	0.1025
Dr.RAW	0.8949	0.7324	0.5924	0.4164	0.4234	0.2821	0.6660	0.4750	0.2572	0.1195
Dark-ISP	0.8888	0.7334	0.5837	0.3390	0.3703	0.2403	0.6420	0.4360	0.2603	0.1063
Ours	**0.8953**	**0.7343**	**0.6646**	**0.4602**	**0.4291**	**0.3223**	**0.6865**	**0.4896**	**0.3528**	**0.1336**

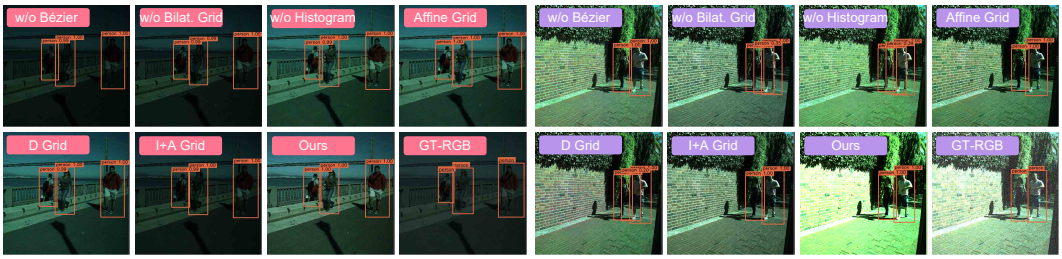
Table 3: Efficiency.

Inference	Time&Mem.	
Method	ms	GB
Linear-RAW	7.20	0.23
RAW-Adapter	10.00	0.24
Dr.RAW	10.52	0.53
Dark-ISP	20.45	0.99
Ours	**10.70**	**0.24**

Metric. We report object detection accuracy in mAP@50 and mAP@75. Best results in experiments are shaded as first, second, and third. Linear-RAW denotes direct integration of RAW images.

4.1 Evaluation on Per-Sensor Benchmarks

Table 1 reports mAP@50 and mAP@75 across the six per-sensor benchmarks under both a ResNet-50 and a Swin Transformer backbone. RAWild ranks first in 16 of the 24 benchmark–metric pairs, one of them tied with RAM. RAM leads five of the remaining eight pairs, all with ResNet-50, and AdaptiveISP, RAW-Adapter, and Dr.RAW lead one each. The improvement is more pronounced beyond the conventional 12-bit setting; on ROD and LOD, ResNet-50 mAP@50 improves by 17.4% and 3.7% over the strongest baseline approaches. Against RAM [24] under our evaluation setting, RAWild achieves higher mAP@50 on LOD and ROD, with gains of 2.49 and 9.91 points. Table 11 in Appendix D.1 reports the RAM evaluation setting. RAWild is higher in all three metrics on LOD-Normal, both NOD [57] sensors, and both ROD splits, and in mAP@75 on PASCALRAW. RAM is higher on LOD-Dark and in mAP and mAP@50 on PASCALRAW. Across the three PASCALRAW exposure variants constructed by Cui and Harada [18], the model recovers nearly identical mAP@50 on LOW, NM, and OE, suggesting that the bounded global curve absorbs exposure variation rather than overfitting to a particular setting. Figure 5 shows fewer missed and mislocalized boxes under saturated highlights and severely underexposed regions. We further compare with the Neural-ISP methods ParamISP [43] and InvISP [99] under the same evaluation protocol in Table 10 of



(a) Low-light (b) Over-exposure

Figure 6: Visualization of ablation studies on PASCALRAW lowlight and overexposure variants. Each case presents transformed images of design choices shown in Table 4, Table 5, and Table 6.

Table 4: Ablation of histogram conditioning mechanisms.

Ablation	@50	@75
No Histogram	0.6838	0.4241
Prefix Condition	0.6397	0.4123
AdaLN Condition	**0.6909**	**0.4363**

Table 5: Ablation study of global Bézier curve design.

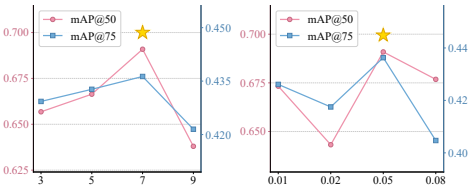
Ablation	@50	@75
No Bézier	0.6650	0.4133
Vanilla Bézier	0.6511	0.4343
Delta Bézier	**0.6909**	**0.4363**

Table 6: Ablation of bilateral grid.

Ablation	@50	@75
No Bilateral Grid	0.6691	0.4346
Vanilla Bilateral Grid	0.6673	0.4149
D	0.6589	**0.4461**
I+A	0.6646	0.4118
D(I+A)	**0.6909**	0.4363

Table 7: Cross-sensor ablation on MultiRAW with a ResNet-50 backbone. Per-sensor columns report mAP@50. †OnePlus 5T appears only in the test split.

	All sensors		Per-sensor mAP@50			
Ablation	@50	@75	ASI	Huawei	iPhone	OnePlus†
No Bézier	0.2670	0.1170	0.2160	0.3190	0.2740	0.2240
No Bilateral Grid	0.2870	0.1280	0.2320	0.3400	0.3000	0.2460
D	0.2990	0.1250	0.2500	**0.3490**	0.3090	0.2460
No Histogram	0.2910	0.1180	0.2390	**0.3490**	0.2950	0.2530
Full model	**0.3528**	**0.1336**	**0.3493**	0.3438	**0.4378**	**0.2822**



(a) Points in Bézier curve (b) k in Bilateral Grid

Figure 7: Ablation on grid activation k and the number of Bézier control points. Stars mark the settings used in all experiments.

Appendix D.1. On LOD, RAWild improves over the stronger Neural-ISP baseline by 13.49 mAP@50 and 15.43 mAP@75 points.

4.2 Evaluation on Mixed-Sensor Benchmarks

Table 2 examines how each method behaves when training and evaluation span heterogeneous sensor captures. Our method attains the highest mAP@50 among task-oriented ISPs in both synthetic and real-world settings, with relative gains of 12.2% on LOD (Syn), 3.1% on PASCALRAW & LOD, and 17.2% on MultiRAW over the strongest baselines. For MultiRAW, this is RAM in Table 10. The margin is largest on MultiRAW, which spans the widest range of bit depths and capture devices in our evaluation. On such heterogeneous mixes, previous pipelines tend to fit their parameters to a dominant sensor mode, whereas our histogram-conditioned prediction adapts to per-image statistics through the global Bézier curve and bilateral grid, which generalizes more reliably across sensor configurations. Against the Neural-ISP baselines in Table 10, RAWild achieves the highest mAP@50 on MultiRAW, with gains of 2.78 points over ParamISP and 3.88 points over InvISP, while ParamISP is 1.04 points higher in mAP@75. RAWild also exceeds RAM by 5.18 mAP@50 points on MultiRAW.

5 Ablation Study

We ablate each component on LOD [35], because its low-light captures leave a wide accuracy range for separating the contribution of each design. We repeat the component ablation on MultiRAW to examine it across sensors. All ablations use the ResNet-50 backbone. Figure 6 visualizes the transformed linear images from each ablated module or design.

Effect of Histogram Conditioning. Table 4 compares ways of injecting the input histogram. AdaLN modulation outperforms both removing the histogram and routing it through a prefix token, consistent with the histogram acting as a sensor-level prior that is most useful when it reshapes every block rather than competing with spatial tokens for attention.

Effect of Global Bézier Curve. Table 5 compares forms of the global tone curve. Our delta parameterization outperforms both removing the curve and a vanilla Bézier that emits absolute control points, with the vanilla variant falling slightly below the no-curve baseline in mAP@50. This is consistent with the head benefiting from a uniform anchor rather than having to learn the identity transform from scratch on every input.

Effect of Spatial Bilateral Grid. Table 6 disentangles the gain–mixing decomposition $\mathbf{D}(\mathbf{I}+\mathbf{A})$. The full form gives the best mAP@50, whereas predicting only $\mathbf{D}$ or only $\mathbf{I}+\mathbf{A}$ falls even below the no-grid baseline in mAP@50, suggesting the gain comes from decoupling per-channel exposure and cross-channel mixing rather than from an increase in parameter count; $\mathbf{D}$ alone reaches the highest mAP@75.

Hyperparameters. Figure 7 shows that seven control points and k=0.05 perform best on LOD, and most other tested values stay within 3.5 mAP@50 points of this setting. Nine control points or k=0.02 lower mAP@50 by about five points. We keep these values fixed for all datasets and backbones, including the cross-sensor ablation below.

Cross-Sensor Ablation. Table 7 repeats the component ablation on MultiRAW, which spans four sensors and 10- to 24-bit RAW. OnePlus 5T appears only in the test split and serves as an unseen sensor. All variants share the detector, adapter configuration, training schedule, and seed. Removing the Bézier curve causes the largest drop, 8.58 mAP@50 and 1.66 mAP@75 points, and the full model achieves the best overall accuracy and the best mAP@50 on the unseen OnePlus 5T. On Huawei P30 Pro, the $\mathbf{D}$-only grid and the model without the histogram reach 0.3490 mAP@50, slightly above the 0.3438 of the full model.

6 Conclusion

We address sensor-agnostic RAW object detection, where differences in bit depth, spectral response, and exposure across cameras prevent a single detector from generalizing. We factor the adapter into a global Bézier tone curve and a local bilateral grid of structured $\mathbf{D}(\mathbf{I}+\mathbf{A})$ matrices, both predicted from the input histogram by a shared transformer, and show across per-sensor and mixed-sensor RAW benchmarks that it outperforms previous task-oriented ISPs in most settings at near-constant inference cost.

Limitations. RAWild remains limited by the information captured in global image statistics and by its per-pixel color transformation. Scene appearance may be confounded with sensor variation, while sensor noise, point spread functions, and demosaicing artifacts are not explicitly modeled. As a tone-mapping front end, RAWild cannot recover information lost to saturation, noise, or quantization at capture, and it does not explicitly normalize these acquisition effects in the RAW data. The continued gains in our scaling study in Appendix D.2 also highlight the dependence on scarce labeled RAW data, and developing a label-balanced, multi-sensor RAW detection dataset would be a valuable direction for future research. The simulation has not yet been evaluated against matched random channel-ratio perturbations or paired captures from the target sensors. Further comparison with these controls would provide a more complete assessment of its physical fidelity. Appendix E discusses these limitations in detail.

Acknowledgments and Disclosure of Funding

This work was partially supported by JST Moonshot R&D Grant Number JPMJPS2011 and by the New Energy and Industrial Technology Development Organization (NEDO) under project JPNP26007. Ziteng Cui was supported by JSPS KAKENHI Grant Number JP26K21296. Shuhong Liu was supported by JST BOOST, Japan Grant Number JPMJBS2418.

References

[1] Mahmoud Afifi, Luxi Zhao, Abhijith Punnappurath, Mohamed A Abdelsalam, Ran Zhang, and Michael S Brown. Time-aware auto white balance in mobile photography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5038–5047, 2025.

[2] Kobus Barnard, Lindsay Martin, Adam Coath, and Brian Funt. A comparison of computational color constancy algorithms. II. Experiments with image data. *IEEE Transactions on Image Processing*, 11(9):985–996, 2002.

[3] Chinmay Belthangady and Loic A Royer. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nature Methods*, 16(12):1215–1225, 2019.

[4] Radu Berdan, Beril Besbinar, Christoph Reinders, Junji Otsuka, and Daisuke Iso. ReRAW: RGB-to-RAW image reconstruction via stratified sampling for efficient object detection on the edge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11833–11843, 2025.

[5] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019.

[6] Mark Buckler, Suren Jayasuriya, and Adrian Sampson. Reconfiguring the imaging pipeline for computer vision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 975–984, 2017.

[7] 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*, pages 3291–3300, 2018.

[8] Jiawen Chen, Sylvain Paris, and Frédo Durand. Real-time edge-aware image processing with the bilateral grid. *ACM Transactions on Graphics*, 26(3):103–es, 2007.

[9] Jiawen Chen, Andrew Adams, Neal Wadhwa, and Samuel W Hasinoff. Bilateral guided upsampling. *ACM Transactions on Graphics*, 35:1–8, 2016.

[10] Jiji Chen, Hideki Sasaki, Hoyin Lai, Yijun Su, Jiamin Liu, Yicong Wu, Alexander Zhovmer, Christian A Combs, Ivan Rey-Suarez, Hung-Yu Chang, et al. Three-dimensional residual channel attention networks denoise and sharpen fluorescence microscopy image volumes. *Nature Methods*, 18(6):678–687, 2021.

[11] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open MMLab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[12] 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.

[13] Yihong Chen, Zheng Zhang, Yue Cao, Liwei Wang, Stephen Lin, and Han Hu. RepPoints v2: Verification meets regression for object detection. In *Advances in Neural Information Processing Systems*, volume 33, pages 5621–5631, 2020.

[14] CIE (International Commission on Illumination). Recommendations on uniform colour-spaces-colour difference equations, psychometric colour terms, 1978.

[15] Commission Internationale de l'Eclairage. Method of measuring and specifying colour rendering properties of light sources. Technical Report CIE 013.3-1995, CIE, 1995.

[16] Commission Internationale de l'Eclairage. Colorimetry. Technical Report CIE 015:2018, CIE, 2018.

[17] Marcos V Conde, Javier Vazquez-Corral, Michael S Brown, and Radu Timofte. NILUT: Conditional neural implicit 3D lookup tables for image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1371–1379, 2024.

[18] 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.

[19] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask AET with orthogonal tangent regularity for dark object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2553–2562, 2021.

[20] Ziteng Cui, Jianfei Yang, and Tatsuya Harada. RAW-Adapter: Adapting pre-trained visual model to camera RAW images and a benchmark. *arXiv preprint arXiv:2503.17027*, 2025.

[21] Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Mobile computational photography: A tour. *Annual Review of Vision Science*, 7(1):571–604, 2021.

[22] Steven Diamond, Vincent Sitzmann, Frank Julca-Aguilar, Stephen Boyd, Gordon Wetzstein, and Felix Heide. Dirty pixels: Towards end-to-end image processing and perception. *ACM Transactions on Graphics*, 40(3):1–15, 2021.

[23] Graham D Finlayson. Corrected-moment illuminant estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1904–1911, 2013.

[24] Shani Gamrian, Hila Barel, Feiran Li, Masakazu Yoshimura, and Daisuke Iso. Beyond RGB: Adaptive parallel processing for RAW object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5547–5557, 2025.

[25] Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand. Deep joint demosaicking and denoising. *ACM Transactions on Graphics*, 35(6):1–12, 2016.

[26] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics*, 36 (4):1–12, 2017.

[27] Jiasheng Guo, Xin Gao, Yuxiang Yan, Guanghao Li, and Jian Pu. Dark-ISP: Enhancing RAW image processing for low-light object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9583–9593, 2025.

[28] Yuduo Guo, Hao Zhang, Mingyu Li, Fujiang Yu, Yunjing Wu, Yuhan Hao, Song Huang, Yongming Liang, Xiaojing Lin, Xinyang Li, et al. Deeper detection limits in astronomical imaging using self-supervised spatiotemporal denoising. *Science*, page eady9404, 2026.

[29] Khurram Azeem Hashmi, Goutham Kallempudi, Didier Stricker, and Muhammad Zeshan Afzal. FeatEnhancer: Enhancing hierarchical features for object detection and beyond under low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6725–6735, 2023.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[31] G. E. Healey and R. Kondepudy. Radiometric CCD camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(3):267–276, 1994. doi: 10.1109/34.276126.

[32] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pająk, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, et al. FlexISP: A flexible camera image processing framework. *ACM Transactions on Graphics*, 33(6):1–13, 2014.

[33] Yashar D Hezaveh, Laurence Perreault Levasseur, and Philip J Marshall. Fast automated analysis of strong gravitational lenses with convolutional neural networks. *Nature*, 548(7669): 555–557, 2017.

[34] Mingbo Hong, Shen Cheng, Haibin Huang, Haoqiang Fan, and Shuaicheng Liu. You only look around: Learning illumination-invariant feature for low-light object detection. *Advances in Neural Information Processing Systems*, 37:87136–87158, 2024.

[35] Yang Hong, Kaixuan Wei, Linwei Chen, and Ying Fu. Crafting object detection in very low light. In *Proceedings of the British Machine Vision Conference*, volume 1, page 3, 2021.

[36] 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 *Advances in Neural Information Processing Systems*, 2025.

[37] 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.

[38] Jun Jiang, Dengyu Liu, Jinwei Gu, and Sabine Süsstrunk. What is the space of spectral sensitivity functions for digital color cameras? In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 168–179. IEEE, 2013.

[39] Xin Jin, Pengyi Jiao, Zheng-Peng Duan, Xingchao Yang, Chongyi Li, Chun-Le Guo, and Bo Ren. Lighting every darkness with 3DGS: Fast training and real-time rendering for HDR view synthesis. *Advances in Neural Information Processing Systems*, 37:80191–80219, 2024.

[40] Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In *European Conference on Computer Vision*, pages 429–444. Springer, 2016.

[41] Eric Kee, Adam Pikielny, Kevin Blackburn-Matzen, and Marc Levoy. Removing reflections from RAW photos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 161–171, 2025.

[42] Wontae Kim and Nam Ik Cho. Image-adaptive 3D lookup tables for real-time image enhancement with bilateral grids. In *European Conference on Computer Vision*, pages 91–108. Springer, 2024.

[43] Woohyeok Kim, Geonu Kim, Junyong Lee, Seungyong Lee, Seung-Hwan Baek, and Sunghyun Cho. ParamISP: Learned forward and inverse ISPs using camera parameters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26067–26076, 2024.

[44] Hoang M Le, Brian Price, Scott Cohen, and Michael S Brown. GamutMLP: A lightweight MLP for color loss recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18268–18277, 2023.

[45] Gongzhe Li, Linwei Qiu, Peibei Cao, Fengying Xie, Xiangyang Ji, and Qilin Sun. Real-time scene-adaptive tone mapping for high-dynamic range object detection. In *Advances in Neural Information Processing Systems*, 2025.

[46] Ruikang Li, Yujin Wang, Shiqi Chen, Fan Zhang, Jinwei Gu, and Tianfan Xue. DualDn: Dual-domain denoising via differentiable ISP. In *European Conference on Computer Vision*, pages 160–177. Springer, 2024.

[47] Zhihao Li, Ming Lu, Xu Zhang, Xin Feng, M Salman Asif, and Zhan Ma. Efficient visual computing with camera RAW snapshots. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4684–4701, 2024.

[48] Zhihao Li, Yufei Wang, Alex Kot, and Bihan Wen. From chaos to clarity: 3DGS in the dark. *Advances in Neural Information Processing Systems*, 37:94971–94992, 2024.

[49] Zhong-Yu Li, Xin Jin, Bo-Yuan Sun, Chun-Le Guo, and Ming-Ming Cheng. Towards RAW object detection in diverse conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8859–8868, 2025.

[50] Panjun Liu, Jiacheng Li, Lizhi Wang, Zheng-Jun Zha, and Zhiwei Xiong. MLP embedded inverse tone mapping. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1283–1291, 2024.

[51] Shuhong Liu, Xining Ge, Ziteng Cui, Liuzhuozheng Li, Gengjia Chang, Jun Liu, Ziying Gu, Dong Li, Xuangeng Chu, Lin Gu, and Tatsuya Harada. FluxFlow: Conservative flow-matching for astronomical image super-resolution. *arXiv preprint arXiv:2605.03749*, 2026.

[52] Shuhong Liu, Xining Ge, Ziying Gu, Quanfeng Xu, Lin Gu, Ziteng Cui, Xuangeng Chu, Jun Liu, Dong Li, and Tatsuya Harada. Denoising the deep sky: Physics-based CCD noise formation for astronomical imaging. *arXiv preprint arXiv:2601.23276*, 2026.

[53] Tie Liu, Yuhui Quan, Yingna Su, Yang Guo, Shu Liu, Haisheng Ji, Qi Hao, Yulong Gao, Yuxia Liu, Yikang Wang, et al. Astronomical image denoising by self-supervised deep learning and restoration processes. *Nature Astronomy*, 9(4):608–615, 2025.

[54] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[55] Calvin S McCamy, Harold Marcus, James G Davidson, et al. A color-rendition chart. *Journal of Applied Photographic Engineering*, 2(3):95–99, 1976.

[56] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022.

[57] Igor Morawski, Yu-An Chen, Yu-Sheng Lin, and Winston H. Hsu. NOD: Taking a closer look at detection under extreme low-light conditions with night object detection dataset. In *Proceedings of the British Machine Vision Conference*. British Machine Vision Association, 2021. doi: 10.5244/C.35.317.

[58] Seonghyeon Nam, Abhijith Punnappurath, Marcus A Brubaker, and Michael S Brown. Learning sRGB-to-Raw-RGB de-rendering with content-aware metadata. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17704–17713, 2022.

[59] Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Lucien E Weiss, Onit Alalouf, Tal Naor, Reut Orange, Tomer Michaeli, and Yoav Shechtman. DeepSTORM3D: Dense 3D localization microscopy and PSF design by deep learning. *Nature Methods*, 17(7): 734–740, 2020.

[60] Alex Omid-Zohoor, Christopher Young, David Ta, and Boris Murmann. Toward always-on mobile object detection: Energy versus performance tradeoffs for embedded HOG feature extraction. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(5):1102–1115, 2017.

[61] Chawin Ounkomol, Sharmishtaa Seshamani, Mary M Maleckar, Forrest Collman, and Gregory R Johnson. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature Methods*, 15(11):917–920, 2018.

[62] Hao Ouyang, Zifan Shi, Chenyang Lei, Ka Lung Law, and Qifeng Chen. Neural camera simulators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7700–7709, 2021.

[63] Wei Ouyang, Andrey Aristov, Mickaël Lelek, Xian Hao, and Christophe Zimmer. Deep learning massively accelerates super-resolution localization microscopy. *Nature Biotechnology*, 36(5):460–468, 2018.

[64] Jin-Hwi Park, Chanhwi Jeong, Junoh Lee, and Hae-Gon Jeon. Depth prompting for sensor-agnostic depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9859–9869, 2024.

[65] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[66] 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.

[67] Max Planck. Über das Gesetz der Energieverteilung im Normalspektrum. In *Von Kirchhoff bis Planck: Theorie der Wärmestrahlung in historisch-kritischer Darstellung*, pages 178–191. Springer, 1978.

[68] Irwin G Priest. A proposed scale for use in specifying the chromaticity of incandescent illuminants and various phases of daylight. *Journal of the Optical Society of America*, 23(2): 41–45, 1933.

[69] Abhijith Punnappurath and Michael S Brown. Learning raw image reconstruction-aware deep image compressors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4): 1013–1019, 2019.

[70] Abhijith Punnappurath and Michael S Brown. Spatially aware metadata for raw reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 218–226, 2021.

[71] Abhijith Punnappurath, Abdullah Abuolaim, Abdelrahman Abdelhamed, Alex Levinshtein, and Michael S Brown. Day-to-night image synthesis for training nighttime neural ISPs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10769–10778, 2022.

[72] Abhijith Punnappurath, Luxi Zhao, Hoang Le, Abdelrahman Abdelhamed, SaiKiran Kumar Tedla, and Michael S. Brown. Improved mapping between illuminations and sensors for RAW images. *Journal of the Optical Society of America A*, 42(8):1182–1190, August 2025. doi: 10.1364/JOSAA.561327. URL https://opg.optica.org/josaa/abstract.cfm?URI=josaa-42-8-1182.

[73] Chang Qiao, Di Li, Yuting Guo, Chong Liu, Tao Jiang, Qionghai Dai, and Dong Li. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. *Nature Methods*, 18(2):194–202, 2021.

[74] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020.

[75] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Transactions on Graphics*, 21(3):267–276, 2002. doi: 10.1145/566654.566575.

[76] Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski. *High dynamic range imaging: Acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010.

[77] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[78] SanerYee. Deep understanding Tesla FSD: Photon to control. Medium, December 2023. URL https://saneryee-studio.medium.com/deep-understanding-tesla-fsd-photon-to-control-48577fab8070.

[79] Yukihiro Sasagawa and Hajime Nagahara. YOLO in the dark – domain adaptation method for merging multiple models. In *European Conference on Computer Vision*, pages 345–359. Springer, 2020.

[80] David Serrano-Lozano, Luis Herranz, Michael S Brown, and Javier Vazquez-Corral. NamedCurves: Learned image enhancement via color naming. In *European Conference on Computer Vision*, pages 92–108. Springer, 2024.

[81] Artur Speiser, Lucas-Raphael Müller, Philipp Hoess, Ulf Matti, Christopher J Obara, Wesley R Legant, Anna Kreshuk, Jakob H Macke, Jonas Ries, and Srinivas C Turaga. Deep learning enables fast and dense single-molecule localization with high accuracy. *Nature Methods*, 18 (9):1082–1090, 2021.

[82] Jiayu Sun, Ke Xu, Youwei Pang, Lihe Zhang, Huchuan Lu, Gerhard Hancke, and Rynson Lau. Adaptive illumination mapping for shadow detection in raw images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12709–12718, 2023.

[83] Sabine Süsstrunk, Robert Buckley, and Steve Swen. Standard RGB color spaces. In *Color and Imaging Conference*, volume 7, pages 127–134. Society of Imaging Science and Technology, 1999.

[84] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 839–846, 1998. doi: 10.1109/ICCV.1998.710815.

[85] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[86] Hongda Wang, Yair Rivenson, Yiyin Jin, Zhensong Wei, Ronald Gao, Harun Günaydın, Laurent A Bentolila, Comert Kural, and Aydogan Ozcan. Deep learning enables cross-modality super-resolution in fluorescence microscopy. *Nature Methods*, 16(1):103–110, 2019.

[87] Nan Wang, Lixing Xiao, Yuantao Chen, Weiqing Xiao, Pierre Merriaux, Lei Lei, Ziyang Yan, Saining Zhang, Shaocong Xu, Chongjie Ye, Bohan Li, Zhaoxi Chen, Tianfan Xue, and Hao Zhao. Unifying appearance codes and bilateral grids for driving scene Gaussian splatting. In *Advances in Neural Information Processing Systems*, volume 38, pages 29827–29858, 2025.

[88] Xinzhe Wang, Kang Ma, Qiankun Liu, Yunhao Zou, and Ying Fu. Multi-object tracking in the dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–392, 2024.

[89] Yuehao Wang, Chaoyi Wang, Bingchen Gong, and Tianfan Xue. Bilateral guided radiance field processing. *ACM Transactions on Graphics*, 43(4):1–13, 2024.

[90] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. Raw image reconstruction with learned compact metadata. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18206–18215, 2023.

[91] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. Beyond learned metadata-based raw image reconstruction. *International Journal of Computer Vision*, 132(12):5514–5533, 2024.

[92] Yujin Wang, Tianyi Xu, Fan Zhang, Tianfan Xue, and Jinwei Gu. AdaptiveISP: Learning an adaptive image signal processor for object detection. *Advances in Neural Information Processing Systems*, 37:112598–112623, 2024.

[93] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2758–2767, 2020.

[94] Martin Weigert, Uwe Schmidt, Tobias Boothe, Andreas Müller, Alexandr Dibrov, Akanksha Jain, Benjamin Wilhelm, Deborah Schmidt, Coleman Broaddus, Siân Culley, et al. Content-aware image restoration: Pushing the limits of fluorescence microscopy. *Nature Methods*, 15 (12):1090–1097, 2018.

[95] Qi Wu, Janick Martinez Esturo, Ashkan Mirzaei, Nicolas Moenne-Loccoz, and Zan Gojcic. 3DGUT: Enabling distorted cameras and secondary rays in Gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[96] Günther Wyszecki and Walter Stanley Stiles. *Color science: Concepts and methods, quantitative data and formulae*. John Wiley & Sons, 2000.

[97] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[98] Haiyang Xie, Xi Shen, Shihua Huang, Qirui Wang, and Zheng Wang. SimROD: A simple baseline for raw object detection with global and local enhancements. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 11051–11059, 2026.

[99] Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6287–6296, 2021.

[100] Ruikang Xu, Chang Chen, Jingyang Peng, Cheng Li, Yibin Huang, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Toward RAW object detection: A new benchmark and a new model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13384–13393, 2023.

[101] Masakazu Yoshimura, Junji Otsuka, Atsushi Irie, and Takeshi Ohashi. DynamicISP: Dynamically controlled image signal processor for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12866–12876, 2023.

[102] Ke Yu, Zexian Li, Yue Peng, Chen Change Loy, and Jinwei Gu. ReconfigISP: Reconfigurable camera image processing pipeline. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[103] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. CycleISP: Real image restoration via improved data synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2696–2705, 2020.

[104] Vahid Zehtab, David B Lindell, Marcus A Brubaker, and Michael S Brown. Efficient neural network encoding for 3D color lookup tables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9772–9779, 2025.

[105] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2058–2073, 2022. doi: 10.1109/TPAMI.2020.3026740.

[106] Chen Zhang, Wencheng Han, Yang Zhou, Jianbing Shen, Cheng-zhong Xu, and Wentao Liu. Leveraging frame affinity for sRGB-to-RAW video de-rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25659–25668, 2024.

[107] Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Cruz Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. *arXiv preprint arXiv:2512.07237*, 2025.

[108] Yuxuan Zhang, Bo Dong, and Felix Heide. All you need is RAW: Defending against adversarial attacks with camera image pipelines. In *European Conference on Computer Vision*, pages 323–343. Springer, 2022.

[109] Chenhao Zheng, Ayush Shrivastava, and Andrew Owens. EXIF as language: Learning cross-modal associations between images and camera metadata. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6945–6956, 2023.

[110] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

[111] Simone Zini, Alex Gomez-Villa, Marco Buzzelli, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. Planckian jitter: Countering the color-crippling effects of color jitter on self-supervised training. In *International Conference on Learning Representations*, 2023.

A Simulated RAW Augmentation

To improve the generalization of our detection model across heterogeneous imaging conditions, we construct a physics-based RAW image simulation pipeline that synthesizes plausible sensor outputs under diverse camera spectral sensitivities, illuminants, exposures, and sensor characteristics. Jointly varying these physically meaningful factors follows Planckian Jitter [111], which shows that physically constrained illumination changes preserve color information better than conventional color jitter, and responds to the sensor- and illumination-specific RAW capture that Punnappurath et al. [72] identify as a major data bottleneck. All operations are performed in linear RAW space after black-level subtraction, and the output is a simulated linear RAW image normalized to $[0, 1]$. Table 2 reports detection results on our synthesized dataset, which is built from PASCALRAW [60], LOD [35], and ROD [100]. We exclude AODRAW [49] because it already contains diverse real-world degradations such as adverse weather and challenging illumination.

Camera Spectral Sensitivity Database. We use the 28-camera spectral sensitivity database of Jiang et al. [38], which provides per-channel sensitivity curves $S_c^{(i)}(\lambda)$ for $c \in \{R, G, B\}$ sampled at 33 wavelengths from 400 nm to 720 nm in 10 nm increments. Each curve is peak-normalized to unity. Following Jiang et al. [38], we observe that the space of real camera sensitivities is well approximated by a low-dimensional linear subspace, and fitting a two-component PCA independently for each channel explains most of the inter-camera variance as shown in Figure 8.

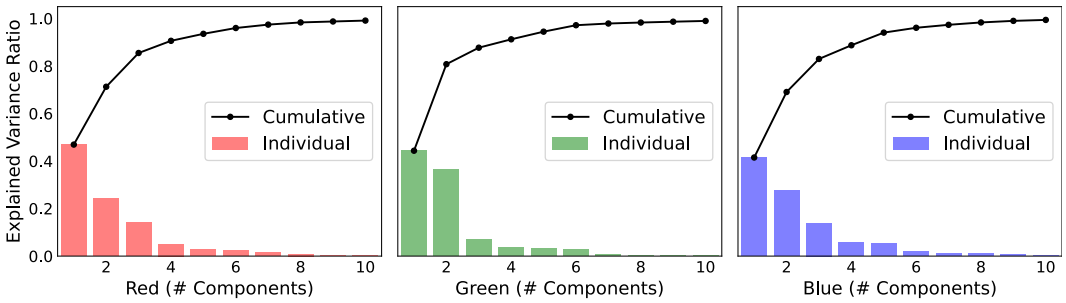


Figure 8: Per-channel PCA explained variance for the 28-camera spectral sensitivity database.

Offline Precomputation. Let $\mathbf{X}^c \in \mathbb{R}^{28\times 33}$ denote the matrix of peak-normalized sensitivity curves for channel c. We fit PCA with $k = 2$ components as:

$$\mathbf{X}^c = \mathbf{1}(\boldsymbol{\mu}^c)^\top + \mathbf{1}(\boldsymbol{\sigma}^c)^\top \odot \mathbf{Z}^c\mathbf{V}^{c\top} + \mathbf{E}^c, \tag{9}$$

where $\mathbf{V}^c \in \mathbb{R}^{33\times 2}$ are the principal components and $\mathbf{Z}^c \in \mathbb{R}^{28\times 2}$ are the corresponding coefficients. We record the per-component empirical mean $\bar{\mathbf{z}}^c$, standard deviation $\hat{\sigma}^c$, and extrema $(\mathbf{z}^c_{\min}, \mathbf{z}^c_{\max})$ from the 28 real cameras.

We also precompute the ColorChecker response of a surrogate source camera. Since the true spectral sensitivity of the source camera is generally unavailable, we approximate it by the mean sensitivity across all 28 cameras:

$$\bar{S}(\lambda) = \frac{1}{28}\sum_{i=1}^{28} S^{(i)}(\lambda). \tag{10}$$

This choice is preferred over the CIE 1931 2° standard observer [16] as the source proxy. As illustrated in Figure 9, the CIE $\bar{z}$ function peaks at 1.77, well above the typical blue-channel response

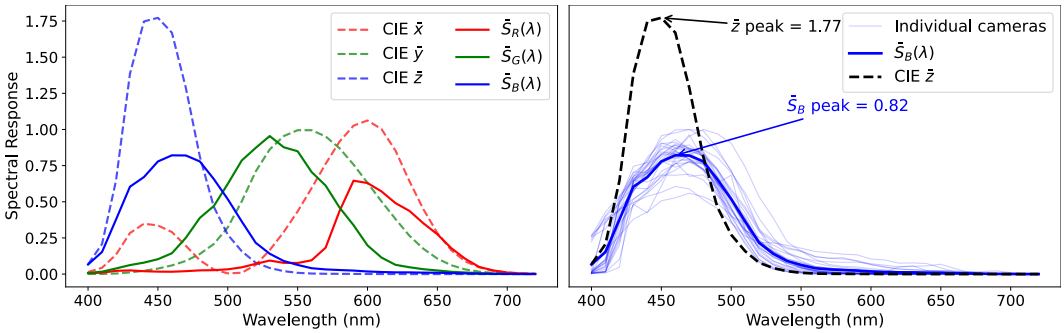


Figure 9: Comparison of the CIE 1931 2° standard observer CMFs (dashed) and the 28-camera mean spectral sensitivity $\bar{S}(\lambda)$ (solid) used as the source camera proxy. The CIE $\bar{z}$ function peaks at 1.77, far above the typical blue-channel response of real cameras (≈0.82), motivating our use of $\bar{S}(\lambda)$ to avoid compressing the blue-channel dynamic range in the cross-camera transform.

of real cameras (≈0.82), which would artificially suppress the blue-channel dynamic range in the estimated cross-camera matrix. Under D65 illumination [15], the ColorChecker response of the source proxy is

$$\mathbf{P}_A = \left(\mathbf{R} \odot \mathbf{d}_{65}\mathbf{1}^\top\right)^\top \bar{\mathbf{S}} \cdot \Delta\lambda \in \mathbb{R}^{24\times 3}, \tag{11}$$

where $\mathbf{R} \in \mathbb{R}^{33\times 24}$ contains the Macbeth ColorChecker reflectance spectra [55], $\mathbf{d}_{65} \in \mathbb{R}^{33}$ is the D65 spectral power distribution, and $\Delta\lambda = 10\,\text{nm}$. $\mathbf{P}_A$ is then normalized so that the green column equals unity.

Exposure Sampling. A global exposure gain α is sampled log-uniformly in stop units:

$$u \sim \mathcal{U}(u_{\min},\, u_{\max}), \qquad \alpha = 2^u, \tag{12}$$

with $u_{\min} = -3$ and $u_{\max} = 3$, spanning a six-stop dynamic range. This log-uniform prior reflects the fact that perceptually equal exposure steps correspond to multiplicative gains [76].



Figure 10: Visualization of RAW images synthesized using different exposure values (EVs).

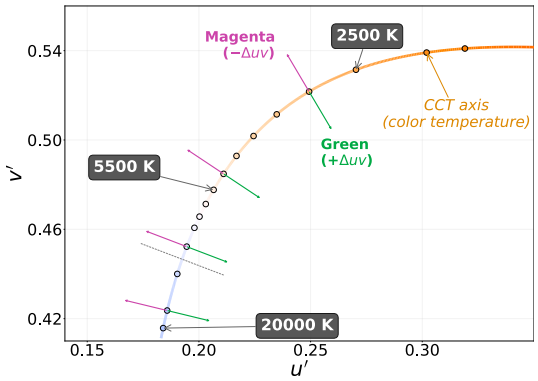


Figure 11: Visualization of Planckian locus on CIE 1976 $u'v'$ chromaticity diagram.

Color Temperature and Tint Sampling. We sample the illuminant color temperature in the mired scale [68], also known as the micro-reciprocal degree scale. This parameterization is commonly used for color-temperature shifts because equal steps in mired produce more uniform perceptual changes than equal steps in Kelvin, especially along the Planckian locus [96]. As shown in Figure 11, we uniformly sample the mired value and convert it to the corresponding color temperature as

$$m \sim \mathcal{U}(m_{\min},\, m_{\max}), \qquad T = \frac{10^6}{m}\,\mathrm{K}, \tag{13}$$

with $m_{\min} = 50$ corresponding to $T = 20{,}000\,\mathrm{K}$ for extreme daylight and $m_{\max} = 400$ corresponding to $T = 2{,}500\,\mathrm{K}$ for tungsten illumination. Given the sampled temperature, we compute the corresponding spectral power distribution using Planck's law of black-body radiation [67],

$$B(\lambda, T) = \frac{2\pi h c^2}{\lambda^5 \left[\exp\left(\dfrac{hc}{k_B T \lambda}\right) - 1\right]}, \tag{14}$$

where h is the Planck constant, c the speed of light, and k_B the Boltzmann constant. The resulting spectrum $I_T(\lambda)$ is peak-normalized to unity. Figure 12 illustrates the spectral shift from warm to cool illuminants across the sampled range.

To account for real-world deviations from the ideal Planckian locus, such as those produced by fluorescent or LED sources, we additionally sample a tint offset $\Delta uv \sim \mathcal{U}(-0.04, 0.04)$ along the green–magenta axis of the CIE $u'v'$ uniform chromaticity diagram [14]. Rather than applying this offset via spectral modification, which requires inverting an overdetermined 33×3 system and yields numerically unstable results, we represent tint as a diagonal RGB gain applied after the cross-camera transform,

$$\mathbf{g}_{\text{tint}} = \left[\max(1 - 0.5 \cdot s \cdot \Delta uv,\ 0.1),\ \ 1.0,\ \ \max(1 - s \cdot \Delta uv,\ 0.1)\right]^{\top}, \tag{15}$$

where $s = 15.0$ is a fixed scaling constant. The asymmetric weighting, where R receives half the correction of B, prevents over-reddening under strong magenta tints, as reflected in the gain curves shown in Figure 13. The perceptual effects of color temperature and tint variation on a real RAW image are illustrated in Figures 14 and 15, where the simulated illuminant shifts from warm tungsten to cool daylight and from green to magenta across the full sampled ranges.

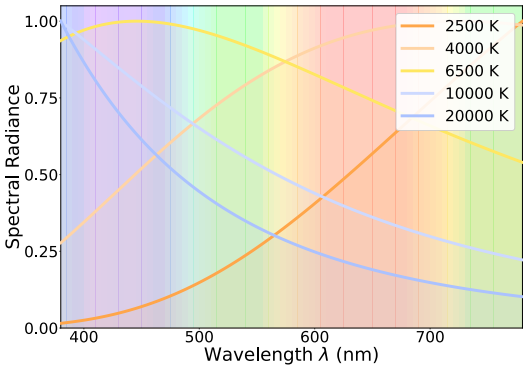


Figure 12: Spectral power distributions $B(\lambda, T)$ of black-body radiators at representative color temperatures spanning our sampling range from 2,500 K to 20,000 K, computed via Planck's law and peak-normalized to unity.

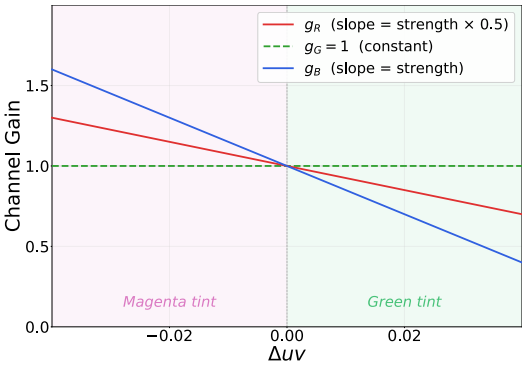


Figure 13: Diagonal RGB gains $\mathbf{g}_{\text{tint}}$ as a function of tint offset Δuv. The R channel receives half the correction of the B channel, preventing over-reddening under strong magenta tints while keeping the G gain fixed at unity.

Target Camera Spectral Response Sampling. A virtual target camera sensitivity $S_B(\lambda) \in \mathbb{R}^{33\times 3}$ is sampled from the learned low-dimensional manifold. For each channel c, we draw PCA coefficients from a truncated Gaussian:

$$\boldsymbol{\zeta}^c \sim \mathcal{N}\big(\bar{\mathbf{z}}^c,\ \mathrm{diag}(\hat{\sigma}^c)^2\big), \quad \text{clipped to } [\mathbf{z}^c_{\min}, \mathbf{z}^c_{\max}]. \tag{16}$$



Figure 14: Visualization of color temperature variation on a real RAW image, simulated across the full sampled range from warm tungsten illumination at 2,500 K to extreme cool daylight at 20,000 K.

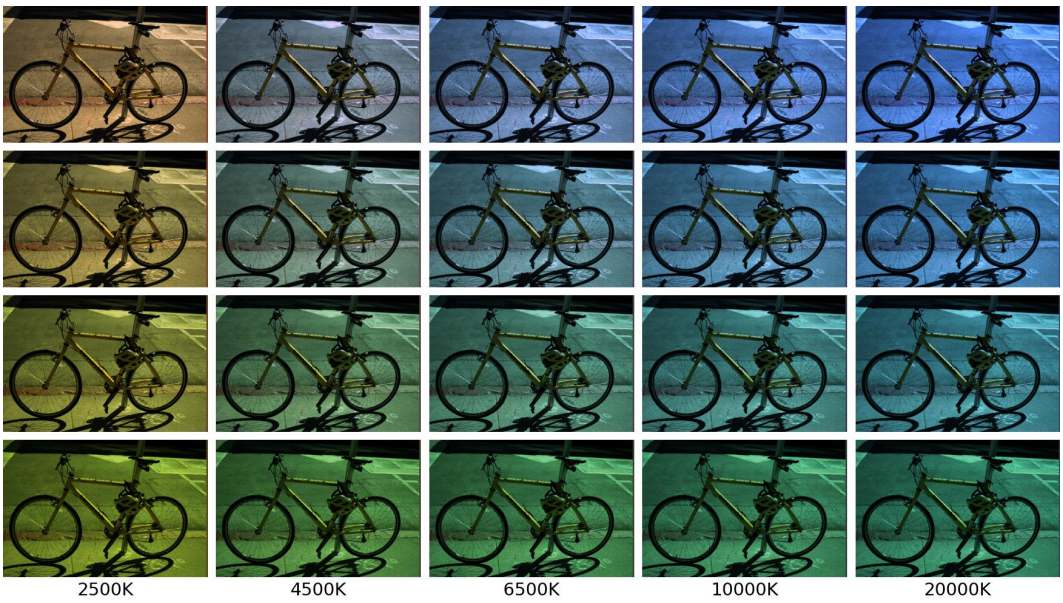


Figure 15: Visualization of temperature and tint variation on a real RAW image, simulated across the full sampled range from strong magenta at $\Delta uv = -0.04$ to strong green at $\Delta uv = +0.04$ along the green–magenta axis of the CIE $u'v'$ chromaticity diagram.

The sensitivity curve is then reconstructed and constrained to be physically realizable:

$$\tilde{S}^c(\lambda) = \boldsymbol{\sigma}^c \odot \mathbf{V}^c \boldsymbol{\zeta}^c + \boldsymbol{\mu}^c, \qquad S_B^c(\lambda) = \frac{\max(\tilde{S}^c(\lambda), 0)}{\max_\lambda \tilde{S}^c(\lambda) + \epsilon}. \tag{17}$$

The Gaussian prior concentrates samples near the center of the observed camera distribution, while the clipping to $[\mathbf{z}^c_{\min}, \mathbf{z}^c_{\max}]$ ensures the reconstructed curves remain within the per-component range spanned by real cameras. As shown in Figure 16, the synthetic sensitivities faithfully reproduce the shape and spread of the diverse real cameras across all three channels, while providing substantially denser coverage of the plausible sensitivity space.

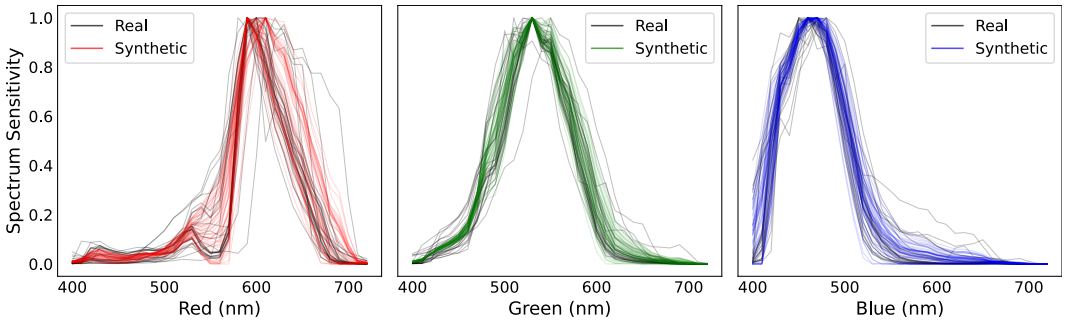


Figure 16: Visualization of spectral sensitivity of real cameras and our synthesized responses.

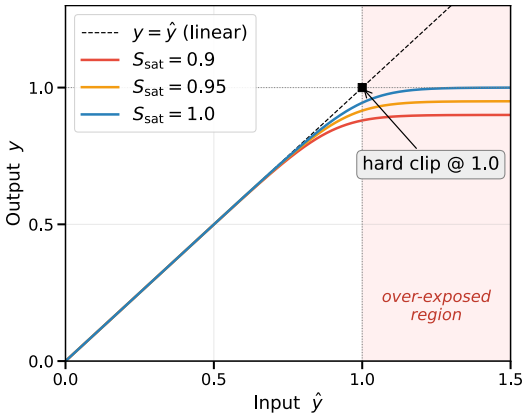


Figure 17: Illustration of the sensor saturation model. For inputs below the saturation threshold S_{sat}, the response follows a smooth compressive curve; beyond this point, the response asymptotically approaches S_{sat} through the tanh roll-off. Lower S_{sat} values produce stronger compression and an earlier onset of saturation.

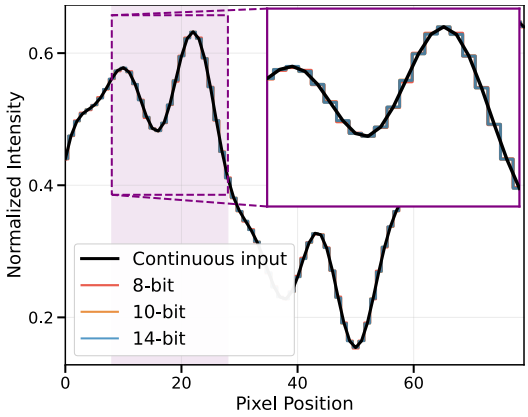


Figure 18: Effect of ADC bit depth on a representative 1D intensity profile. Coarser quantization (8-bit) introduces visible staircase artifacts relative to the continuous reference, while finer quantization (10-bit and 14-bit) progressively recovers the smooth signal shape.

Cross-camera Color Transform. We compute a 3×3 linear matrix $\mathbf{M}_{A \to B,T}$ that jointly accounts for the change in camera spectral sensitivity and the change in illuminant from D65 to I_T. The ColorChecker response of the target camera under illuminant I_T is

$$\mathbf{P}_B = \left(\mathbf{R} \odot I_T \mathbf{1}^\top\right)^\top \mathbf{S}_B \cdot \Delta\lambda \in \mathbb{R}^{24\times 3}, \tag{18}$$

normalized so that the green column equals unity. The cross-camera matrix is estimated by least-squares regression over the 24 ColorChecker patches:

$$\mathbf{M}_{A\to B,T} = \arg\min_{\mathbf{M}} \|\mathbf{P}_A \mathbf{M}^\top - \mathbf{P}_B\|_F^2. \tag{19}$$

This bridge approach, which uses a reflectance chart to relate two cameras under different illuminants, follows the methodology established in camera color characterization [2, 23].

Overall Transformation. Having sampled the exposure α, the cross-camera matrix $\mathbf{M}_{A\to B,T}$, and the tint gain $\mathbf{g}_{\text{tint}}$, we further model sensor non-idealities via a channel-crosstalk matrix $\boldsymbol{\varepsilon}$, with off-diagonal entries drawn from $\mathcal{U}(-0.05, 0.05)$ and diagonal set to zero, and per-channel black-level offsets $\boldsymbol{\beta} \sim \mathcal{U}(-0.02, 0.02)^3$. Before applying the transform, we remove the inherent green bias of Bayer RAW images via channel-wise white balance:

$$\mathbf{g}_{\text{wb}} = \bar{c}_G / \bar{\mathbf{c}}, \qquad \mathbf{x}_{\text{bal}} = \mathbf{x} \odot \mathbf{g}_{\text{wb}}, \tag{20}$$

where $\bar{\mathbf{c}}$ denotes per-channel spatial means. The full transform matrix is then assembled as

$$\mathbf{M}_{\text{full}} = \alpha\big(\text{diag}(\mathbf{g}_{\text{tint}})\, \mathbf{M}_{A\to B,T} + \boldsymbol{\varepsilon}\big), \tag{21}$$

and applied pixel-wise:

$$\mathbf{y} = \mathbf{M}_{\text{full}}\, \mathbf{x}_{\text{bal}} + \boldsymbol{\beta}. \tag{22}$$

The white balance is then reversed using target camera B's own gray-world multipliers $\mathbf{g}_{\text{wb}}^B = r_{B,G}/\mathbf{r}_B$ with $\mathbf{r}_B = \mathbf{S}_B^\top I_T \cdot \Delta\lambda$, rather than the source camera's $\mathbf{g}_{\text{wb}}$, so that the output retains target camera B's natural green dominance under illuminant I_T rather than re-imprinting the source camera's spectral bias:

$$\hat{\mathbf{y}} = \mathbf{y} \oslash \mathbf{g}_{\text{wb}}^B. \tag{23}$$

Highlight Roll-off and Quantization. To simulate the soft saturation behavior of real sensors approaching full-well capacity [31], we apply a tanh-based roll-off:

$$\hat{\mathbf{y}}' = S_{\text{sat}} \tanh\left(\frac{\hat{\mathbf{y}}}{S_{\text{sat}}}\right), \quad S_{\text{sat}} \sim \mathcal{U}(0.9, 1.0), \tag{24}$$

where the $\tanh$ function provides a smooth asymptotic upper bound that approaches but never exceeds S_{sat}, while any sub-zero values arising from the black-level offset $\boldsymbol{\beta}$ are clipped to zero, as illustrated in Figure 17. Finally, the output is quantized to a bit depth $b \in \{10, 12, 14, 16\}$ chosen uniformly at random:

$$\tilde{\mathbf{y}} = \frac{1}{2^b - 1} \left\lfloor \frac{\hat{\mathbf{y}}'}{S_{\text{sat}}} (2^b - 1) + 0.5 \right\rfloor, \tag{25}$$

where the floor operation is applied element-wise. The effect of bit depth on signal fidelity is shown in Figure 18. Since every operation is a pointwise color transformation with no spatial warping, ground-truth bounding box annotations remain valid without modification.

Scope. This simulation is designed to create structured sensor and illumination shifts for cross-sensor evaluation. It begins with captured linear RAW images and uses a mean source sensitivity proxy with a global 3×3 mapping. It does not recover scene radiance, scene spectra, or exact target sensor responses. The synthetic benchmarks evaluate detection robustness under broader structured shifts, while the lack of rigorous evaluation of simulation fidelity remains a limitation.

B Additional Related Works

RAW Imaging Pipelines for High-Level Vision. Early reconfigurable imaging pipelines [6] demonstrate that bypassing most ISP stages preserves the accuracy of CNN-based vision while reducing sensor power, motivating the use of RAW data as a direct input to perception networks. Subsequent works learn end-to-end pipelines that fuse a lightweight neural front end with a downstream classifier or detector [20, 22, 101]. Beyond accuracy, RAW data also provides natural robustness signals, with Zhang et al. [108] mapping RGB inputs through a learned RAW-domain ISP to defend against adversarial perturbations across classification, segmentation, and detection.

Low-Light RAW Object Detection. Low-light scenarios drive a substantial body of RAW-specific detection research. Sasagawa and Nagahara [79] merge a learned dark-image enhancer with YOLO via glue layers and knowledge distillation. Cui et al. [19] introduce a multitask auto-encoding-transformation framework that disentangles ISP and illumination factors from object features. Hashmi et al. [29] learn hierarchical feature enhancers trained jointly with the detector, while Hong et al. [34] extract illumination-invariant kernels grounded in the Lambertian image-formation model. Beyond static images, Wang et al. [88] address multi-object tracking on dark RAW video with adaptive low-pass downsampling and degradation-suppression learning.

RAW Segmentation and Other High-Level Tasks. Beyond object detection, RAW data has been studied for additional high-level tasks. Chen et al. [12] perform instance segmentation directly on RAW low-light captures with feature-noise suppression, smooth convolutions, and a paired LIS benchmark. Sun et al. [82] formulate shadow detection on RAW HDR data through a many-to-many adaptive illumination mapping that preserves multi-scale contrast often lost after ISP tone compression. These efforts show that the linear photon evidence retained in RAW captures benefits dense prediction tasks well beyond bounding-box detection.

sRGB-to-RAW Reconstruction. A complementary research thread reverses the camera pipeline to synthesize realistic RAW from sRGB inputs, enabling RAW-side training without paired sensor captures. The unprocessing framework [5] inverts each ISP step in closed form to render large RGB datasets back to the sensor domain for raw denoiser training at scale. CycleISP [103] introduces a cycle-consistent learned ISP and inverse-ISP pair that produces realistic noisy RAW for image restoration. Invertible ISP [99] replaces the camera pipeline with a normalizing flow whose forward pass renders sRGB and inverse pass recovers near-perfect RAW with no metadata. A parallel line stores compact metadata alongside sRGB to anchor accurate per-pixel reconstruction, evolving from raw-reconstruction-aware sRGB compressors [69] and spatially-aware sample pairs [70] to content-aware metadata [58] and end-to-end learned latent metadata [90, 91], with a recent extension to video sequences [106].

Synthetic RAW for Downstream Tasks. Synthesized RAW data has been widely used to scale learning beyond paired captures. Physics-based noise modeling [93] calibrates sensor shot, read, row, and quantization noise to match real low-light RAW statistics. Neural Camera Simulators [62]

learn a controllable RAW synthesizer conditioned on capture parameters such as exposure and ISO. Day-to-night RAW synthesis [71] relights daytime RAW captures into nighttime variants for training neural ISPs. ReRAW [4] reconstructs sensor-specific RAW from large RGB datasets via stratified sampling for efficient detector pre-training on edge sensors. These inverse-ISP and synthetic-RAW designs are complementary in spirit to our physics-based simulation, while our adapter focuses on a richer global-local tonal representation conditioned on per-image RAW statistics.

RAW for Scientific Imaging. Beyond consumer photography, RAW captures play a central role in scientific instruments where every photon carries quantitative meaning. In fluorescence microscopy, content-aware image restoration [94] pioneered the use of deep networks on raw microscope volumes for denoising and isotropy recovery, followed by label-free virtual fluorescence prediction from transmitted light [61], cross-modality super-resolution [86], accelerated single-molecule localization microscopy [59, 63, 81], Fourier-channel attention super-resolution [73], and three-dimensional residual channel attention denoising [10], all surveyed in [3]. In astronomy, deep networks operate directly on raw CCD frames for galaxy-scale gravitational lensing analysis [33], self-supervised denoising of ground-based images [53], and self-supervised spatiotemporal denoising that pushes deeper detection limits [28]. Closer to our setting, physics-based CCD noise modeling tailored to deep-sky observation [52] and conservative flow-matching for ground-to-space astronomical super-resolution [51] demonstrate that the linear photon evidence retained in scientific RAW captures supports both restoration and downstream measurement. These developments underline that sensor-agnostic RAW modeling, the focus of our work, is increasingly relevant across imaging domains far beyond ordinary RGB photography.

C Experimental Protocols

C.1 MultiRAW Dataset

MultiRAW [47] is a multi-device RAW detection benchmark of 7,208 images captured by four sensors, ASI 294MC Pro, Huawei P30 Pro, iPhone XS Max, and OnePlus 5T, spanning heterogeneous CFA patterns, resolutions, and bit depths from 10 to 24 bits, with per-sensor ISP renderings and XML annotations. Its 11-class taxonomy is long-tailed and device-coupled, with overlapping categories such as the merged `bicycle`, `motorcycle`, and `tricycle`. OnePlus 5T appears only in the test split, and the Huawei P30 Pro annotations are shifted from the objects in the RAW images.

We therefore re-associate RAW images with their metadata, correct the label misalignment, keep the original split, and select the five classes shared by all four sensors, namely `car`, `traffic-sign`, `traffic-light`, `person`, and `bicycle`, with class-aware balancing toward a target of 3,500 active annotated instances per class. The resulting split contains 1,991 unique training and 2,064 unique test RAW images, whose sensor composition is listed in Table 8. Class-aware expansion produces 6,197 training-manifest entries, and all methods in Table 2 use the same manifest. The training and test splits share no images, and OnePlus 5T, which is absent from training, provides an unseen-sensor evaluation.

Table 8: Sensor composition of the MultiRAW split used in Table 2.

Sensor	Unique train	Train share	Balanced manifest entries	Unique test	Test share
Huawei P30 Pro	842	42.29%	3,524	834	40.4%
ASI 294MC Pro	817	41.03%	2,341	823	39.9%
iPhone XS Max	332	16.68%	332	322	15.6%
OnePlus 5T	0	0%	0	85	4.1%
Total	1,991	100%	6,197	2,064	100%

C.2 Baseline Protocols

Default-ISP. The Default-ISP baseline is dataset-dependent. For PASCALRAW-NM, LOD, and AODRAW, it uses the official paired display images for training and evaluation: the Nikon D3200 JPEG provided by PASCALRAW, the short-exposure `RGB-dark` sRGB images provided by LOD,

and the paired camera-generated sRGB JPEG provided by AODRAW, respectively. The AODRAW images are downsampled by a factor of three using the official bilinear preprocessing.

PASCALRAW-LOW/OE are synthetic low-light and overexposure variants constructed following RAW-Adapter [18]. We first normalize the 12-bit camera RGB by 4095, then apply fixed white-balance gains $[2.4492, 1, 1.0508]$ and the fixed camera-RGB-to-linear-sRGB matrix

$$\begin{bmatrix} 1.8517 & -0.7916 & -0.0601 \\ -0.1562 & 1.6474 & -0.4912 \\ 0.0178 & -0.5437 & 1.5259 \end{bmatrix}. \tag{26}$$

Before clipping, pixels with two saturated input channels are blended toward an achromatic highlight with strength 0.8, while pixels with all three channels saturated are fully neutralized. The result is then clipped to $[0, 1]$, gamma-corrected with exponent $1/2.2$, and quantized to 8 bits. No automatic exposure adjustment is applied.

The public ROD release does not provide corresponding SDR images. We therefore use a reference software ISP: 24-bit normalization, per-channel black-level estimation using the 0.1 percentile, median-based gray-world gains clipped to $[0.25, 4]$, an exposure adjustment that maps the 99.5th-percentile luminance to 0.65, ACES tone mapping, sRGB encoding, a saturation factor of 1.05, and 8-bit quantization.

Neural-ISP. We train ParamISP [43] and InvISP [99] with their public implementations on the training images of each dataset and evaluate them with the same protocol as RAWild (Table 10). Both ISPs receive the same packed linear RAW input as the other methods, so ParamISP skips its Bayer demosaicing stage. Each ISP is trained from scratch with the same reconstruction budget of 100 epochs of 1,024 randomly sampled paired patches, and its checkpoint is selected by the paired reconstruction loss on validation pairs. The ISP renderings are then used to train a RetinaNet-R50 detector for 20 epochs at batch size 4, and we report the epoch with the highest mean of mAP@50 and mAP@75. ParamISP additionally takes camera parameters from EXIF; on MultiRAW, the missing values of the ASI camera, which records no optical EXIF, are imputed with training-set medians.

RAM. We port RAM [24] into our MMDetection RetinaNet-R50 pipeline and train it for 20 epochs under our evaluation setting, with the same packed RAW input as the other methods. For each dataset in Tables 1 and 10, we report the best observed value of each metric over the 20 epochs.

For Table 11, we follow the main-table setting of RAM: a Faster R-CNN detector [77] with a ResNet-18 backbone and instance normalization, trained from scratch for 100 epochs with SGD (learning rate 0.02, momentum 0.9, weight decay 10^{-4}), a 500-iteration linear warmup, and tenfold learning-rate decays at epochs 29, 49, and 79. Training uses random resizing and horizontal flipping. We use the dataset splits and annotations released with RAM for PASCALRAW, LOD, NOD [57], and ROD, and report COCO-style mAP, averaged over IoU thresholds from 0.50 to 0.95, together with mAP@50 and mAP@75. The RAM results in Table 11 come from training RAM from scratch in this setting, and we verify our port of the RAM module by loading the official Faster R-CNN ResNet-50 checkpoint released for ROD. For both methods, we report the epoch with the highest mAP.

C.3 Computational Cost

Table 9 complements Table 3 with front-end parameters and FLOPs at the nominal input resolutions 667×400 and 1333×800. Parameters and FLOPs cover the method-specific front end, while the RetinaNet detector is shared by all methods and excluded. FLOPs are counted with FVCore, with one fused multiply-add counted as one FLOP. Inference time and GPU memory cover the full detector. They are measured on an NVIDIA H100 with a batch size of one on the LOD test images, which the nominal 667×400 resolution resizes to 600×400. GPU memory is the memory occupied by the full detector during inference, including its weights and intermediate activations but not the CUDA context, averaged over the whole inference run. RAWild maintains low GFLOPs because coefficient prediction operates on a token grid downsampled by a factor of 16, while the full-resolution computation only involves grid slicing and a 3×3 matrix application. Dark-ISP has fewer parameters, but its convolutional operations are repeatedly applied to high-resolution feature maps, resulting in substantially higher FLOPs and latency.

Table 9: Computational cost of the method-specific front ends with a ResNet-50 RetinaNet detector. Shading ranks the learned front ends in each column, and bold marks the best value.

Method	Front-end params	GFLOPs 667×400	GFLOPs 1333×800	Time (ms)	Memory (GB)
Linear-RAW	0	0.000	0.000	7.20	0.23
RAW-Adapter [18]	0.575M	3.173	**12.203**	**10.00**	**0.24**
Dr.RAW [36]	3.220M	32.211	110.088	10.52	0.53
Dark-ISP [27]	**0.269M**	35.458	136.377	20.45	0.99
Ours	1.101M	**1.940**	14.145	10.70	**0.24**

D Additional Results

D.1 Comparison with Additional Baselines

Neural-ISP Methods and RAM. Tables 10 and 11 give the results discussed in Section 4, under our evaluation setting and the RAM evaluation setting. Appendix C.2 gives the protocols.

Table 10: Comparison with additional baselines on LOD and MultiRAW with a ResNet-50 backbone under our evaluation setting. For RAM, we report the best observed value of each metric over 20 epochs.

Method	LOD @50	LOD @75	MultiRAW @50	MultiRAW @75
RAM [24]	0.6660	0.4470	0.3010	0.1300
ParamISP [43]	0.5400	0.2820	0.3250	0.1440
InvISP [99]	0.5560	0.2820	0.3140	0.1290
Ours	**0.6909**	**0.4363**	**0.3528**	**0.1336**

Table 11: Comparison with RAM [24] under the RAM evaluation setting (Appendix C.2). Values are in percent.

Dataset	RAM mAP	RAM @50	RAM @75	**Ours** mAP	**Ours** @50	**Ours** @75
PASCALRAW	**65.7**	**92.7**	77.7	65.3	92.1	**79.3**
LOD-Dark	**25.9**	**49.4**	**24.8**	24.9	47.0	23.0
LOD-Normal	30.2	53.4	30.4	**31.3**	**53.9**	**32.1**
NOD-Nikon	31.2	55.8	30.7	**31.8**	**56.3**	**31.9**
NOD-Sony	32.2	58.3	34.0	**34.2**	**59.2**	**34.9**
ROD-Day	25.4	41.1	27.4	**26.8**	**43.6**	**28.4**
ROD-Night	33.6	56.1	34.9	**35.2**	**60.8**	**35.8**

Conventional Tone Mapping. A piecewise-linear (PWL) tone curve followed by a global 3D LUT computes

$$y(p) = L\big(T(I(p))\big), \tag{27}$$

where the curve T and the lookup table L are the same at every pixel, so $I(p) = I(q)$ implies $y(p) = y(q)$. RAWild computes

$$y(p) = M\big(p, \ell(p) \mid \phi(I)\big)\, g\big(I(p) \mid \phi(I)\big), \tag{28}$$

where $\phi(I)$ denotes the image features and quantile statistics from which the curve and the grid are predicted for each image (Section 3.4). The matrix is indexed by pixel position and luminance, so two pixels with the same RAW value in different grid cells can be mapped to different outputs. A global 3D LUT cannot represent this position-dependent mapping. The global operators also differ: a PWL curve has a piecewise-constant derivative, whereas the Bézier curve of Equation (2) is a smooth Bernstein polynomial whose identity-anchored control points are predicted for each image.

Table 12 replaces the curve, the grid, or both with conventional global operators (a PWL curve and the Reinhard operator [75]) and local operators (a 3D LUT [105] and bilateral filtering [84]) on LOD. All variants use the same RetinaNet-R50 detector, initialization, detection head, input resolution, and training protocol, and only the front-end operators change. The full design achieves the best mAP@50 and mAP@75. It exceeds the strongest two-stage alternatives by 1.49 mAP@50 points (Bézier with a 3D LUT) and 0.83 mAP@75 points (PWL with the Bilateral Grid).

D.2 Training-Set Scaling on LOD

We train on nested, class-stratified subsets of the LOD training set that contain 12.5%, 25%, 50%, and 75% of the images (225, 450, 900, and 1,350 images), where each subset is contained in the next

Table 12: Conventional global and local tone-mapping operators on LOD with a ResNet-50 RetinaNet detector. Rows give the global operator and column groups the local operator; – marks combinations that were not evaluated. With neither operator, the model is Linear-RAW (Table 1). Bézier alone is the No Bilateral Grid variant of Table 6, and the Bilateral Grid alone is the No Bézier variant of Table 5.

	Local operator							
	None		Bilateral filtering		3D LUT		Bilateral Grid	
Global operator	@50	@75	@50	@75	@50	@75	@50	@75
None	0.5710	0.3713	0.5780	0.3410	0.5970	0.3120	0.6650	0.4133
PWL	0.5830	0.3370	–	–	–	–	0.6720	0.4280
Reinhard	0.5940	0.3310	–	–	–	–	0.6720	0.4100
Bézier	0.6691	0.4346	0.6730	0.3890	0.6760	0.4130	**0.6909**	**0.4363**

larger one. Every run uses 72,000 optimizer updates, the same seed, and the same evaluation protocol, and we select the checkpoint with the highest mean of mAP@50 and mAP@75. Table 13 compares RAWild with Dr.RAW [36] and RAW-Adapter [18]. RAWild achieves the highest mAP@50 at every scale and the highest mAP@75 at 12.5%, 25%, and 75%, while Dr.RAW reaches a higher mAP@75 at 50%. For all three methods, both metrics increase steadily with the amount of training data.

Table 13: Detection accuracy on LOD when training on nested subsets of the training set.

	RAW-Adapter		Dr.RAW		Ours	
Training data	@50	@75	@50	@75	@50	@75
12.5% (225)	0.3776	0.1608	0.4030	0.1993	**0.4326**	**0.2122**
25% (450)	0.4638	0.2160	0.4550	0.2197	**0.4725**	**0.2600**
50% (900)	0.5511	0.3021	0.5601	**0.3278**	**0.5781**	0.3120
75% (1,350)	0.6294	0.3150	0.6201	0.3646	**0.6336**	**0.3750**

D.3 Qualitative Results

Object Detection. Figure 19 compares RAWild with Linear-RAW, RAW-Adapter, Dark-ISP, and Dr.RAW on AODRAW [49], LOD [35], and ROD [100], which together cover adverse weather, extreme low light, and 24-bit HDR driving scenes. RAWild gives fewer missed instances in dark regions, fewer false positives under haze and rain, and tighter boxes on small or low-contrast objects in HDR scenes.

Ablation Study. Figure 20 shows the ablation variants of Section 5 on real LOD samples under normal, low-light, and over-exposure conditions. Without the Bézier curve, the global tonal range stays misaligned and detections in dark or saturated areas drift or disappear. The simpler Full affine, $\mathbf{D}$, and $\mathbf{I}+\mathbf{A}$ grids give inconsistent boxes around small or low-contrast objects, and removing the histogram conditioning makes predictions noisier, most visibly under over-exposure. The full model recovers cleaner tonal structure and gives the most consistent detections in all three regimes.

D.4 RAW Semantic Segmentation

Implementation Details. We follow the exposure-curated ADE20K [110] protocol of RAW-Adapter [18], with Low-light (*LOW*), Normal (*NM*), and Over-exposure (*OE*) splits. The network input is an offline-processed 8-bit RGB image derived from RAW data and normalized to $[0, 1]$, following the released protocol. We insert the same histogram-conditioned global-local adapter described in Sections 3.2 to 3.4 into SegFormer with MiT-B0, MiT-B3, and MiT-B5 backbones [97] and keep the decoder unchanged. We compare with Linear-RAW, RAW-Adapter [18], Dr.RAW [36], and Dark-ISP [27], the baselines with compatible segmentation implementations under the same MMSegmentation framework.

Training runs for 200k iterations using AdamW (base learning rate 6×10^{-5}, β_1=0.9, β_2=0.999, weight decay 0.01), with a 1.5k-iteration linear warmup followed by polynomial decay. We use



Figure 19: Additional qualitative object detection results on AODRAW [49], LOD [35], and ROD [100]. The examples highlight adverse weather, severe low-light noise, and simultaneous highlights and shadows, under which our method produces cleaner detections and more stable localization than prior task-oriented ISP baselines.

a global batch size of 16 on a single H100 GPU and apply a $10\times$ learning-rate multiplier to the adapter and decode head. MiT-B0 and MiT-B5 are initialized from the corresponding ImageNet-pretrained SegFormer checkpoints, while MiT-B3 is warm-started from an earlier 120k checkpoint. Data augmentation uses random resize to $(2048, 512)$ with ratio range $[0.5, 2.0]$, random crop to 512×512, and random horizontal flipping. We report mean Intersection-over-Union (mIoU) with single-scale sliding-window inference and no test-time augmentation.

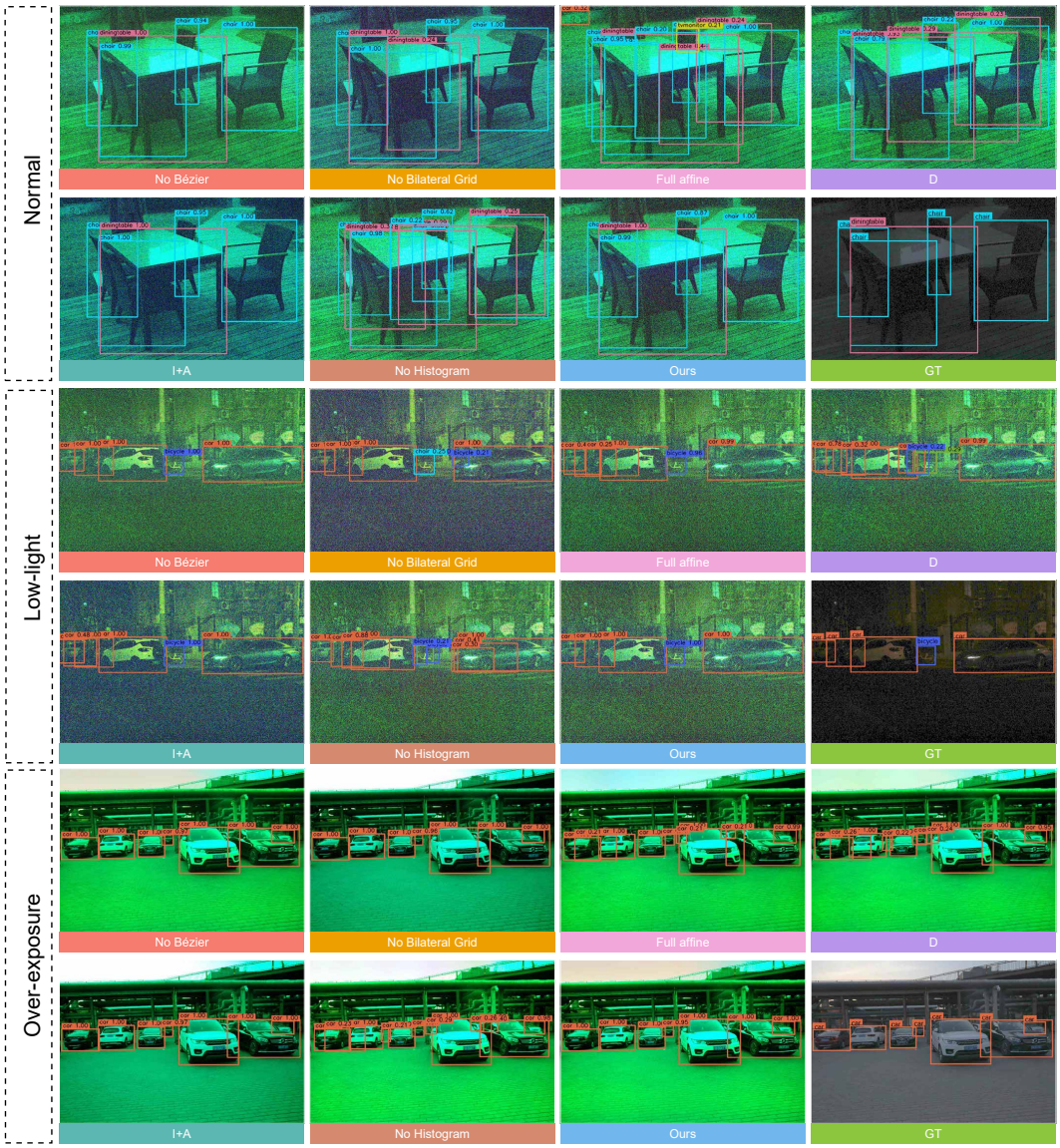


Figure 20: Qualitative ablation on real LOD samples under normal, low-light, and over-exposure conditions. Each block compares *No Bézier*, *No Bilateral Grid*, *Full affine* (the vanilla bilateral grid in Table 6), **D**, **I**+**A**, *No Histogram*, and *Ours* with the ground truth. The full model yields the most stable detections across all three regimes.

Quantitative Results. Table 14 reports mIoU across the three exposure splits and three backbone scales. Our method consistently improves over all baselines in every setting. The gain is largest with the lightweight MiT-B0 backbone, where we surpass the strongest baseline by 6.66, 4.00, and 4.43 mIoU points under *LOW*, *NM*, and *OE*, respectively. The global Bézier curve stabilizes image-level tonal and exposure statistics, while the bilateral grid refines boundary-sensitive local color and contrast, so the two components together improve dense prediction without modifying the segmentation decoder.

Qualitative Results. Figure 21 shows qualitative comparisons. Under low-light inputs, baseline methods often produce fragmented semantic regions or lose small objects in dark areas. Under over-exposed inputs, saturated regions and compressed local contrast make object boundaries difficult to recover. Our method produces more coherent semantic regions and better preserves object boundaries

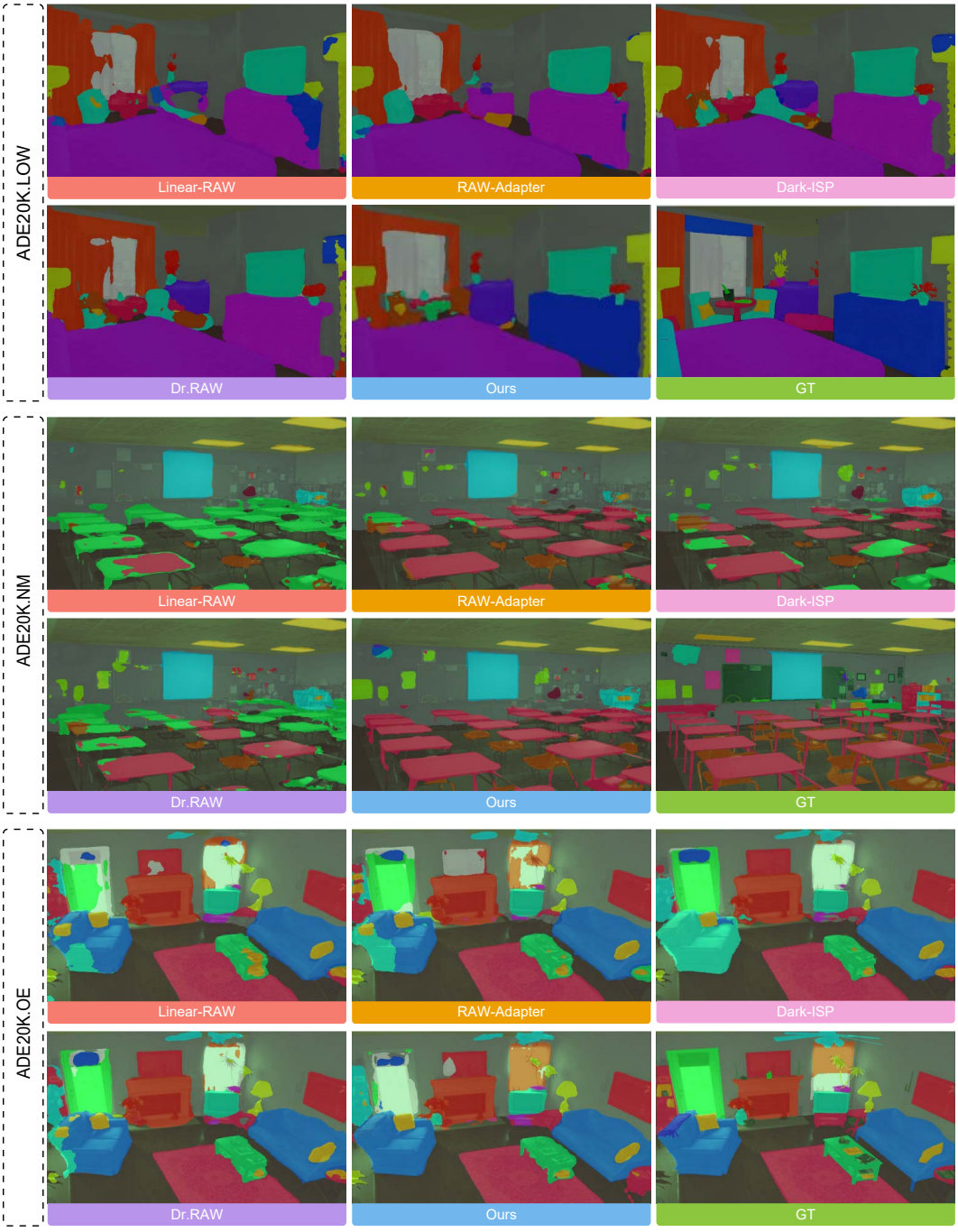


Figure 21: Qualitative transfer results on exposure-curated ADE20K under three illumination regimes: Low-light (*LOW*), Normal (*NM*), and Over-exposure (*OE*). Our method yields more coherent regions and cleaner semantic boundaries across all three settings.

Table 14: Semantic segmentation on exposure-curated ADE20K [110] with MiT-series backbones [97], evaluated by mIoU.

	LOW			NM			OE		
Method	B0	B3	B5	B0	B3	B5	B0	B3	B5
Linear-RAW	0.2027	0.3538	0.3653	0.2632	0.4403	0.4546	0.2691	0.4229	0.4404
RAW-Adapter	0.1885	0.3453	0.3636	0.2939	0.4457	0.4541	0.2683	0.4284	0.4385
Dr.RAW	0.2206	0.3581	0.3821	0.3134	0.4477	0.4666	0.2929	0.4343	0.4475
Dark-ISP	0.2202	0.3559	0.3806	0.3131	0.4432	0.4680	0.2917	0.4283	0.4488
Ours	**0.2872**	**0.3957**	**0.4082**	**0.3534**	**0.4516**	**0.4708**	**0.3372**	**0.4381**	**0.4560**

across *LOW*, *NM*, and *OE*, indicating that the same histogram-conditioned global-local adapter transfers from box prediction to dense prediction under RAW exposure variation.

E Detailed Limitations

Scene and Sensor Statistics. The global quantile descriptor cannot fully disentangle scene-induced from sensor-induced statistics. For example, a genuinely warm-colored scene and a warm illuminant can produce similar global distributions, so the adapter may partially alter true scene color.

Per-Pixel Transformation. The output transformation remains a per-pixel color mapping despite being spatially adaptive. It does not explicitly model sensor noise, point spread functions, or demosaicing artifacts, which provide complementary sources of cross-sensor variation.

Labeled RAW Data. Labeled RAW detection data remain scarce. In our fixed-budget scaling study (Appendix D.2), mAP@50 and mAP@75 increase consistently with training-set size for all three compared methods, suggesting that more labeled RAW data could yield further gains.

Output Range. One limitation of the current output transformation is that some values can fall outside $[0, 1]$ before Equation (6) clips them for input to the downstream backbone. The Bézier output remains within the normalized range, but the Bilateral Grid subsequently applies spatially varying gains and channel mixing. Its diagonal gain can exceed 1 and move some values outside the normalized range, which are then clipped to the endpoints. Validation on LOD shows that only 3.77% of channel elements are affected, and the task-optimized grid makes their impact on detection marginal. However, the potential reduction of tonal detail at the endpoints remains a limitation.

F Assets and Licenses

Table 15 lists the datasets, pretrained models, and libraries used in this work and the assets we release, with the license stated by each source as of November 4, 2026. "Not stated" means that we found no explicit license for the asset at the listed source. For baselines, we use the public implementations of InvISP [99] (MIT) and ParamISP [43] (not stated). Our use of these assets is for non-commercial research.

Table 15: Assets used and released in this work, with their licenses and sources.

Asset	License	Source
Datasets		
PASCALRAW [60] and its exposure variants [18]	CC BY-NC-SA 3.0	purl.stanford.edu/hq050zr7488
LOD [35]	Not stated	github.com/ying-fu/LODDataset
ROD [100]	CC BY-NC 4.0	openi.pcl.ac.cn/innovation_contest/innov202305091731448/datasets
AODRAW [49]	LICENSE file: CC BY-NC 4.0; README (code): CC BY-NC-SA 4.0	github.com/lzyhha/AODRaw
MultiRAW [47]	Not stated	njuvision.github.io/rho-vision
NOD [57]	Repository: MIT; images: not stated	github.com/igor-morawski/RAW-NOD
ADE20K [110] and its exposure variants [18]	Images: non-commercial research; annotations: BSD 3-Clause	ade20k.csail.mit.edu/terms
Spectral sensitivities [38]	CC BY-NC-SA 4.0	www.gujinwei.org/research/camspec/db.html
RAM splits and annotations [24]	MIT	github.com/SonyResearch/RawAdaptationModule
Pretrained models		
RetinaNet R50 (COCO)	Apache 2.0	github.com/open-mmlab/mmdetection
RepPointsV2 Swin-T (COCO)	MIT	github.com/SwinTransformer/storage/releases/tag/v1.0.9
MiT-B0/B3/B5 [97]	NVIDIA Source Code License	github.com/NVlabs/SegFormer
RAM Faster R-CNN R50 (ROD) [24]	MIT	github.com/SonyResearch/RawAdaptationModule
Libraries		
PyTorch 2.1.0	BSD 3-Clause	github.com/pytorch/pytorch
MMDetection 3.3.0	Apache 2.0	github.com/open-mmlab/mmdetection
MMSegmentation 1.2.1	Apache 2.0	github.com/open-mmlab/mmsegmentation
MMCV 2.1.0, MMEngine 0.10.4	Apache 2.0	github.com/open-mmlab/mmcv, github.com/open-mmlab/mmengine
fvcore	Apache 2.0	github.com/facebookresearch/fvcore
timm	Apache 2.0	github.com/huggingface/pytorch-image-models
colour (with CIE and ColorChecker data)	BSD 3-Clause	github.com/colour-science/colour
Released with this work		
Code	MIT; third-party components remain under their licenses	github.com/I2WM/RAWild
Detection checkpoints	CC BY-NC-SA 4.0	github.com/I2WM/RAWild
Segmentation checkpoints	NVIDIA Source Code License	github.com/I2WM/RAWild
Simulated PASCALRAW, ROD	CC BY-NC-SA 4.0	github.com/I2WM/RAWild
Simulated LOD, repackaged MultiRAW	Not stated	github.com/I2WM/RAWild

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the contributions, and Sections 3 to 5 support them with the method description and experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Limitations paragraph in Section 6 summarizes the limitations of the method, and Appendix E discusses them in detail.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 and Appendix C describe the training settings, data splits, and baseline protocols, and Appendix A describes the simulation pipeline.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code, pretrained checkpoints, and simulated data are available at https://github.com/I2WM/RAWild, with instructions for installation, training, and evaluation. All benchmarks used in the paper are publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The implementation details in Section 4 and Appendix C specify the data splits, optimizer, hyperparameters, and baseline protocols.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The tables report point estimates without error bars or statistical significance tests.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C.3 reports inference time and GPU memory measured on an NVIDIA H100 GPU, and Appendix D.4 reports that the segmentation models are trained on a single H100 GPU. Training one detector with the released LOD configuration (20 epochs, batch size 4) on a single NVIDIA Quadro RTX 8000 takes about 4.6 hours with ResNet-50 and about 10 hours with Swin Transformer.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: The paper proposes a generic front end for RAW object detection, evaluated on public benchmarks and not tied to a specific application or deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released models are object detection front ends trained on public benchmarks and pose no high risk of misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original papers of the datasets and baselines. Appendix F summarizes the sources and available licensing information for the datasets, pretrained models, and main libraries used in this work, and marks the assets with no stated license. Our use of these assets is for non-commercial research.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the code under the MIT license at `https://github.com/I2WM/RAWild`, and third-party components remain under their licenses; the README documents installation, training, and evaluation. Appendix F lists the released checkpoints, simulated data, and repackaged data with their licensing information.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used only for writing and formatting and are not part of the method or experiments.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.