

RAWild

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

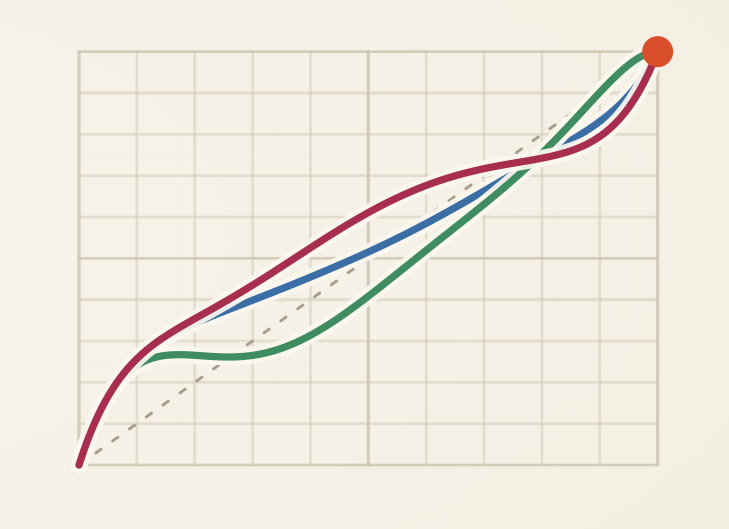
Shuhong Liu^{1,2*} · Gengjia Chang^{2*} · Jun Liu² · Xuangeng Chu^{1,2}
Yinqiang Zheng¹ · Tatsuya Harada^{1,3} · Ziteng Cui^{1†}

1 The University of Tokyo · 2 I2WM · 3 RIKEN

* Equal contribution · † Corresponding author



i2wm.github.io/RAWild



RAWild.

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

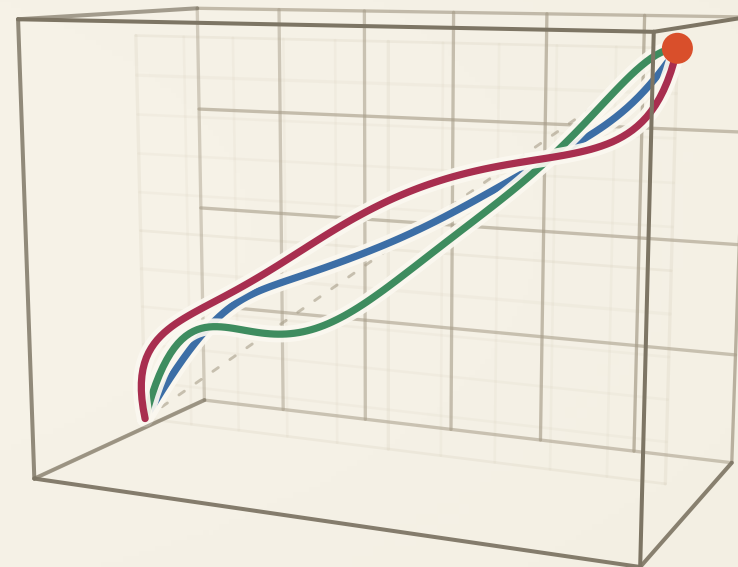
Shuhong Liu^{1,2*} · Gengjia Chang^{2*} · Jun Liu² · Xuangeng Chu^{1,2}
Yinqiang Zheng¹ · Tatsuya Harada^{1,3} · Ziteng Cui^{1†}

1 The University of Tokyo · 2 I2WM · 3 RIKEN

* Equal contribution · † Corresponding author

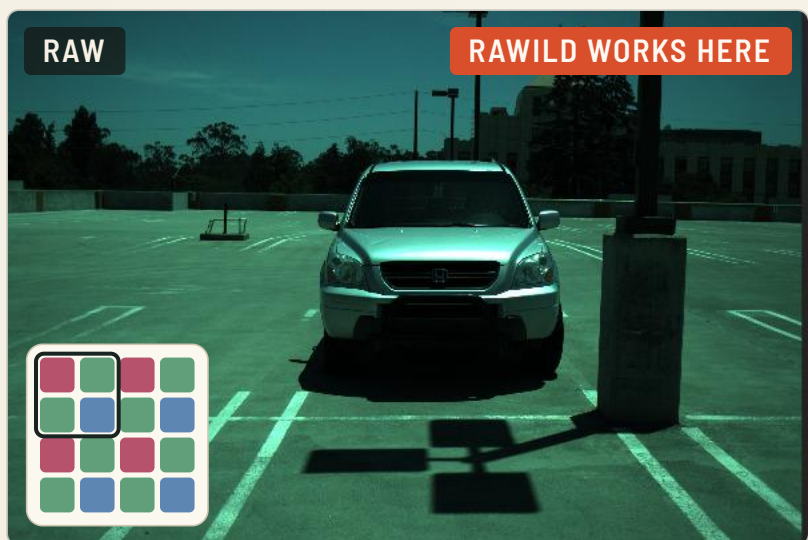


i2wm.github.io/RAWild



What is Camera RAW?

Cameras turn RAW into 8-bit sRGB for the eye; detecting on RAW avoids that lossy step.



Linear · 12-bit · one color per pixel

Linear

Values grow with the light.

Camera ISP

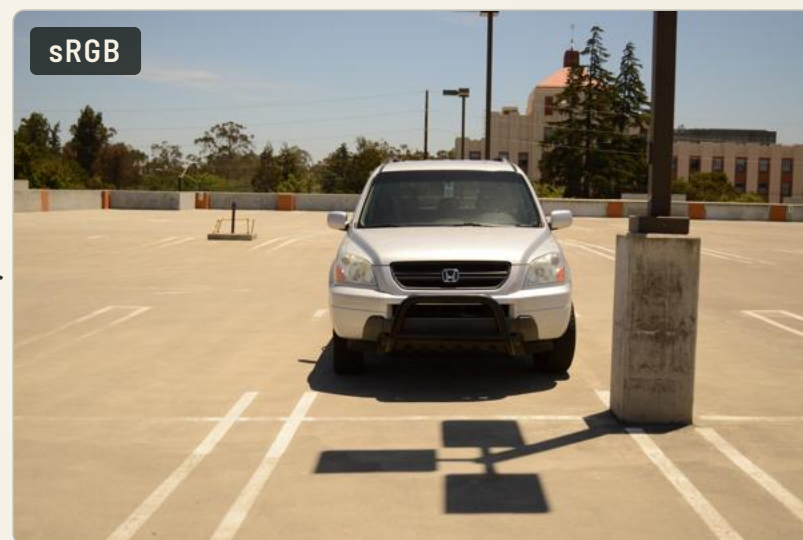
Denoise

Demosaic

White balance

Tone curve

Quantize to 8 bits



Tone-mapped · 8-bit · made for the eye

No ISP loss

Not yet tone-mapped or cut to 8 bits.

Different RAW distributions across sensors

Same task, four cameras from 12 to 24 bits: their RAW pixel distributions differ sharply.

12 bit

Nikon D3200



14 bit

Sony α7 III



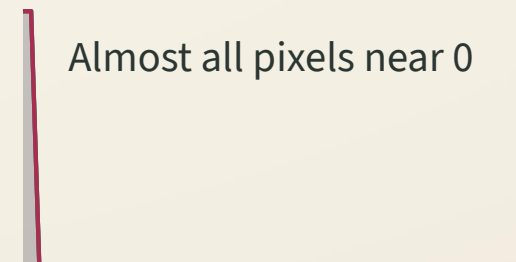
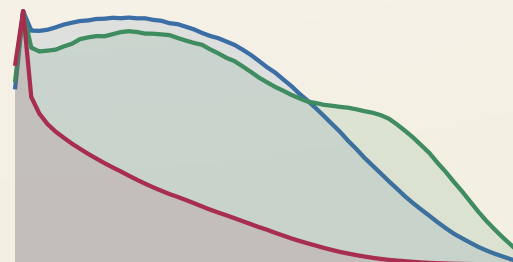
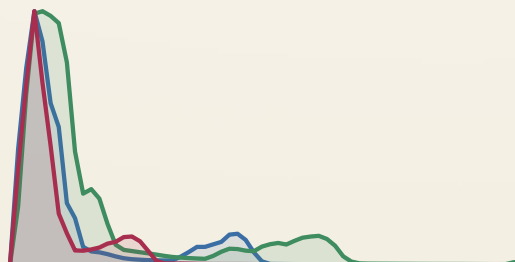
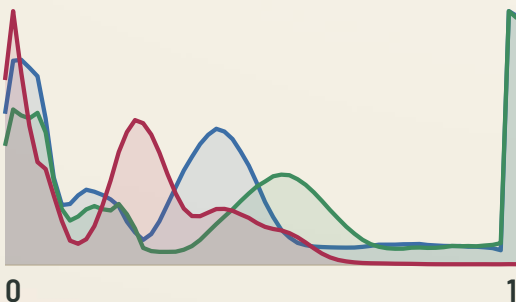
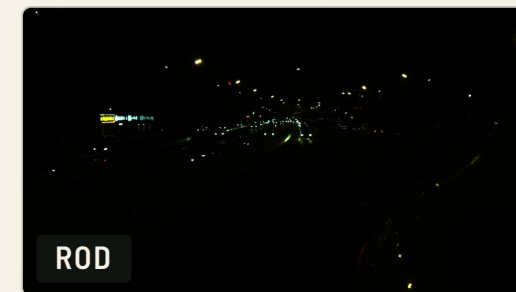
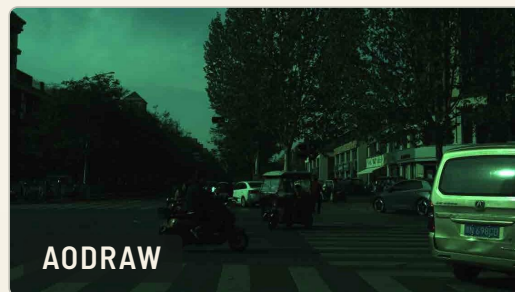
16 bit

Canon 5D IV



24 bit

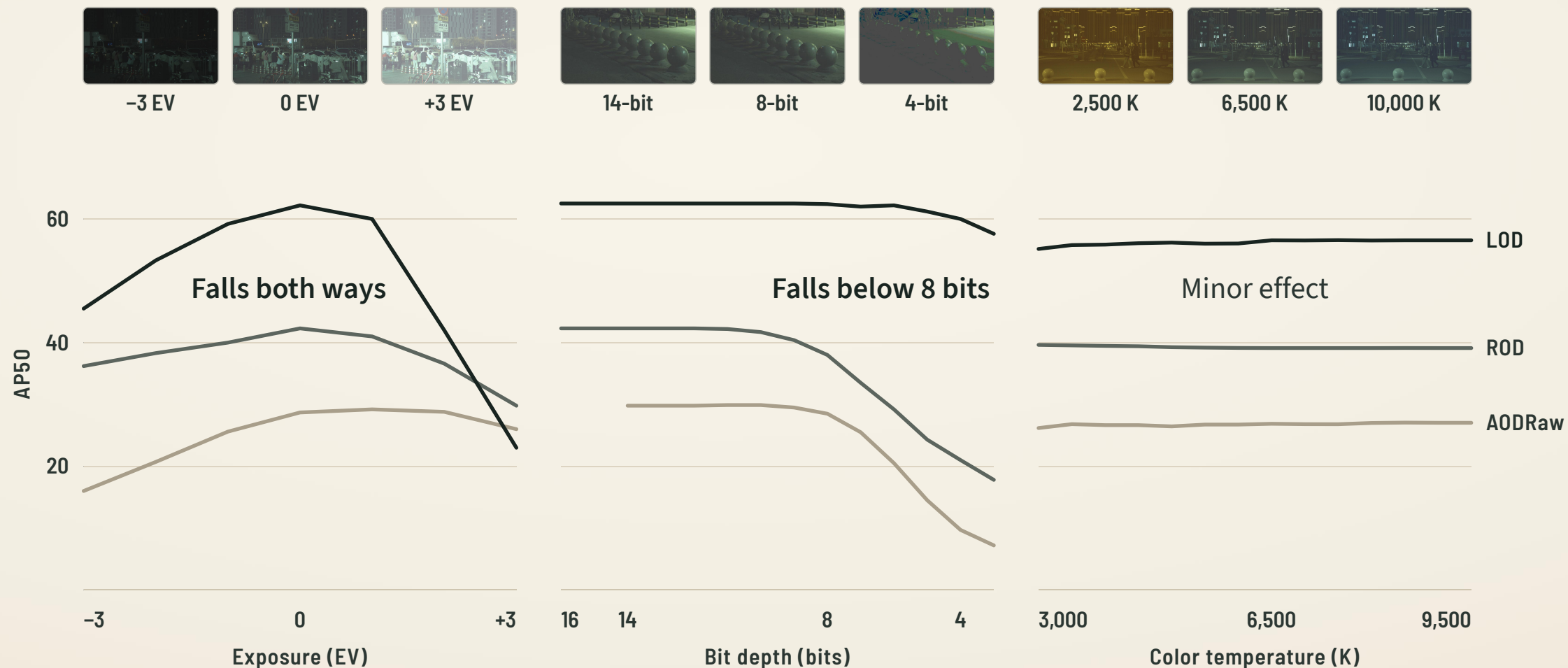
Sony IMX490



RAWild reads each image's histogram

ISP-relevant factors affect detection

Exposure, bit depth, color: fixed at capture, then rendered by an ISP. On linear RAW, shifts can cost AP50 (paper Fig. 3).



Existing ISP modules are tied to specific sensors

They learn a full ISP or a chain of ISP-style blocks; RAWild re-predicts two lightweight operators per image.

Existing methods

Learn the whole ISP end to end

ReconfigISP · AdaptiveISP · Dark-ISP

Insert ISP-style blocks by stage

RAW-Adapter · RAM

Both stay tied to one sensor's settings and a handcrafted chain of operators, so one pipeline does not carry over across illuminations, camera responses and bit depths.

What RAWild contributes

A global curve, then a local grid

A Bézier curve g corrects global tone, a bilateral grid M refines local color; a histogram-guided predictor sets both for every image.

Two stable reparameterizations

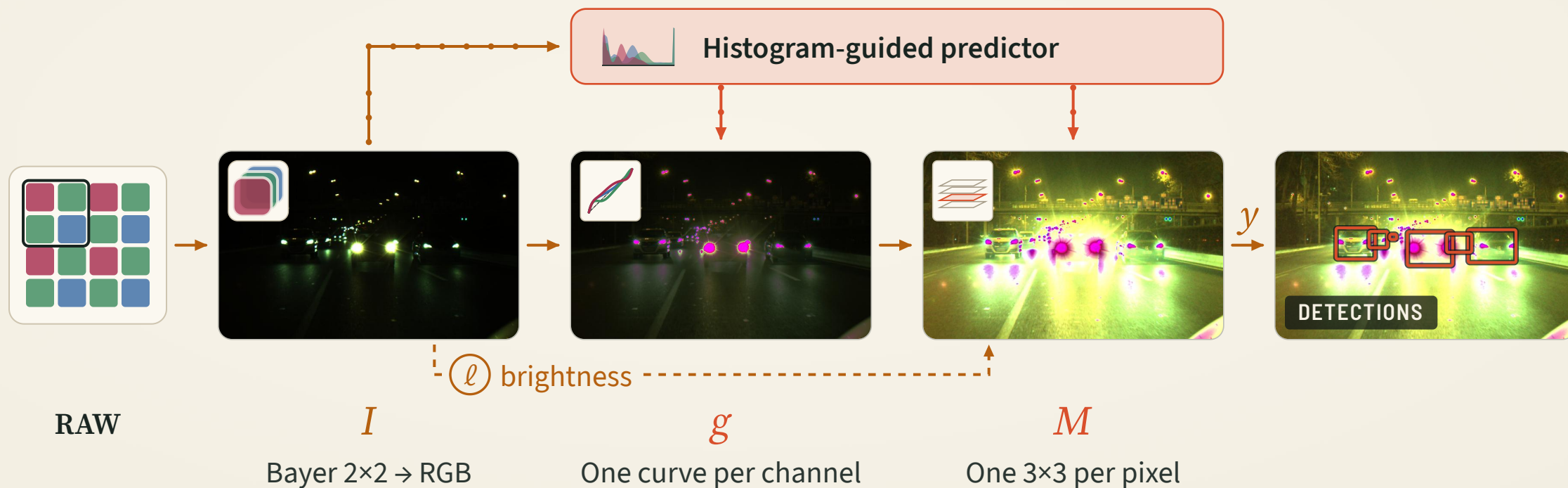
Delta-residual Bézier and gain-mixing grid $M = D(I + A)$: both start as the identity and stay bounded, so training is stable across sensors.

Physics-based RAW simulation

Virtual sensors from 28 real camera spectra; tested on 6 real-capture benchmarks (per-sensor and mixed) and 3 synthetic sets, 10 to 24 bits.

RAWild intuition: a global curve, then a local grid

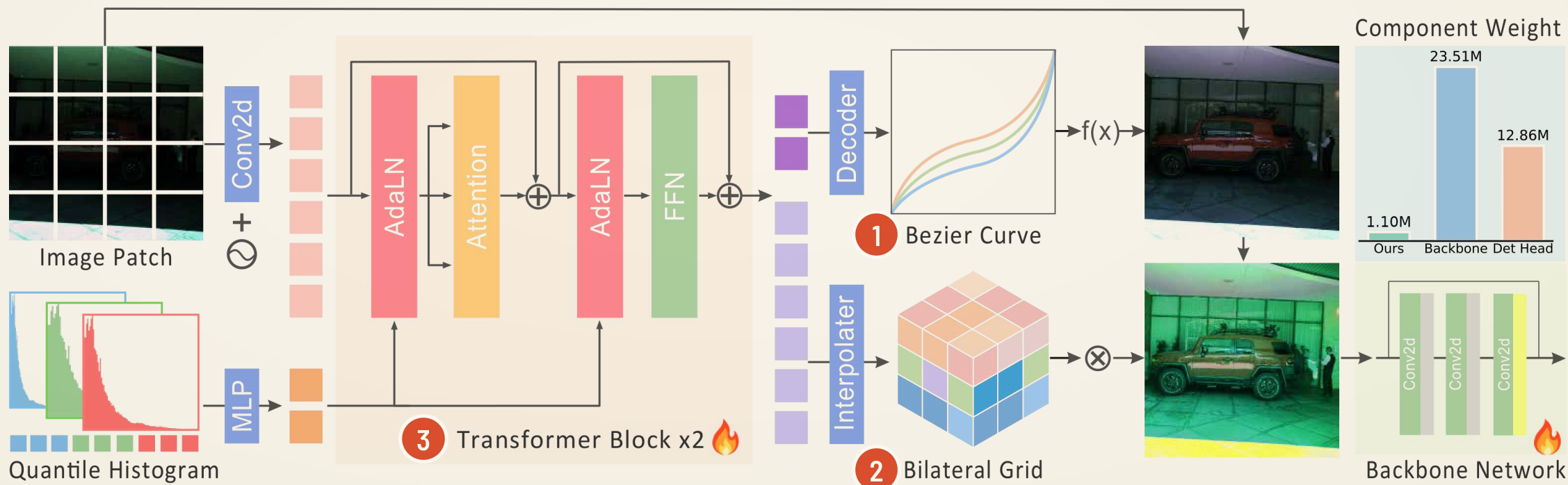
RAWild turns each RAW image I into a detector-ready y : a global curve g , then a local 3×3 matrix M .



$$y(p) = M(p, \ell(p)) g(I(p))$$

Method overview

Paper Fig. 4: a small histogram-guided Transformer predicts both operators for every image.



1 Global Bézier curve

Pooled tokens give per-channel offsets δ ; the curve is $g(f(x))$ in the figure).

2 Local bilateral grid

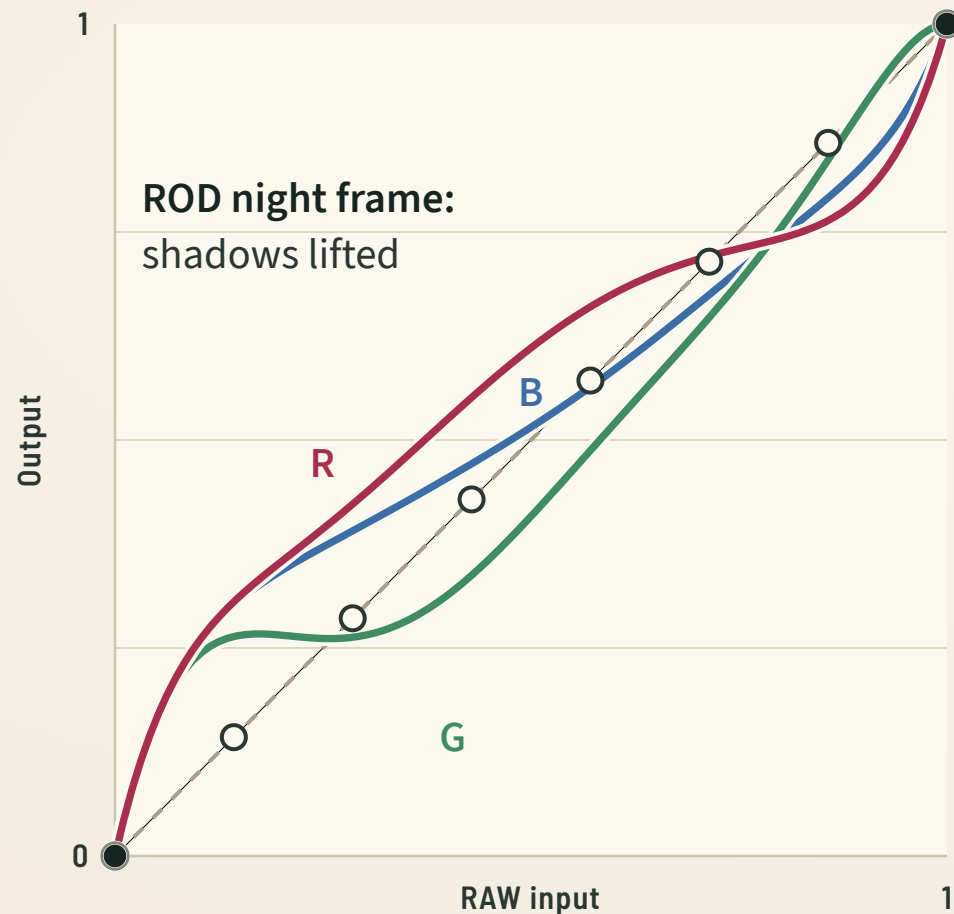
Per-token outputs fill a grid of 3×3 matrices M , sliced by position and brightness.

3 Histogram-guided predictor

64 quantiles per channel scale and shift every Transformer block (AdaLN).

1 Global Bézier curve

g bends each channel's tone with a Bézier curve. It starts as the identity and learns only bounded offsets $\frac{1}{2} \tanh \delta$.

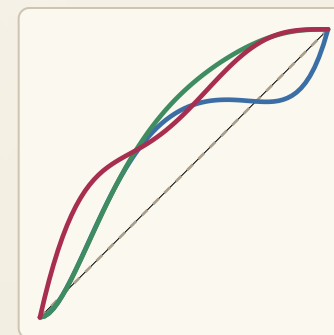
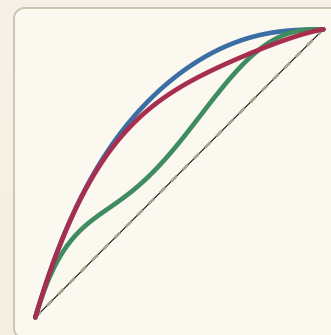
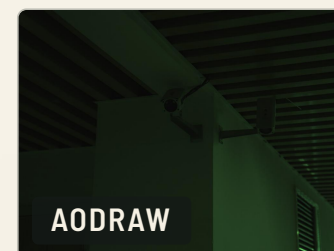
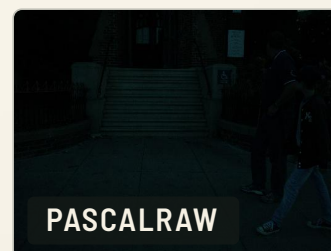


$$p_i = \text{clip}(i/n + \frac{1}{2} \tanh \delta_i) \quad i = 1 \dots n-1$$

○ $\delta = 0$: the identity

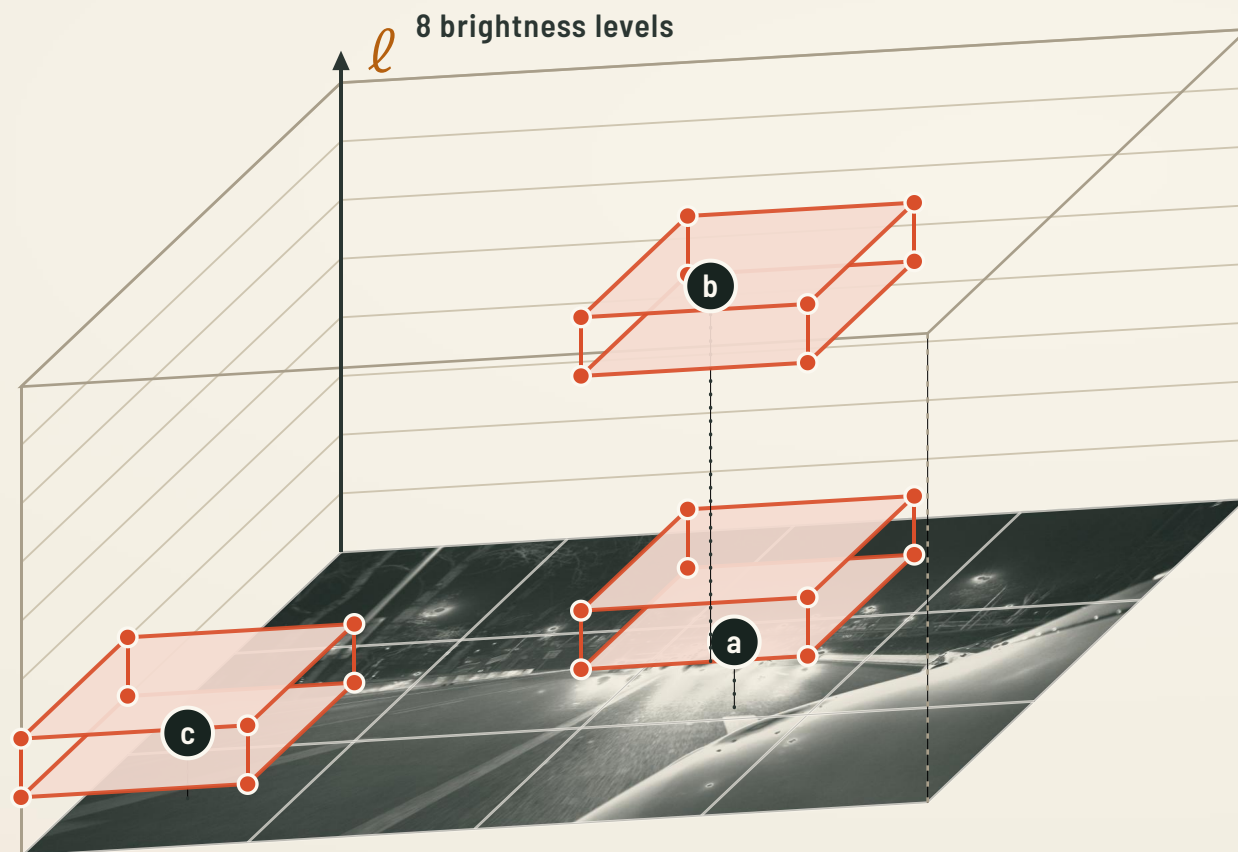
● Ends pinned at 0 and 1

Other sensors, other curves

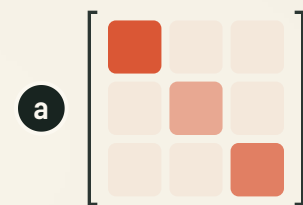


2 Local bilateral grid

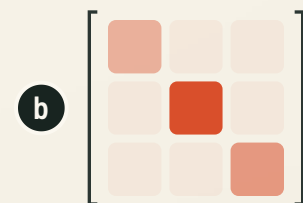
M lives in a bilateral grid: each pixel looks up its 3×3 matrix by position and brightness ℓ .



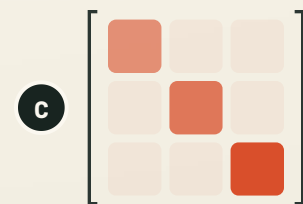
16 × 16 px cells (drawn enlarged)



Road pixel:
its own 3×3



Brighter, same cell:
a new 3×3



Same ℓ , other cell:
a new 3×3

2 Grid matrices: gain and mixing kept apart

Each 3×3 is a diagonal gain D times (identity + small mixing A): it stays invertible and starts as the identity.

$$\begin{array}{ccc}
 \mathbf{D} & \mathbf{I} + \mathbf{A} & \mathbf{M} \\
 \left[\begin{array}{ccc} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{array} \right] & \cdot \left[\begin{array}{ccc} 1 & a_{12} & a_{13} \\ a_{21} & 1 & a_{23} \\ a_{31} & a_{32} & 1 \end{array} \right] & = \left[\begin{array}{ccc} d_1 & d_1 a_{12} & d_1 a_{13} \\ d_2 a_{21} & d_2 & d_2 a_{23} \\ d_3 a_{31} & d_3 a_{32} & d_3 \end{array} \right] \\
 \text{Gain } \times 0.37 \text{ to } \times 2.72 & \text{Mixing } \leq 5\% \text{ of gain} & \text{Identity at start,} \\
 \text{never flips sign} & \text{per entry} & \text{stays invertible}
 \end{array}$$

Gain only
−1.0 vs no grid **65.9**

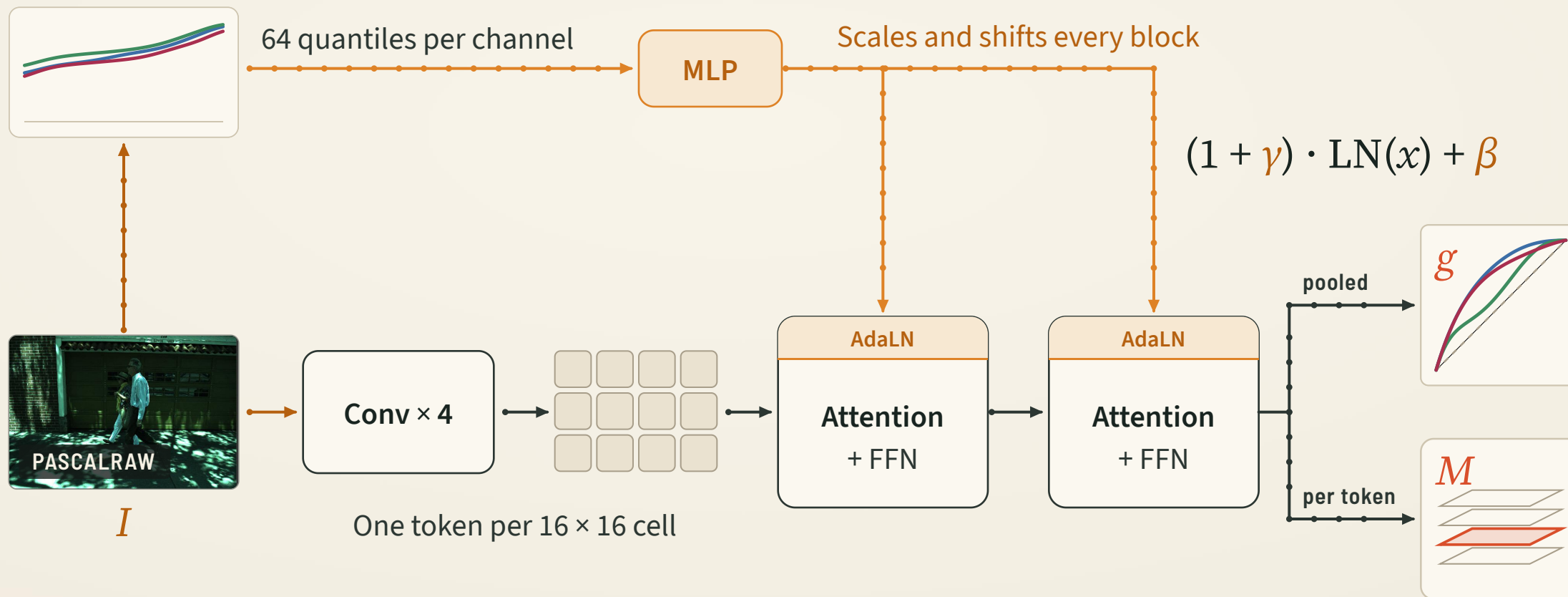
Mixing only
−0.4 vs no grid **66.5**

Both: $D(I + A)$
+2.2 vs no grid **69.1**

AP50 on LOD (Table 6). No grid: 66.9. Vanilla grid: 66.7. Either part alone does worse than no grid.

3 Histogram-guided predictor

A small Transformer reads the image; the image's 64-quantile histogram modulates every block. It outputs δ and the grid.



Stage-by-stage outputs on a night frame

One ROD frame through RAWild: the curve lifts the shadows, then the grid corrects color locally.

Linear RAW

INPUT

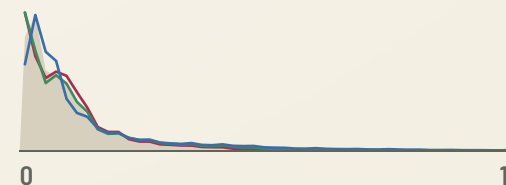
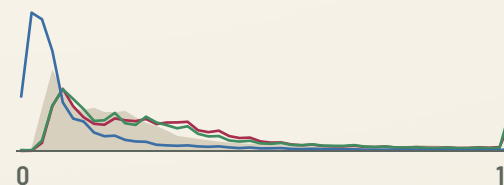
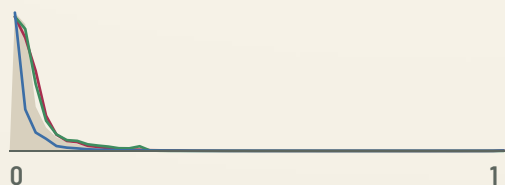
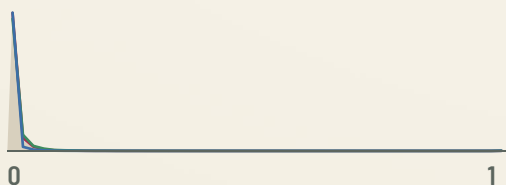
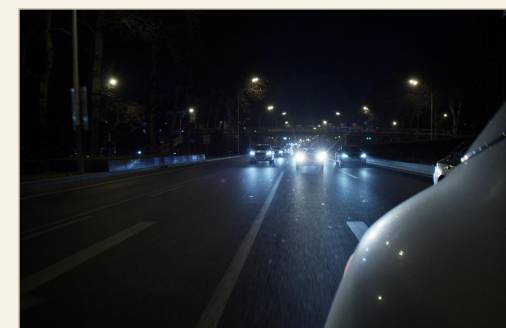
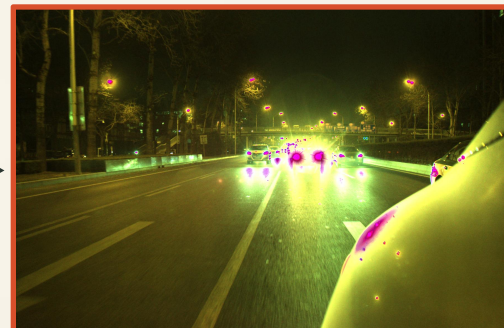
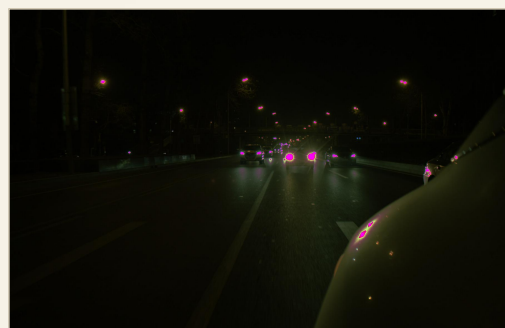
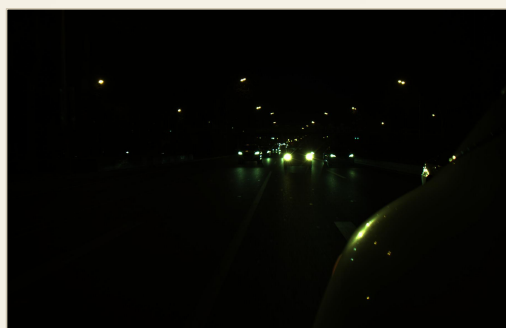
1 + Curve

2 + Grid

DETECTOR INPUT

For the eye

REFERENCE



Nearly black: almost every pixel sits close to 0.

The curve spreads the dark tones, channel by channel.

The grid corrects color locally. This is what the detector sees.

A camera-style rendering made for people, shown only for comparison.

Tuned for the detector, not the eye: RAWild's output does not need to look natural.

Simulation: motivation and pipeline

Mixed-sensor RAW with balanced labels is scarce, so RAWild re-renders labeled RAW as new virtual sensors.

THE GAP

Labeled RAW covers few sensors

Most RAW benchmarks use one camera; balanced multi-device RAW is scarce.

THE IDEA

Simulate new sensors

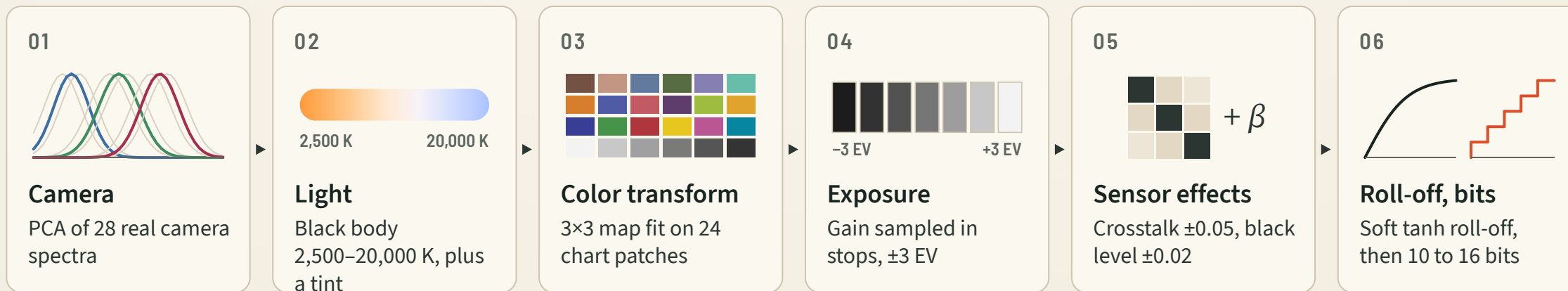
Sample a camera, a light, an exposure and a bit depth, then apply them to every pixel.

THE PAYOFF

Cross-sensor sets with free labels

Nothing moves spatially, so every box stays valid: PASCALRAW, LOD and ROD (Syn).

PIPELINE · SAMPLED PER IMAGE, APPLIED PER PIXEL



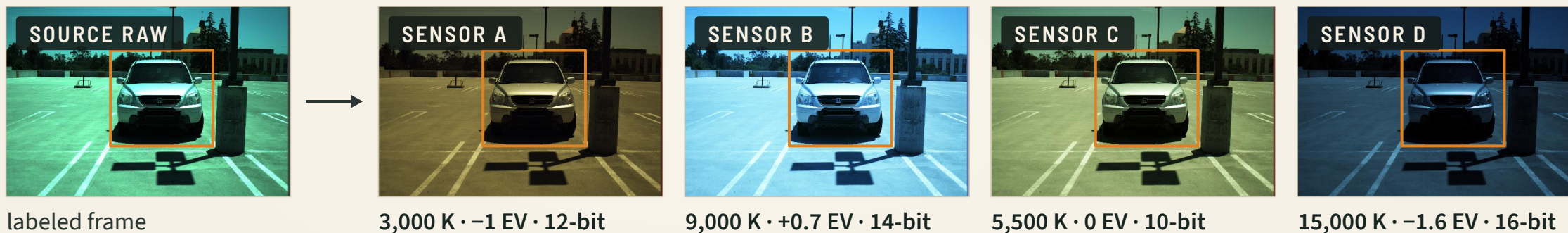
$$y = \alpha (\text{diag}(g_{\text{tint}}) M_{A \rightarrow B, T} + \varepsilon) x_{\text{bal}} + \beta \quad \text{then divide by the new camera's white-balance gains, roll off, quantize}$$

Paper App. A: x_{bal} is the labeled source RAW after gray-world white balance; y is then divided by the new camera's gray-world gains.

Simulation results

Each draw is a new virtual sensor with the same boxes. On LOD (Syn), RAWild beats the best baseline by 7.2 AP50.

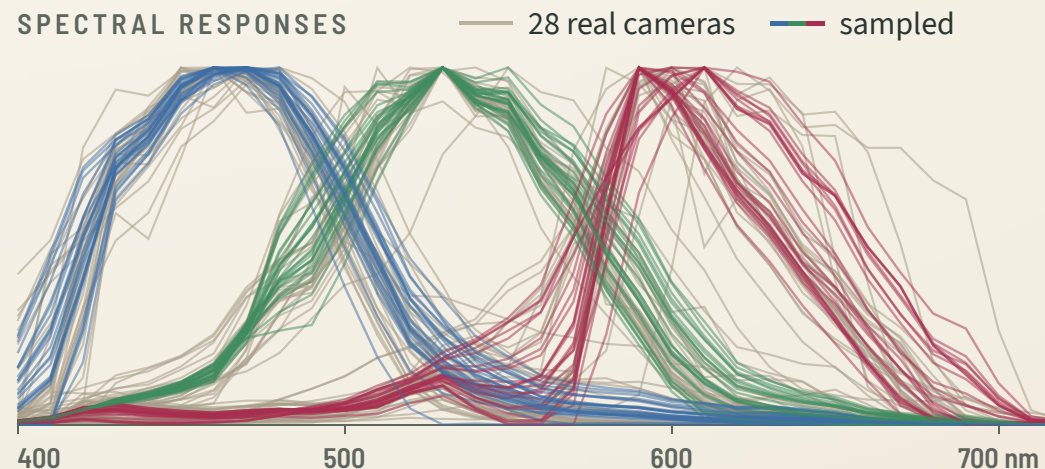
ONE LABELED PASCALRAW FRAME, FOUR VIRTUAL SENSORS



ONE FACTOR AT A TIME

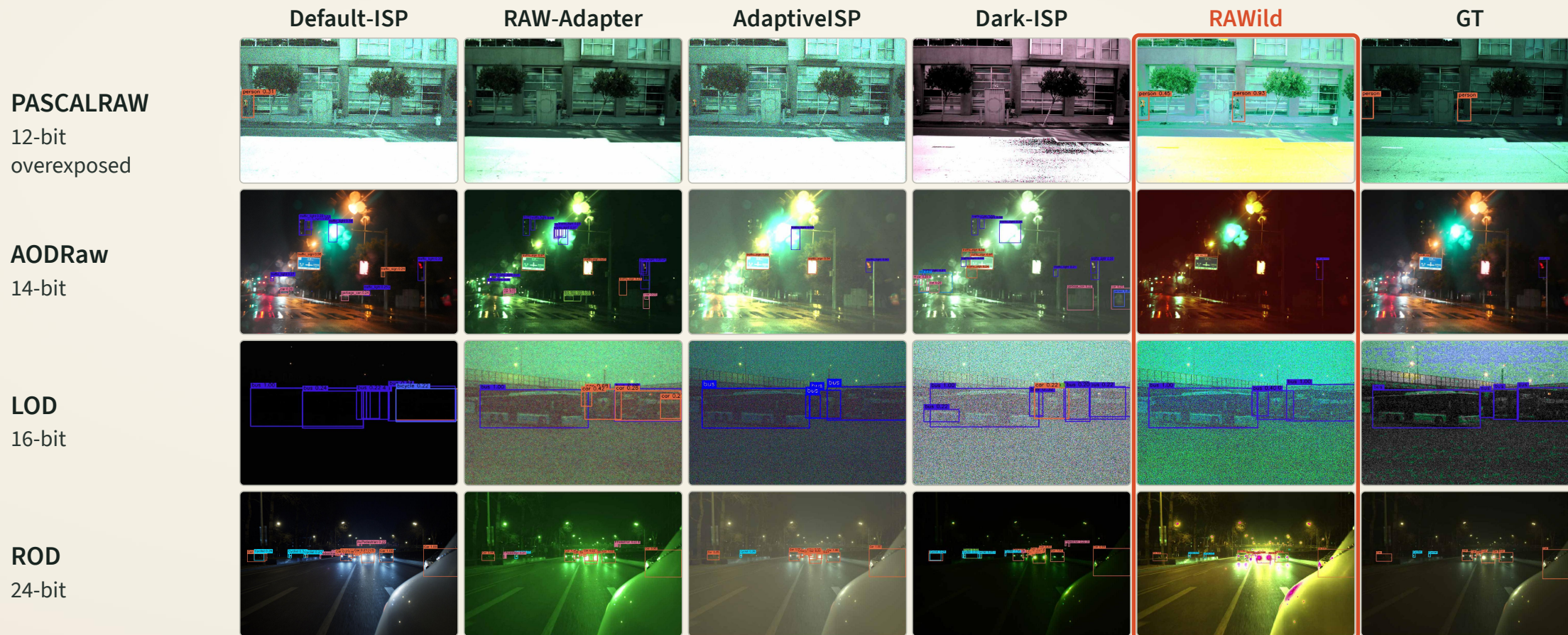


SPECTRAL RESPONSES



Experiments overview: detections on four sensors

Paper Fig. 5, one frame per benchmark (12 to 24 bits): baselines miss or mislocalize objects in glare and in the dark.



RAWild column: detector input, not for viewing. **Default-ISP:** official sRGB for LOD (short exposure, hence dark) and AODRaw, fixed software ISP for PASCALRAW, reference ISP for ROD.

Per-sensor benchmarks

One model per benchmark, RetinaNet with ResNet-50: on par at 12 to 14 bits, the largest gains on 16- and 24-bit sensors.

AP50 (%)	PASCALRAW · 12-bit			AODRaw · 14-bit	LOD · 16-bit	ROD · 24-bit
	LOW LIGHT	NORMAL	OVER-EXP.	SONY A7 III	CANON 5D IV	SONY IMX490
Linear-RAW	76.7	86.6	87.5	23.0	57.1	30.4
Default-ISP	83.9	89.6	89.7	32.4	58.8	41.8
ReconfigISP	85.6	89.5	89.9	29.0	53.8	34.4
RAW-Adapter	86.5	89.4	89.9	29.3	60.8	39.5
AdaptiveISP	88.7	90.1	90.1	33.8	61.4	43.8
Dark-ISP	71.4	89.4	88.4	32.4	55.6	40.9
Dr.RAW	88.7	89.8	89.8	36.3	64.1	47.2
RAM	89.8	90.4	90.2	36.5	66.6	45.5
RAWild	89.1	90.2	90.2	36.2	69.1	55.4
<i>Δ vs. best baseline</i>	-0.7	-0.2	0.0	-0.3	+2.5	+8.2

Bold: best in each column at the precision shown. RAM was run under our protocol (author response); the other rows are paper Table 1. Δ uses unrounded values.

Mixed-sensor benchmarks

Each column, one model trained across its sensors: three synthetic sets from our simulation, two real mixes.

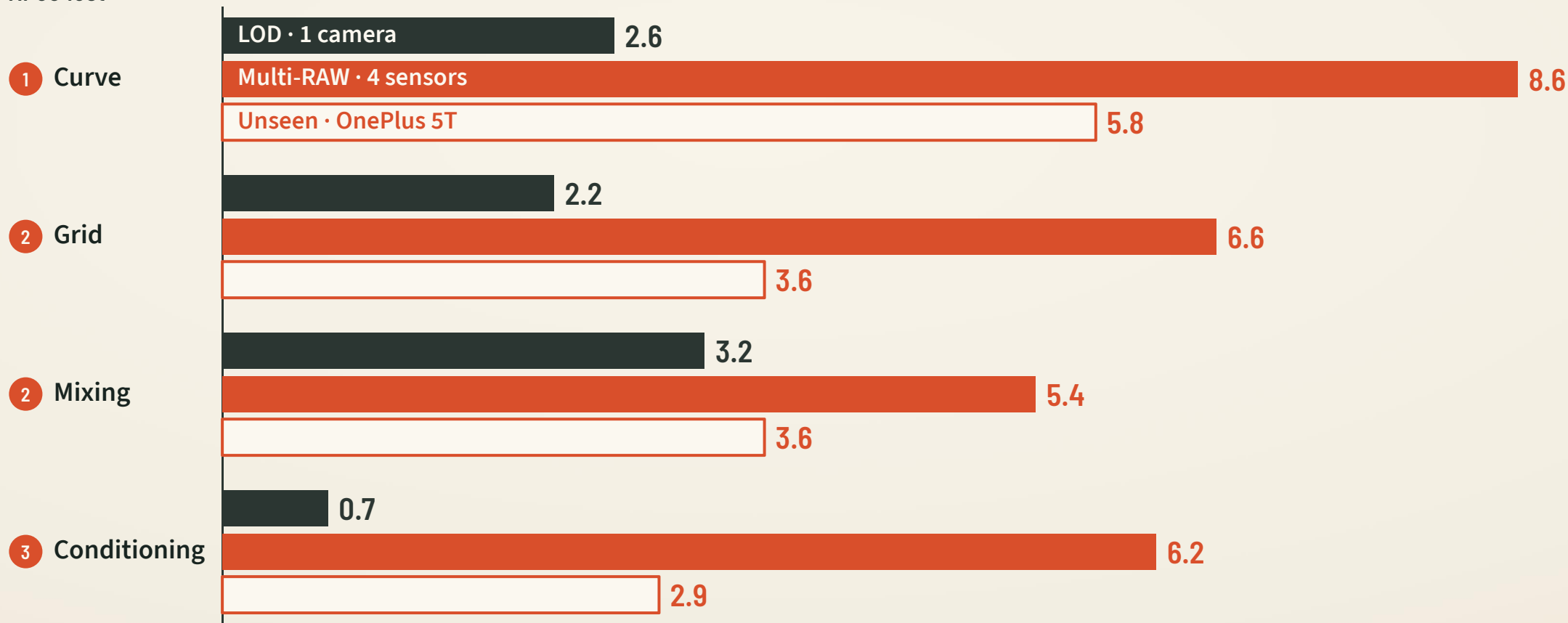
	Synthetic sensors · our simulation			Real mixed sensors	
AP50 (%)	PASCALRAW	LOD	ROD	PASCALRAW + LOD	MULTI-RAW
Linear-RAW	89.2	53.1	35.0	60.6	25.4
RAW-Adapter	88.7	58.7	36.4	65.4	23.3
Dark-ISP	88.9	58.4	37.0	64.2	26.0
Dr.RAW	89.5	59.2	42.3	66.6	25.7
RAM	—	—	—	—	30.1
InvISP	—	—	—	—	31.4
ParamISP	—	—	—	—	32.5
RAWild	89.5	66.5	42.9	68.7	35.3
<i>Δ vs. best baseline</i>	0.0	+7.2	+0.6	+2.1	+2.8

Bold: best in each column at the precision shown. RAM, InvISP, ParamISP: Multi-RAW results from the author response; the other rows are paper Table 2. Δ uses unrounded values.

Ablation: removing one component at a time

Every part matters: a little on one camera, more across four sensors, and still on an unseen one.

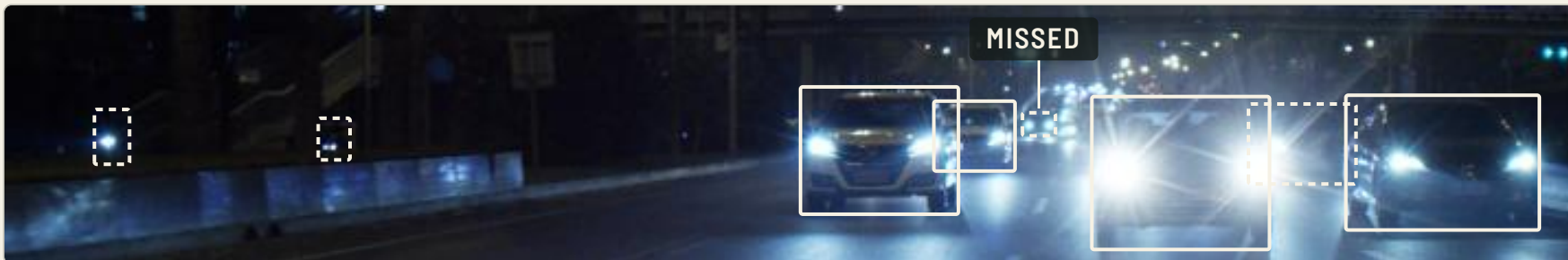
AP50 lost



Mixing: a gain-only grid (D only). **Conditioning:** no histogram on LOD (paper Table 4), w/o conditioning on Multi-RAW. **Multi-RAW and OnePlus 5T:** author response.

Example: small objects in a night scene

Zoom on the ROD night frame: dashed boxes are labeled objects that a method missed (score ≥ 0.5).



Dark-ISP

4 of 8



Dr.RAW

5 of 8



RAWild

8 of 8

score ≥ 0.5 · IoU ≥ 0.5

Example: one model on four devices

Multi-RAW: trained on three devices and tested on four, including the unseen OnePlus 5T.



TRAIN / TEST IMAGES

ASI 294MC Pro	817 / 823
Huawei P30 Pro	842 / 834
iPhone XS Max	332 / 322
OnePlus 5T	0 / 85

+2.8

AP50 over ParamISP, the strongest baseline, on the full four-device test set (35.3 vs. 32.5)

Takeaways

A global curve, then a local grid

Sensor shifts act at two scales: a Bézier curve fixes global tone, a bilateral grid fixes local color. Both start as the identity.

Re-predicted for every image

A small Transformer reads each image's histogram and predicts both operators; the whole detector runs in 10.7 ms on an H100.

New sensors by simulation

Physics-based RAW simulation turns labeled RAW into virtual sensors for training and testing.

+8.2

AP50 over the strongest baseline on 24-bit ROD
(RetinaNet, ResNet-50)

+2.8

AP50 over the strongest baseline on Multi-RAW:
one model, four devices

Limitation: real multi-sensor RAW with balanced labels is still scarce; such a dataset is worth building.

RAWild.

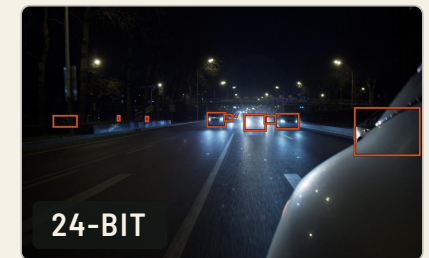
Toward Sensor-Agnostic RAW Object Detection
via Physics-Guided Curve and Grid Modeling

Thank you.

i2wm.github.io/RAWild

[arXiv 2605.05941](https://arxiv.org/abs/2605.05941) · github.com/I2WM/RAWild

Supported by JST Moonshot R&D JPMJPS2011 · JSPS KAKENHI JP26K21296 · JST BOOST JPMJBS2418



Matched controls and cost

LOD · same detector and training, only the front end changes

Front end	AP50	AP75
Bézier + bilateral grid	69.1	43.6
Bézier + 3D LUT	67.6	41.3
Bézier + bilateral filtering	67.3	38.9
PWL + bilateral grid	67.2	42.8
Reinhard + bilateral grid	67.2	41.0
Bézier only	66.9	43.5
Bilateral grid only	66.5	41.3
3D LUT only	59.7	31.2
Reinhard only	59.4	33.1
PWL only	58.3	33.7
Bilateral filtering only	57.8	34.1
Linear-RAW	57.1	37.1

RAWild.

Cost on H100 · GFLOPs at 667×400

Method	Params (M)	GFLOPs	ms	GB
Linear-RAW	0	0.0	7.2	0.23
RAW-Adapter	0.6	3.2	10.0	0.24
Dr.RAW	3.2	32.2	10.5	0.53
Dark-ISP	0.3	35.5	20.5	0.99
RAWild	1.1	1.9	10.7	0.24

Params, GFLOPs: the front end only; the RetinaNet detector is shared.

ms, GB: inference time and GPU memory of the whole detector.

Sources: paper Table 3; params and GFLOPs from the author response.

Multi-RAW, sensor by sensor

VARIANT	Multi-RAW · all sensors			AP50 per sensor		
	AP50	AP75	ASI	HUAWEI	IPHONE	ONEPLUS (UNSEEN)
w/o curve	26.7	11.7	21.6	31.9	27.4	22.4
w/o grid	28.7	12.8	23.2	34.0	30.0	24.6
w/o mixing (D only)	29.9	12.5	25.0	34.9	30.9	24.6
w/o conditioning	29.1	11.8	23.9	34.9	29.5	25.3
RAWild (full)	35.3	13.4	34.9	34.4	43.8	28.2

Same RetinaNet-ResNet-50 detector, schedule, seed and evaluation for every variant; only the named component changes. OnePlus 5T has no training images (0 train / 85 test).
Bold: best in each column at the precision shown. Source: author response.

Under RAM's own protocol

DATASET · %	RAM			RAWild		
	mAP	AP50	AP75	mAP	AP50	AP75
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

RAWild leads all three metrics on five of the seven sets; RAM leads on LOD-Dark, and on PASCALRAW in mAP and AP50 (RAWild leads its AP75).
RAM's benchmarks, metrics and detector; RAM run from its official checkpoints. Bold: the better of the two. Source: author response.