

Real-Time Scene-Adaptive Tone Mapping for High-Dynamic Range Object Detection

—Supplementary Material—

In this file, we provide the following supplementary studies:

- Details of Implemented HDR ISP pipeline Sec. 1.
- The proof of approximation to Dynamic Range Sec. 2.
- The proof of scaling-invariant Tone Mapping Sec. 3.
- More Ablation Studies. Sec. 4.
- More Vision Comparison. Sec. 5.
- Real World Evaluation. Sec. 6.

1 HDR ISP Details

In this section, we describe the HDR ISP pipeline used in the comparison method [1]. This pipeline consists of a series of operations, as illustrated in Fig. 1. We follow the implementation described in [1, 2], with modifications applied to the modules preceding the tone-mapping algorithms. The intermediate results of key components (indicated by the red dashed lines in Fig. 1) are shown in Fig. 2, which demonstrate how the HDR data is transformed into a visually appealing LDR image after undergoing several nonlinear operations. Next, we introduce the key steps in this process.

Black-level Correction: We should subtract an offset from all pixels so that pixels receiving no light have a value of zero. This offset is obtained from optically shielded pixels on the sensor.

$$I_{blc} = I - I_{bl} \quad (1)$$

where I_{bl} is the black level.

Anti-Aliasing filter: An anti-aliasing filter is a type of low-pass filter that prevents aliasing components from being sampled.

$$I_{aaf} = I * k_{aaf} \quad (2)$$

where k_{aaf} is 5×5 filter kernel, having non-zero elements only at the corners and center. The kernel

is defined as: $k_{aaf} = 1/16 \cdot \begin{bmatrix} 1 & \dots & 1 \\ \dots & 8 & \dots \\ 1 & \dots & 1 \end{bmatrix}$.

Auto White Balance: The AWB (Auto White Balance) module is responsible for adjusting the image to ensure that the four (RGGB) channels are linearly scaled, so that grays in the scene correspond to grays in the image. These scaling factors are calculated using the Gray-World white balance algorithm, which adjusts the pixel values based on the gray-world assumption. This assumption posits that the average of all color channels should produce a neutral gray image.

$$\begin{bmatrix} I_{R'} \\ I_{Gr'} \\ I_{Gb'} \\ I_{B'} \end{bmatrix} = \begin{bmatrix} g_R & 0 & 0 & 0 \\ 0 & g_{Gr} & 0 & 0 \\ 0 & 0 & g_{Gb} & 0 \\ 0 & 0 & 0 & g_B \end{bmatrix} \begin{bmatrix} I_R \\ I_{Gr} \\ I_{Gb} \\ I_B \end{bmatrix} \quad (3)$$

where the parameters g_R, g_{Gr}, g_{Gb}, g_B are color gain, which can be derived by the gray world algorithm.

Demosaic: Demosaicing converts a Bayer raw image into a full-resolution linear RGB image, preserving texture details. We use a combination of techniques from the Malvar algorithm [3].

$$I_{R,G,B} = \text{Demosaic}(I_{(R,Gr,Gb,B)}) \quad (4)$$

where $I_{(R,Gr,Gb,B)}$ denote the single channel bayer array, $I_{R,G,B}$ denote the complete 3-channels RGB image.

Local Tone Mapping: The LTM (Local Tone-Mapping) block simulates the exposure fusion algorithm [2, 4] by brightening darker areas while ensuring that brighter content remains unsaturated.

$$I_{ltm} = \sum_i^n I_i \cdot w_i \quad (5)$$

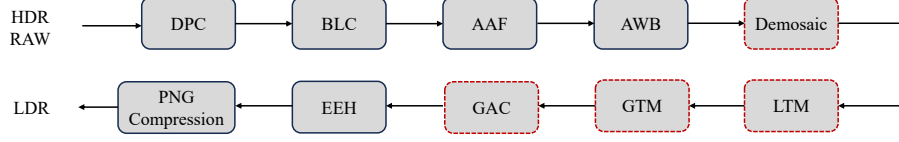


Figure 1: The key components of the HDR ISP pipeline, see text for details. We visualize the results for the modules marked with a red dotted line.

where I_i is the synthetic exposure image, and w represents the corresponding weights calculated based on image content.

Global Tone Mapping: The GTM (Global Tone-Mapping) block follows the LTM, enhancing the overall luminance. An S-shaped contrast-enhancing tone curve [5] is applied to the linear sRGB image. This curve is then concatenated with the sRGB color component transfer function, which transforms the image from linear sRGB to nonlinear sRGB. The Reinhard tone curves [5] can be expressed as follows:

$$\bar{I}_w = \frac{1}{N} \exp \left(\sum \log (\delta + I) \right) \quad (6)$$

$$I_{\text{gtm}} = \frac{I \cdot \left(1 + \frac{I}{\bar{I}_w^2} \right)}{1 + I} \quad (7)$$

where δ is a small value to avoid numerical overflow.

Gamma Correction: The GAC (gamma correction) is used to match the non-linear characteristics of a display device or human perception. We adopt the correction function Eq. (8) recommended in ITU-R BT. 709 standard [6], which is widely used in commodity cameras today.

$$I_{\text{gamma}} = \begin{cases} 12.92 \cdot I, & I \leq 0.00304, \\ 1.055 \cdot I^{1/2.4} - 0.055, & I > 0.00304. \end{cases} \quad (8)$$

Edge Enhancement: The EEH (Edge Enhancement) module enhances image details and edges, improving image clarity and visual appeal. It is particularly useful for accentuating finer image structures. The module can be expressed as follows:

$$I_{\text{sharpen}} = p_s \cdot I + (1 - p_s) \cdot I_{\text{blurred}}, \quad (9)$$

2 The Proof of the Approximation to Dynamic Range

1. Dynamic Range Ratio for a Single Gaussian Component. For a Gaussian component with mean μ and standard deviation σ , define the dynamic range:

$$d = \frac{\mu + \sigma}{\mu - \sigma}, \mu > \sigma. \quad (10)$$

This dynamic range measures the distance of the component relative to its mean.

2. Relating d to μ and σ . Solve for μ in terms of d and σ :

$$\mu = \sigma \cdot \frac{d+1}{d-1}. \quad (11)$$

3. Second Moment and Mean of a Single Component For a Gaussian component:

$$\mathbb{E}(x^2) = \sigma^2 + \mu^2, \quad \mathbb{E}(x) = \mu.$$

Thus:

$$\mathbb{E}(x^2) + \mathbb{E}(x) = \sigma^2 + \mu^2 + \mu.$$

Substitute $\mu = \sigma \cdot \frac{d+1}{d-1}$:

$$\mathbb{E}(x^2) + \mathbb{E}(x) = \sigma^2 + \left(\sigma \cdot \frac{d+1}{d-1} \right)^2 + \sigma \cdot \frac{d+1}{d-1}. \quad (12)$$



Figure 2: Visual comparison of key steps in the HDR ISP pipeline.

62 Simplify:

$$\mathbb{E}(x^2) + \mathbb{E}(x) = \sigma^2 \left[1 + \frac{(d+1)^2}{(d-1)^2} \right] + \sigma \cdot \frac{d+1}{d-1}. \quad (13)$$

63

64 4. Bounding $\mathbb{E}(x^2) + \mathbb{E}(x)$ using R Expand the squared term

$$\frac{(d+1)^2}{(d-1)^2} = \frac{d^2 + 2d + 1}{d^2 - 2d + 1} = 1 + \frac{4d}{(d-1)^2}. \quad (14)$$

65 Substitute back

$$\mathbb{E}(x^2) + \mathbb{E}(x) = \sigma^2 \left[2 + \frac{4d}{(d-1)^2} \right] + \sigma \cdot \frac{d+1}{d-1} \quad (15)$$

66

67 5. Inequality Analysis Using the AM-GM (Arithmetic and Geometric Means) Inequality:

$$\frac{\mu + \sigma}{2} \geq \sqrt{\mu\sigma} \implies \mu + \sigma \geq 2\sqrt{\mu\sigma}. \quad (16)$$

68 For the dynamic range d :

$$d = \frac{\mu + \sigma}{\mu - \sigma} \geq \frac{2\sqrt{\mu\sigma}}{\mu - \sigma}. \quad (17)$$

69 This implies that d grows as the overlap between μ and σ increases, amplifying $\mathbb{E}(x^2) + \mathbb{E}(x)$.

70 6. For a GMM with K components:

$$\mathbb{E}(x^2) + \mathbb{E}(x) = \sum_{i=1}^K \pi_i (\sigma_i^2 + \mu_i^2 + \mu_i). \quad (18)$$

71 Expressing each μ_i in terms of $d_i = \frac{\mu_i + \sigma_i}{\mu_i - \sigma_i}$:

$$\mathbb{E}(x^2) + \mathbb{E}(x) = \sum_{i=1}^K \pi_i \left[\sigma_i^2 + \left(\sigma_i \cdot \frac{d_i + 1}{d_i - 1} \right)^2 + \sigma_i \cdot \frac{d_i + 1}{d_i - 1} \right]. \quad (19)$$

72 Ignoring first-order terms and constants and focusing on the dominant components, we obtain the
73 final expression:

$$\mathbb{E}(x^2) + \mathbb{E}(x) \propto \sum_{i=1}^K \pi_i \left(\frac{d_i + 1}{d_i - 1} \right)^2, \quad d_i = \frac{\mu_i + \sigma_i}{\mu_i - \sigma_i} \quad (20)$$

74

75 3 Detailed Proof of Scaling-Invariant Tone Mapping

76 We construct the tone mapper TM_{SI} by a neural network composed of $\{\text{Conv-BN-ReLU}\}$ with L
77 layers and remove all bias terms within this network. Then we start proof the TM is functionally
78 equivalent to a local tone mapping operator.

79 **1.Convolution:**For input x and kernel K_i :

$$\text{conv}_i(\alpha x) = \alpha \cdot \text{conv}_i(x), \quad \text{where } \text{conv}_i(x) = K_i * x. \quad (21)$$

Table 1: Performance comparison with the radiance range of neural photometric on RoD Dataset. **Bold** denotes the default setting.

Radiance Range	[1e3, 1e6]	[1e3, 1e7]	[1e3, 1e8]	[1e4, 1e7]	[1e4, 1e8]	[1e5, 1e7]	[1e5, 1e8]
mAP	49.7	49.8	49.9	49.8	49.7	49.6	49.6
mAR	58.6	58.7	58.7	58.7	58.7	58.6	58.5

Table 2: Quantitative comparison of different pretraining losses.

Method	Pretrained Loss	mAP	AP50	AP75	contrast
Faster R-CNN [7]	L1	41.3	67.1	48.2	0.08
	NLPD [8]	49.8	73.3	55.6	0.411

80 **2.Batch Normalization:**For input x and kernel K_i :

$$\text{BN}_i(x) = \gamma_i \cdot \frac{x - \mu_i(x)}{\sigma_i(x)}, \quad (22)$$

81 where $\mu_i(\alpha x) = \alpha \cdot \mu_i(x)$ and $\sigma_i(\alpha x) = \alpha \cdot \sigma_i(x)$.

82 **3.ReLU:**For input x :

$$r(x) = \max(x, 0). \quad (23)$$

83 Then we start proving this bias-free network is scaling-invariant:

$$\begin{aligned}
\text{TM}_{\text{SI}}(\alpha \cdot x) &= r \circ \text{BN}_L \circ K_L * \dots \circ r \circ \text{BN}_1 \circ K_1 * (\alpha y) \\
&= r \circ \text{BN}_L \circ K_L * \dots \circ r \circ \text{BN}_1 \circ (\alpha \cdot K \cdot y) && \# \text{Convolution Linearity} \\
&= r \circ \text{BN}_L \circ K_L * \dots \circ r \left(\gamma_1 \cdot \frac{\alpha \cdot K_1 * x - \alpha \cdot \mu_1}{\alpha \cdot \sigma_1} \right) && \# \text{BN statistic scaling} \\
&= r \circ \text{BN}_L \circ K_L * \dots \circ r \left(\gamma_1 \cdot \alpha \cdot \frac{K_1 * x - \mu_1}{\sigma_1} \right) \\
&= r \circ \text{BN}_L \circ K_L * \dots \circ \alpha \cdot r \left(\gamma_1 \cdot \frac{K_1 * x - \mu_1}{\sigma_1} \right) && \# \text{Homogeneity} \\
&= \alpha \cdot r \circ \text{BN}_L \circ K_L * \dots \circ r \circ \text{BN}_1 \circ K_1 * x \\
&= \alpha \cdot \text{TM}_{\text{SI}}(x).
\end{aligned} \quad (24)$$

84 where \circ denotes the cascading of network layers. In the scaling-invariant transformation, all operators
85 apply a linear transformation on local neighborhoods. Hence, this neural network is naturally a
86 local tone mapping model. Here, BN_i adaptively adjusts gains using local statistics (μ_i, σ_i) , and the
87 cascade of BN_i and ReLU activation r mimics tone curves that compress highlights and shadows
88 while preserving midtones. Since all these linear transformations are applied within a local window,
89 they effectively function as local tone mapping.

90 4 More Ablation Studies

91 **Ablation of Radiance Range.** In the proposed neural photometric calibration, we set the radiance
92 range as a hyperparameter. We conduct an ablation study on radiance and detection performance,
93 with the results shown in Table 1. The findings demonstrate that our neural photometric calibration is
94 not sensitive to the radiance range, highlighting the method’s robustness in handling varying lighting
95 conditions.

96 **Ablation of Pretraining Loss.** We conduct an ablation study on pretraining loss to demonstrate its
97 effect on convergence performance. We compare the pretraining loss between NLPD [8] and L1 loss,
98 with the results shown in Table 2. The L1 loss is supervised by HDR ISP results. The experiment
99 shows that NLPD pretraining enables the tone mapper to enhance details, thereby improving detection
100 performance.

101

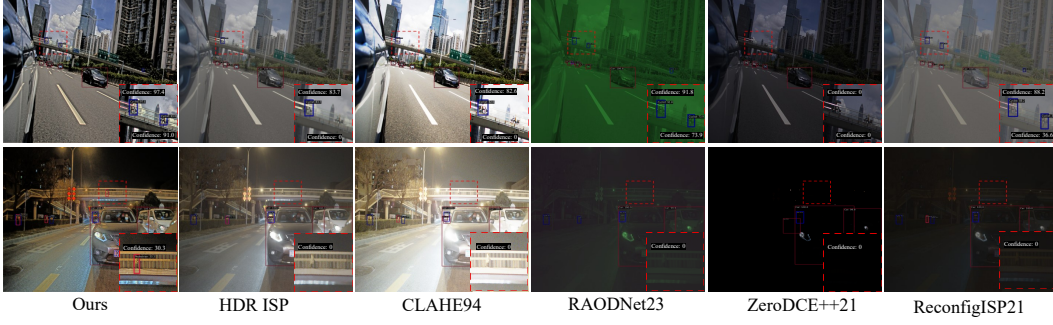


Figure 3: Visual comparison of different methods on HDR RAW inputs. The first row shows day scenes, while the second row presents night scenes. Our method outperforms the comparison methods. Please zoom in for confidence scores and class predictions.

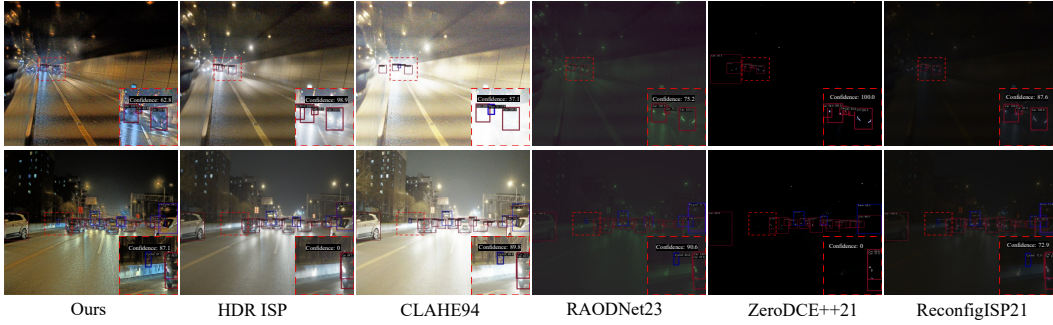


Figure 4: Visual comparison of different methods on HDR RAW inputs. The first row shows day scenes, while the second row presents night scenes. Our method outperforms the comparison methods. Please zoom in for confidence scores and class predictions.

5 More Visual Comparison.

We show more visual results in Fig. 3 and Fig. 4. Specifically, we visualize the detection results of the comparison methods with confidence scores greater than 0.3, in different scenarios of the RoD dataset [9]. In the first row of Fig. 3, the tunnel environment is shown, where our method effectively reduces false detections. In Fig. 4, shows a typical HDR scene, where our method detects small objects that other methods fail to identify.

6 Real World Evaluation.

HDR Video Validation. We collect HDR RAW video sequences from autonomous driving scenes to validate the proposed method, using YOLOv3 [10] as the base detector. The entire pipeline is evaluated on the Nvidia Jetson AGX Orin (16-bit float). We assess the input HDR RAW images at 2K resolution (2048×1080) and 4K resolution (4096×2160), with the proposed lightweight method **ours (Lite)** achieving 113 FPS at 2K resolution and 45 FPS at 4K resolution. Additionally, we have created a video demo (**video_demo.mp4**) in the supplementary file to showcase the detection results on the video sequences. Please refer to the attachment. We will release the video sequences and corresponding annotations once the dataset is complete.

References

- [1] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016.

- 121 [2] Antoine Monod, Julie Delon, and Thomas Veit. An Analysis and Implementation of the HDR+ Burst
122 Denoising Method. *Image Processing On Line*, 11:142–169, 2021. [https://doi.org/10.5201/](https://doi.org/10.5201/ipol.2021.336)
123 [ipol.2021.336](https://doi.org/10.5201/ipol.2021.336).
- 124 [3] Henrique S Malvar, Li-wei He, and Ross Cutler. High-quality linear interpolation for demosaicing of
125 bayer-patterned color images. In *2004 IEEE International Conference on Acoustics, Speech, and Signal*
126 *Processing*, volume 3, pages iii–485. IEEE, 2004.
- 127 [4] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high
128 dynamic range photography. In *Computer graphics forum*, pages 161–171. Wiley Online Library, 2009.
- 129 [5] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for
130 digital images. *ACM Trans. Graph.*, 21(3):267–276, 2002.
- 131 [6] Matthew Anderson, Ricardo Motta, Srinivasan Chandrasekar, and Michael Stokes. Proposal for a standard
132 default color space for the internet—srgb. In *Color and imaging conference*, volume 4, pages 238–245.
133 Society of Imaging Science and Technology, 1996.
- 134 [7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object
135 detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*,
136 39(6):1137–1149, 2016.
- 137 [8] Valero Laparra, Alex Berardino, Johannes Ballé, and Eero P Simoncelli. Perceptually optimized image
138 rendering. *Journal of the Optical Society of America A*, 34(9):1511–1525, 2017.
- 139 [9] Ruikang Xu, Chang Chen, Jingyang Peng, Cheng Li, Yibin Huang, Fenglong Song, Youliang Yan, and
140 Zhiwei Xiong. Toward raw object detection: A new benchmark and a new model. In *Proceedings of the*
141 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13384–13393, 2023.
- 142 [10] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*,
143 2018.