FLEXHDR: LEARNING ADAPTIVE EXPOSURE CONTROL FOR SPLIT-PIXEL HDR IMAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Split-pixel sensors offer a promising single-shot solution for High Dynamic Range (HDR) imaging, especially in safety-critical applications like autonomous driving. However, conventional auto-exposure methods typically use shared exposure settings across subframes, overlooking their distinct sensitivities and noise characteristics and thereby limiting dynamic range utilization and downstream perception performance. In this work, we propose FlexHDR, a task-driven neural exposure control framework that incorporates sensor priors to predict exposure parameters for individual subframes. Our method combines a differentiable image formation model for closed-loop exposure optimization during training with a cross-attention module that fuses subframes in feature space for downstream perception. To support reproducible evaluation, we develop a calibrated split-pixel camera simulator and a synthetic HDR scene radiance dataset covering diverse automotive scenes. Experiments demonstrate consistent improvements in object detection and instance segmentation over existing auto-exposure and tone-mapping methods under challenging automotive HDR conditions. These results highlight the benefits of independent per-subframe exposure control and end-to-end training for robust HDR perception.

1 INTRODUCTION

High dynamic range (HDR) imaging has become indispensable for computer vision applications operating in uncontrolled environments, such as robotic manipulation, autonomous driving, and 3D scene perception. Its primary strength lies in capturing the full spectrum of scene radiance, which frequently exceeds the dynamic range of conventional sensors. Modern sensing techniques employ multiple exposures, specialized photodiodes, variable-gain circuits, and dedicated ISP pipelines (Darmont, 2013; Banterle et al., 2017; Brown & Kim, 2019; Mosleh et al., 2020) to optimize the dynamic range of the capture, thereby enabling HDR imaging. These techniques are critical for maintaining reliable visual perception under challenging lighting conditions, including direct sunlight, specular reflections, and deep shadows.

Popular multi-exposure fusion techniques extend dynamic range via temporal multiplexing (Mertens et al., 2007), but they often suffer from motion artifacts and flicker in dynamic scenes (Gallo et al., 2009). To address this, spatially multiplexed HDR sensors (Willassen et al., 2015; Asatsuma et al., 2019), such as the Sony IMX490 and Omni OV50H, employ a split-pixel design to capture scene luminance across the spatial domain. By acquiring and merging multiple subframes with distinct responsivities in a single exposure (Ohta, 2020), these sensors effectively extend the dynamic range while inherently avoiding motion artifacts and flicker.

However, split-pixel sensors are highly sensitive to exposure settings, necessitating precise control to ensure that the subframes cover the entire dynamic range. Although split-pixel sensors enable subframe-specific exposure control, existing auto-exposure (AE) methods, whether handcrafted (Shim et al., 2018; Shin et al., 2019; Tedla et al., 2023) or data-driven (Onzon et al., 2021; Tomasi et al., 2021; Zhang et al., 2024), still apply uniform exposure parameters across all subframes. These strategies effectively ignore the unique noise and signal characteristics of each subframe. Moreover, even task-supervised methods drive exposure through predefined control equations that bind exposure time and analog gain into a single update rule, preventing independent allocation across subframes. Optimizing per-subframe exposure time and gain for spatial split-pixel

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **(a) Scene Radiance Simulation.** Independent light-source components are rendered from 3D assets and combined with different weights to synthesize diverse illumination conditions. **(b) Sensor Simulation.** A split-pixel sensor simulator converts scene radiance into split-pixel captures. **(c) FlexHDR vs. Traditional HDR.** FlexHDR is compared with conventional HDR imaging under challenging high-contrast, dynamic, and low-light scenes.

HDR sensors (Asatsuma et al., 2019; Jähne, 2020) is a complex task. It involves balancing signal-to-noise ratio (SNR), contrast preservation, and temporal consistency across subframes, all while accounting for the specific requirements of the target vision model. Conventional exposure control focuses on low-level image statistics, while downstream vision models exploit high-level semantic information. Bridging this gap requires jointly optimizing the imaging process and the vision model for downstream task performance.

To this end, we propose a neural exposure control framework for split-pixel HDR sensors. We introduce a calibrated, differentiable split-pixel imaging simulator with a lightweight neural controller that predicts the exposure parameters for each subframe under real-camera constraints. The simulator maps scene radiance to sensor measurements while accounting for sensor noise and analog amplification curves calibrated using standard test charts (Jähne, 2020; Wei et al., 2020). In addition, we construct a realistic synthetic HDR scene-radiance dataset covering diverse autonomous-driving scenarios. We employ physically based rendering of 3D assets to generate clean scene radiance that closely reflects real-world illumination conditions. By coupling the simulator with the radiance dataset, we can evaluate exposure control for downstream vision tasks, such as object detection (Ge et al., 2021) and instance segmentation (He et al., 2017), enabling task-driven optimization of the exposure policy. As shown in Fig. 1, we further compare our imaging results with conventional multi-exposure HDR fusion methods under challenging conditions. Our method learns optimal exposure combinations and produces improved imaging and perception results. Experiments on synthetic and real-world automotive HDR benchmarks further demonstrate that our method consistently outperforms existing handcrafted and learning-based AE baselines while maintaining real-time efficiency. These results demonstrate that exposure control for split-pixel HDR sensors fundamentally benefits from jointly optimizing sensor-level capture parameters, imaging physics, and downstream perception objectives. Our contributions are summarized as follows:

- We propose a neural exposure control framework for split-pixel HDR sensors that optimizes subframe-specific exposure parameters for downstream visual perception.
- We develop a calibrated split-pixel imaging simulator and construct a realistic synthetic HDR radiance dataset to evaluate exposure control under challenging illumination.
- We validate our method on a synthetic dataset and real-world scenes, demonstrating consistent improvements over auto-exposure control and tone-mapping baselines.

2 Related Work

HDR imaging and split-pixel sensors. The dynamic range of real-world scenes often exceeds the capture capability of conventional image sensors, making a single standard exposure insufficient in challenging conditions such as tunnel entrances and exits (Kriss, 2015; Darmont, 2013;

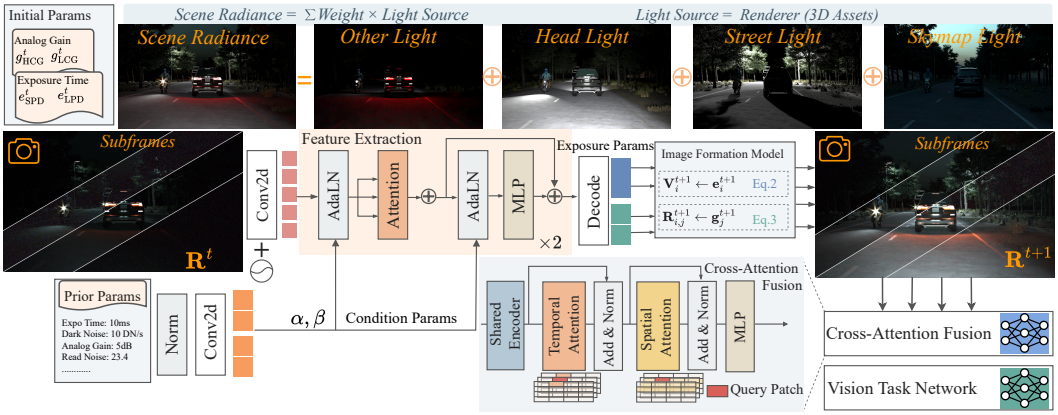

Figure 2: **Overview of our method.** We construct scene radiance as a weighted sum of rendered light components and simulate the current split-pixel captures using initial exposure parameters. Conditioned on sensor priors, the neural controller extracts visual features and predicts exposure times and analog gains for the next frame. These parameters drive the differentiable image formation model to generate four subsequent captures, which are fused through cross-attention and passed to a downstream network. The pipeline is optimized end-to-end using task supervision.

Nayar & Mitsunaga, 2000). Classical HDR imaging addresses this limitation through temporal exposure bracketing (Hasinoff et al., 2016; Li et al., 2026; Mertens et al., 2007) and multi-exposure fusion, where multiple images captured with different exposure settings are combined to recover a wider range of scene radiance (Hasinoff et al., 2010; Mertens et al., 2007; Pourreza-Shahri & Kehtarnavaz, 2015). However, because the measurements are acquired at different times, temporal fusion is inherently vulnerable to ghosting, motion artifacts, and flicker in dynamic scenes (Mertens et al., 2007; Debevec & Malik, 2023). Spatially multiplexed HDR sensors introduce a split-pixel design (Takayanagi et al., 2018; Ohta, 2020), achieving exposure bracketing by integrating photodiodes or readout circuits with different responsivities to acquire multiple low-dynamic-range subframes in a single shot. More recent designs further incorporate lateral overflow integration capacitors (LOFIC) (Fujihara et al., 2021) to improve sensor performance and have been deployed in mobile photography. Recent deep optics-based approaches have also explored single-shot HDR capture through learned optical elements (Metzler et al., 2020; Sun et al., 2020). These designs modulate the input energy distribution and extend the sensor's capture dynamic range.

However, existing AE methods apply a single shared exposure setting to all subframes, leaving their distinct saturation limits and noise characteristics unexploited. Our work bridges this gap by optimizing exposure for such spatially multiplexed sensors in challenging HDR scenarios under downstream task supervision.

Auto-exposure for downstream perception. Auto-exposure is a fundamental component of image acquisition. Traditional AE methods typically use heuristic optimization algorithms (Lagarias et al., 1998; Lu et al., 2010) to maximize image-quality metrics (Granados et al., 2013; Shim et al., 2018; Shin et al., 2019). However, these image metrics are often misaligned with downstream perception, leading to suboptimal task performance. Recent learning-based AE methods predict exposure settings from image statistics or semantic cues and optimize them through predefined exposure-control equations under supervision from downstream tasks, such as object detection (Onzon et al., 2021; 2024), stereo matching (Choi et al., 2025), depth estimation (Chang & Wetzstein, 2019), and SLAM (Zhang et al., 2024). However, these methods are not tailored to split-pixel sensors. They typically allocate exposure time and gain using predefined exposure equations (Onzon et al., 2021; Zhang et al., 2024), often favoring longer exposure times to maximize the SNR (Hasinoff et al., 2010). This strategy leaves motion blur and ghosting in dynamic scenes for post-processing to correct. Our method instead predicts exposure time and gain independently per subframe, explicitly modeling motion blur to balance exposure and gain across scenes.

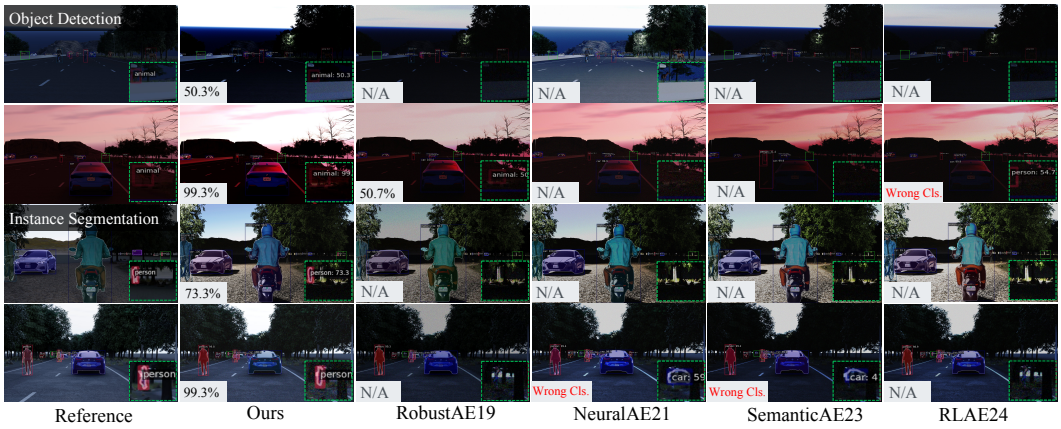

Figure 3: **Evaluation on Synthetic Dataset**. We compare different AE methods on the proposed synthetic dataset. The first two rows show object detection results, while the last two rows present instance segmentation results. Please zoom in to view confidence scores and class predictions.

HDR simulation and perception. Computational imaging workflows typically begin with a camera model and use high-fidelity scene radiance as input to simulate camera behavior, thereby enabling reliable sim-to-real transfer (Onzon et al., 2021; 2024; Kim et al., 2026). However, reproducing realistic scene radiance computationally remains challenging because real-world scenes exhibit extremely high dynamic ranges and complex physical properties. Most existing works start from captured HDR images (Xu et al., 2023; Li et al., 2025b) and treat them as approximations of scene radiance to build closed-loop simulation frameworks for computer vision perception (Kim et al., 2026) and computational photography (Li et al., 2026). However, real HDR captures inherently contain camera noise, and their dynamic range is constrained by the acquisition settings, making them less reliable for evaluating HDR perception systems. Another direction relies on rendering techniques that combine 3D assets to generate scene radiance (Liu et al., 2019; Xie et al., 2024; Choi et al., 2025). However, these datasets often rely on simplified material properties and are constrained by the rendering engine (Dosovitskiy et al., 2017), resulting in a limited output dynamic range and reduced realism and fidelity.

In this work, we construct a high-fidelity scene-radiance dataset using physically based rendering techniques to reproduce realistic scenes, as shown in Fig. 1. Building on this dataset, we couple a differentiable split-pixel imaging simulator with a neural exposure controller that optimizes the exposure parameters of each subframe, enabling end-to-end exposure control for split-pixel sensors.

3 Method

3.1 HDR Imaging Simulator

Scene Radiance Simulation. We simulate scene radiance from 3D assets using physically based rendering (Pharr & Humphreys, 2010), as illustrated in Fig. 1. To enable flexible control over the scene dynamic range, we render the radiance contribution of each light source independently and combine them as

$$\mathbf{\Phi} = \sum\nolimits_c \omega_c \cdot L_c, \tag{1}$$

where c indexes the active and background lighting components. L_c denotes the radiance contribution of component c, and ω_c is its weight. The resulting scene radiance is passed to a physically based image formation model to generate sensor captures that reproduce the imaging characteristics of real cameras. Further details are provided in the supplementary material.

Split-Pixel Image Formation Model. We model a split-pixel HDR sensor with a small photodiode (SPD) and a large photodiode (LPD) per pixel, together with high (HCG) and low (LCG) conversion-gain modes. Given the scene radiance map $\mathbf{\Phi}$, we first apply motion blur and then model the electron count $\mathbf{V}_i$ accumulated by each photodiode $i \in \{\mathrm{SPD}, \mathrm{LPD}\}$ during exposure e_i:

$$\mathbf{V}_i = \min\left(\mathcal{P}_{\mathrm{shot}}\left(\overline{\mathbf{\Phi}} \cdot e_i\right) + \mathcal{P}_{\mathrm{dark},i}\left(\boldsymbol{\mu}_i \cdot e_i\right), M_i\right), \quad \overline{\mathbf{\Phi}} = \mathbf{\Phi} * \mathbf{K}_{\mathrm{motion}}(e_i; m), \tag{2}$$

Table 1: **Quantitative Performance.** We evaluate the proposed method on our synthetic dataset with Mask R-CNN and YOLO-X. We report AP, AP_{50}, AP_{75}, and AR as evaluation metrics. The best and second-best results are marked in **bold** and with an underline, respectively.

Methods	Instance Segmentation (2017)				Object Detection (2021)				Parameters (M)	Latency (ms)
	AP^{seg}	AP^{seg}_{50}	AP^{seg}_{75}	AR^{seg}	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AR^{box}		
Direct HDR	15.77	26.97	25.41	25.84	23.03	41.55	27.18	29.89	-	-
Tone Mapping										
IANet22 (2022)	39.68	61.24	42.16	45.23	41.87	62.96	45.17	49.28	0.25	48.34
RAODNet23 (2023)	40.76	60.17	43.19	46.23	42.25	63.37	46.57	49.34	0.48	54.25
SANet25 (2025a)	38.75	60.86	39.86	45.04	35.57	55.67	39.41	43.12	0.07	15.42
Auto-Exposure										
PIDAE10 (2010)	34.75	57.08	37.61	43.09	37.51	57.96	38.15	42.45	-	10.27
FiboAE13 (2013)	36.31	58.11	38.85	43.05	39.05	61.20	45.07	45.22	-	20.59
RobustAE19 (2019)	36.81	58.61	38.61	43.59	41.42	61.98	45.63	48.25	-	44.24
NeuralAE21 (2021)	41.34	63.27	43.59	46.61	43.82	64.23	48.63	51.35	2.96	103.61
SemanticAE23 (2023)	35.79	57.75	37.22	42.50	38.12	58.92	40.44	44.79	-	58.50
RLAE24 (2024)	34.81	55.33	34.46	40.85	38.77	59.95	40.77	42.71	1.90	99.67
AdaptiveAE25 (2025)	38.04	59.96	39.17	44.14	41.34	60.40	45.06	48.74	10.25	389.59
Ours	**45.72**	**70.02**	**49.13**	**51.74**	**47.38**	**69.49**	**51.36**	**54.65**	1.89	49.43

where $\mathbf{K}_{\text{motion}}(e_i; m)$ is the motion blur PSF parameterized by exposure time e_i and motion kernel m, $*$ denotes convolution, and $\mathcal{P}(\cdot)$ denotes Poisson sampling. Here, $\boldsymbol{\mu}_i$ is the dark-current rate, and M_i is the full-well capacity.

We then amplify and digitize the accumulated signal. For each gain circuit $j \in \{\text{HCG}, \text{LCG}\}$, the resulting subframe is modeled as:

$$\mathbf{R}_{i,j} = \text{round}\left(A_j(\mathbf{V}_i; g_j) + \boldsymbol{\eta}_{\text{read},j}\right), \boldsymbol{\eta}_{\text{read},j} \sim \mathcal{N}(\mathbf{0}, \sigma_j^2), \tag{3}$$

where $A_j(\cdot; g_j)$ denotes analog amplification with gain g_j, and σ_j^2 is the readout noise variance. The rounding operation quantizes the signal to integer values, and we use a straight-through estimator to approximate its gradient during backpropagation. Finally, we apply gray-world automatic white balance (AWB) and gamma correction to all subframes. This process yields four subframes from the SPD/LPD $\times$ HCG/LCG branches, covering complementary luminance ranges.

3.2 Neural Exposure Controller

Feature Extraction. The neural exposure controller predicts the exposure times and analog gains for the next capture from the current captures and their corresponding exposure parameters. We concatenate the four subframes along the channel dimension to form $\mathbf{R}^t$ and extract two conditioned feature sequences:

$$\left(\mathbf{x}_e^t, \mathbf{x}_g^t\right) = \text{FeatureExtraction}\left(\mathbf{z}^t\right), \quad \mathbf{z}^t = \text{Tokenizer}\left(\mathbf{R}^t \downarrow\right), \tag{4}$$

where $\downarrow$ denotes downsampling to 512×512. The tokenizer comprises four convolutional layers, followed by fixed 2D sinusoidal positional encodings (Vaswani et al., 2017). The resulting tokens $\mathbf{z}^t$ initialize two feature streams, each processed by two prior-conditioned Transformer blocks to obtain the exposure-time features $\mathbf{x}_e^t$ and analog-gain features $\mathbf{x}_g^t$. We parameterize the camera settings using an exposure-time vector $\mathbf{e} = [e_{\text{SPD}}, e_{\text{LPD}}]$ and an analog-gain vector $\mathbf{g} = [g_{\text{HCG}}, g_{\text{LCG}}]$.

Prior Feature Modulation. Each branch of a split-pixel HDR sensor exhibits distinct response characteristics. We therefore incorporate sensor priors as a conditional vector, enabling the neural exposure controller to learn branch-specific behavior Kim et al. (2024). Since the sensor parameters span different numerical ranges, we first normalize and encode them before constructing the conditioning vector:

$$\begin{cases} \mathbf{p}_{\text{e}} = [\mathbf{e}, \boldsymbol{\mu}] \\ \mathbf{p}_{\text{g}} = [\mathbf{g}, \boldsymbol{\sigma}] \end{cases} \xrightarrow{\text{Norm}} \begin{cases} \mathbf{p}_{\text{e}} = [\frac{\mathbf{e}}{e_{\max}}, \frac{\boldsymbol{\mu}}{e_{\max}}] \\ \mathbf{p}_{\text{g}} = [\frac{\mathbf{g}}{g_{\max}}, \frac{\boldsymbol{\sigma}}{g_{\max}}] \end{cases}, \tag{5}$$

where the exposure prior $\mathbf{p_e}$ captures shot and dark-current noise, while the gain prior $\mathbf{p_g}$ represents analog amplification and readout noise. Since the maximum exposure time and gain are numerically comparable to the noise parameters, we directly normalize the latter. This normalization brings $\mathbf{e}$ and $\boldsymbol{\mu}$ into similar numerical ranges, stabilizing feature modulation during training.

Table 2: **Ablation Study on Feature Modulation**. The best results are highlighted in **bold**.

Method	AP	AP_{50}	AP_{75}
Instance Segmentation			
w/o Prior Modulation	43.27	66.83	43.96
w/o Norm	39.84	60.79	41.21
w Add Modulation	42.15	65.27	43.34
w Concat Modulation	44.23	67.48	46.02
Ours	**45.72**	**70.02**	**49.13**
Object Detection			
w/o Prior Modulation	46.32	66.98	50.35
w/o Norm	42.96	61.83	47.28
w Add Modulation	45.19	63.82	45.16
w Concat Modulation	46.75	66.94	50.47
Ours	**47.38**	**69.49**	**51.36**

Table 3: **Ablation Study on the Subframe Fusion**. The best results are highlighted in **bold**.

Method	AP	AP_{50}	AP_{75}
Instance Segmentation (2017)			
ExposureFusion07 (2007)	37.26	56.83	44.07
OptimalFusion10 (2010)	32.84	54.19	41.92
DeepHDR17 (2017)	42.05	62.18	43.96
NeuralFusion24 (2024)	42.66	64.80	45.22
Ours	**45.72**	**70.02**	**49.13**
Object Detection (2021)			
ExposureFusion07 (2007)	42.37	56.52	49.36
OptimalFusion10 (2010)	39.14	44.89	42.31
DeepHDR17 (2017)	43.63	65.18	46.25
NeuralFusion24 (2024)	45.46	66.83	49.97
Ours	**47.38**	**69.49**	**51.36**

We then use independent convolution layers to generate the modulation parameters from the exposure time prior $\mathbf{p}_\mathrm{e}$ and analog gain prior $\mathbf{p}_\mathrm{g}$:

$$\boldsymbol{\alpha} = \mathrm{Tanh}(\mathrm{Conv}_\alpha(\mathbf{p}_\mathrm{e}))\,, \boldsymbol{\beta} = \mathrm{Conv}_\beta(\mathbf{p}_\mathrm{g}). \tag{6}$$

Both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have twice as many channels as $\mathbf{x}^t$ and are split into exposure and gain branches: $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_\mathrm{e}, \boldsymbol{\alpha}_\mathrm{g}]$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_\mathrm{e}, \boldsymbol{\beta}_\mathrm{g}]$. Then the prior features are used in Layer Normalization (Ba et al., 2016) to adjust the visual features within the feature extraction network:

$$\mathbf{x}^t_\mathrm{e} = (1 + \boldsymbol{\alpha}_\mathrm{e}) \odot \mathrm{LayerNorm}(\mathbf{x}^t_e) + \boldsymbol{\beta}_\mathrm{e}, \quad \mathbf{x}^t_\mathrm{g} = (1 + \boldsymbol{\alpha}_\mathrm{g}) \odot \mathrm{LayerNorm}(\mathbf{x}^t_g) + \boldsymbol{\beta}_\mathrm{g}, \tag{7}$$

where $\odot$ denotes channel-wise multiplication. The residual modulation term $1 + \boldsymbol{\alpha}$ keeps the transformation close to standard layer normalization, which stabilizes training while enabling time- and gain-dependent sensor priors to adapt the visual features.

3.3 Exposure Parameter Decoding

After feature extraction, the neural controller predicts the exposure times and analog gains for the next frame. We spatially average the output features and apply separate heads for exposure time and gain decoding. We parameterize exposure times in the log domain to enforce their bounds and use a soft-argmax relaxation of discrete gain selection during training:

$$\begin{cases} \mathbf{e}^{t+1} = \exp\left\{\mathrm{Sigmoid}(-\mathbf{x}^t_\mathrm{e}) \log e_\mathrm{min} + \mathrm{Sigmoid}(\mathbf{x}^t_\mathrm{e}) \log e_\mathrm{max}\right\}, \\ \mathbf{g}^{t+1} = \arg\max_n(\tau \cdot \mathbf{y}^t_{\mathrm{g},n}) \approx \sum_{n=0}^{N} n \cdot \mathrm{Softmax}\big(\tau \cdot \mathrm{MLP}(\mathbf{x}^t_\mathrm{g})\big)_n, \end{cases} \tag{8}$$

where $\mathbf{e}^{t+1}$ is constrained to $[e_\mathrm{min}, e_\mathrm{max}]$, and $\mathbf{y}^t_\mathrm{g}$ denotes the gain logits computed from the visual features. Here, N specifies the upper bound ($N_\mathrm{HCG} = 20$ and $N_\mathrm{LCG} = 14$), and $\tau = 100$ controls the sharpness of the soft selection. The predicted exposure times and analog gains are passed to the image formation model to generate the next four subframes $\{\mathbf{R}^{t+1}_{i,j}\}$.

3.4 End-to-End Optimization

Cross-Attention Exposure Fusion. We then use cross-attention to fuse subframes with photometric inconsistencies and spatial misalignment for downstream perception. Given the next capture $\mathbf{R}^{t+1}$ generated with the predicted exposure settings, the pipeline is defined as follows:

$$(d, c, s) = \mathrm{VisionHead}(\mathrm{CrossAttFusion}(\mathbf{R}^{t+1}_{i,j})), \tag{9}$$

where d denotes the object prediction (bounding box or segmentation mask), and c and s denote the corresponding predicted class and confidence score, respectively. We use these symbols to represent the output of the vision perception network. Rather than reconstructing a fused HDR image in the

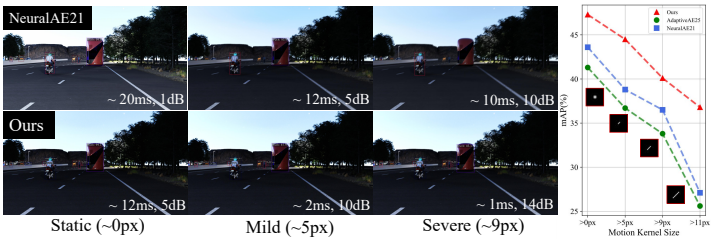

Table 4: Ablation Study on Dynamic Scenes.

Method	Static	Mild	Severe
NeuralAE21	43.67	38.82	36.53
AdaptiveAE25	41.33	36.7	33.2
Ours	**47.38**	**44.52**	**40.11**

Figure 4: **Evaluation in dynamic scenes.** Left: Visual comparisons and performance curves. Right: Quantitative comparison under static, mild-motion, and severe-motion conditions.

pixel-space, we extract features from each subframe and fuse them using *Cross-Attention Fusion*. The *Vision Head* denotes the downstream task-specific head for detection or segmentation.

As shown in Fig. 2, temporal attention aggregates complementary features from local neighborhoods across the four subframes, enabling soft correspondence under spatial misalignment. Spatial attention then captures context through shifted-window interactions. Each attention stage includes residual addition and layer normalization, and a token-wise MLP refines the resulting features.

Task Objective Loss. Our framework is trained end-to-end without explicit handcrafted image-quality losses. Instead, the neural exposure controller, differentiable simulator, and fusion module are optimized directly using downstream perception objectives:

$$\mathcal{L}_{\text{obj}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{loc}}, \quad \mathcal{L}_{\text{ins}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{mask}}. \tag{10}$$

Here, $\mathcal{L}_{\text{obj}}$ denotes the object detection loss used in YOLO-X (Ge et al., 2021), while $\mathcal{L}_{\text{ins}}$ denotes the instance segmentation loss used in Mask R-CNN (He et al., 2017). Specifically, $\mathcal{L}_{\text{cls}}$, $\mathcal{L}_{\text{loc}}$, and $\mathcal{L}_{\text{mask}}$ correspond to the classification, localization, and mask prediction losses, respectively.

4 Dataset and Training

Scene Radiance Dataset. We construct a synthetic HDR scene radiance dataset for autonomous driving, comprising 3,000 high-fidelity radiance images: 2,500 for training and 500 for validation. Following RAODNet (Xu et al., 2023), we render images directly in the linear RGB domain to avoid zipper artifacts caused by demosaicing. The dataset contains 32,000 annotated instances with bounding boxes and segmentation masks across seven classes commonly encountered in driving scenarios: person, animal, car, bus, truck, bicycle, and motorcycle. We set the scene dynamic range to 140 dB to reflect real-world radiance variations (Darmont, 2013). Unless otherwise stated, all experiments are conducted under static conditions. More details are provided in Supplementary.

Parameter Calibration. Our split-pixel simulator generates four subframes per shot, as described in Sec. 3. Following EMVA 1288 (Jähne, 2020), we calibrate dark-current levels μ_i, read-noise parameters σ_j, and analog response functions A_j using gray-patch images captured at varying reflectances and exposures with a LUCID TRI054S camera equipped with a Sony IMX490 HDR sensor (Inc., 2020) and adjustable subframe settings. Exposure time is capped at 20 ms for both photodiode units. High- and low-gain modes are sampled in 1 dB steps over $[0, 20]$ dB and $[0, 14]$ dB, respectively; gain values are converted to linear amplification factors for simulation. All simulated images have a resolution of 1080×1920 and 12-bit precision.

Implementation Details. We evaluate two downstream tasks: instance segmentation with Mask R-CNN (He et al., 2017) and object detection with YOLO-X (Ge et al., 2021). All models are trained with random flipping for data augmentation. To isolate the contribution of the proposed architecture, we retrain all learning-based baselines on our synthetic training dataset. Initial exposure parameters are set to their mid-range values: 10 ms for all exposure times, 10 dB for HCG, and 7 dB for LCG. All experiments are conducted on an NVIDIA RTX 4090 GPU.

Evaluation. For performance evaluation, we employ Average Precision (AP) at different IoU thresholds, as well as AP at IoU thresholds of 0.5 (AP_{50}), 0.75 (AP_{75}), and Average Recall (AR) for both bounding box and segmentation mask evaluations (Zhang et al., 2022; Zhu et al., 2021).

5 Evaluation

Quantitative Evaluation. We compare our method with several categories of exposure control approaches on the synthetic dataset. PIDAE10 (Lu et al., 2010) adapts a PID controller (Lu et al.,

Table 5: **Ablation Study on Exposure Parameter Control.** Bounding-box detection performance across motion levels defined at a reference exposure time. The best results are shown in **bold**.

Method	Static ($\sim$ 0 px)			Mild ($\sim$ 5 px)			Severe ($\sim$ 9 px)		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Shared	44.73	66.32	48.13	42.43	64.56	45.42	40.01	62.57	42.49
Indep. time	45.78	67.10	49.32	43.04	64.97	46.06	39.38	62.52	43.11
Indep. gain	44.68	66.94	48.57	42.23	65.02	45.65	38.92	61.60	41.18
Indep. both (Ours)	**47.38**	**69.49**	**51.36**	**44.52**	**67.21**	**47.98**	**40.11**	**63.32**	**43.91**

2010) to adjust exposure using image-gradient-based metrics. Optimization-based methods (Gupta et al., 2013; Shin et al., 2019; Tedla et al., 2023) cast exposure control as an image-quality optimization problem. Specifically, FiboAE13 (Gupta et al., 2013) maximizes SNR-related metrics with shutter-change constraints to reduce ghosting artifacts, RobustAE19 (Shin et al., 2019) reweights multiple image-quality metrics for challenging scenes, and SemanticAE23 (Tedla et al., 2023) prioritizes semantic regions during exposure estimation. These handcrafted objectives are derived from prior heuristics and may not align well with downstream perceptual tasks. Learning-based methods, including NeuralAE21 (Onzon et al., 2021) and NeuralFusion24 (Onzon et al., 2024), predict exposure parameters from learned image features via predefined exposure equations under task-driven supervision. RL-based methods, such as RLAE24 (Zhang et al., 2024) and AdaptiveAE25 (Xu et al., 2025), further optimize exposure policies through sequential decision-making without requiring differentiable camera models. We compare our method with tone mapping approaches, evaluating IANet (Liu et al., 2022), RAODNet (Xu et al., 2023), and SANet (Li et al., 2025a) on images captured with intermediate exposure settings for a fair comparison. Smaller exposure budgets produce noisy images, whereas larger budgets risk overexposure. Although these methods compress dynamic range and enhance local contrast for downstream perception, their performance remains limited by the SNR of input images. Our method achieves the highest mask and box AP, outperforming the strongest baseline, NeuralAE21, by 4.38 and 3.56 points, respectively.

Qualitative Evaluation. Fig. 3 illustrates the visual comparisons with competing methods. RobustAE19 (Shin et al., 2019) and SemanticAE23 (Tedla et al., 2023) optimize handcrafted image-quality metrics, inadvertently suppressing contrast in task-critical regions. NeuralAE21 (Onzon et al., 2021) and RLAE24 (Zhang et al., 2024) rely on predefined exposure equations that favor long exposures to maximize SNR, introducing noticeable motion blur in dynamic scenes. Our method instead allocates exposure time and analog gain independently for each subframe, yielding sharper object boundaries and higher-confidence predictions under challenging illumination.

6 Ablation Study

Analysis of Feature Modulation. The feature modulation strategies are compared in Tab. 2. Direct feature addition performs worst among the evaluated strategies, while concatenation yields only marginal improvements over the baseline. These results suggest that naive injection of sensor priors is insufficient and highlight the importance of explicit conditional modulation for adapting visual features to sensor-specific imaging conditions. Removing parameter normalization degrades performance, likely due to scale mismatch between sensor priors and network weights.

Analysis of Realistic Simulation. Fig. 5 shows real captures of a standard color checker and their simulated counterparts generated by our simulation model under different exposure settings. We further evaluate the color difference between real and simulated samples using the CIEDE2000 ΔE metric (Luo et al., 2001). To avoid color misalignment introduced by the ISP pipeline, we compute this metric only on gray patches. The

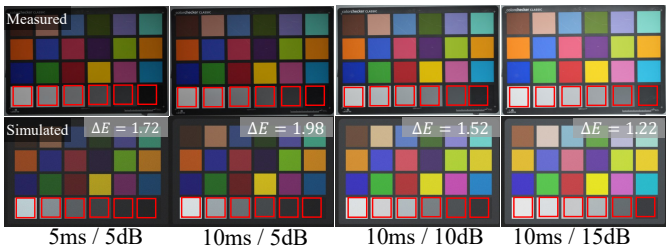

Figure 5: **Simulation Realism on a Standard Test Chart.** We compare real measurements and simulated captures of the standard chart and evaluate ΔE on gray patches.

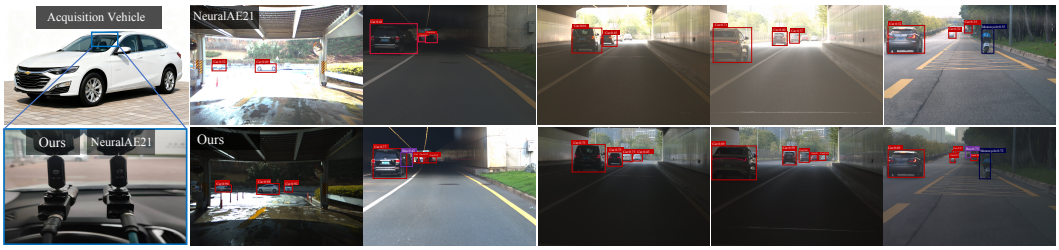

Figure 6: **Evaluation on a Hardware Prototype.** We conduct a side-by-side comparison between the proposed method and NeuralAE21 using real-time vehicle-capture sequences. Our method balances exposure across dark interiors and sunlit exteriors, adapting robustly to rapid illumination changes and preserving task-relevant contrast.

quantitative results show that the discrepancy between simulated and real images remains within an acceptable range, with $\Delta E \leq 2$. These results provide quantitative validation of the simulation.

Analysis of Subframe Fusion Strategy. Tab. 3 compares the proposed fusion strategy with representative methods. ExposureFusion07 (Mertens et al., 2007) and OptimalFusion10 (Hasinoff et al., 2010) fuse subframes in image space without motion compensation, making them susceptible to motion artifacts. DeepHDR17 (Kalantari & Ramamoorthi, 2017) and NeuralFusion24 (Onzon et al., 2024) directly fuse subframes in feature space, but over/under-exposure and sensor noise may still propagate into the extracted features. Our method independently controls the exposure parameters to capture complementary subframes, modulates features using sensor priors, performs feature-space fusion, and is optimized by the downstream task.

Analysis of Model Complexity. Table 1 compares the parameter counts and inference latencies of the evaluated methods. Compared with prior learning-based AE methods (Onzon et al., 2021), our method uses a tokenizer to compress spatial information and feature modulation to encode sensor-specific priors. Together, these components preserve global statistics and local semantic cues while improving performance and reducing computational cost. The exposure controller has approximately 1M parameters and a GPU inference latency of 50 ms. The cross-attention fusion module has approximately 10M parameters, can replace a conventional perception backbone (Onzon et al., 2024), and achieves a favorable performance–latency trade-off.

Analysis of Dynamic Scene. We further evaluate the proposed method under varying motion conditions. As shown in Fig. 4 and Tab. 4, our model predicts shorter exposure times and higher gains for scenes with greater motion, reducing motion blur while maintaining image brightness. In contrast, the comparison methods (Onzon et al., 2021; Xu et al., 2025) rely on predefined exposure equations, which limit their flexibility in allocating the exposure budget in dynamic scenes.

Analysis of Exposure Parameter Control. We evaluate exposure parameter control under different motion conditions in Tab. 5. The *shared* strategy uses a common exposure time and analog gain across all subframes, whereas *independent time / gain* allows only the corresponding parameter to vary across subframes. Independent time control improves AP over shared control under static and mild motion but underperforms it under severe motion. Our method independently controls both parameters and achieves the highest AP across all motion levels. These results highlight the benefit of independently controlling both parameters to balance motion blur and noise: longer exposures increase motion blur, whereas high gain used to compensate for short exposures can amplify noise.

7 Real-World Evaluation

We compare the proposed method with NeuralAE21 (Onzon et al., 2021) using a side-by-side capture rig mounted on a vehicle windshield, as shown in Fig. 6. The rig consists of two independently operating cameras that acquire image streams from separate imagers with aligned fields of view. Each stream is processed online by the Yolo-X detector (Ge et al., 2021) deployed on NVIDIA Jetson AGX Orin platforms. To ensure a fair comparison, we use the same object detector architecture and post-processing pipeline for both methods; only the AE strategy is changed.

8 CONCLUSION

In this work, we presented a task-driven neural exposure control method for spatially multiplexed HDR sensors. By jointly optimizing the exposure time and analog gain of each subframe through a differentiable HDR imaging simulator, the proposed method enables more flexible dynamic range allocation and noise control than conventional exposure-equation-based approaches. We also introduced a photorealistic HDR radiance dataset paired with the proposed simulator, enabling controlled exposure simulation and benchmarking for auto-exposure methods. Experiments on both synthetic and real-world automotive HDR datasets demonstrated that our method consistently outperforms traditional and learned auto-exposure baselines while maintaining real-time operation.

AI USE STATEMENT

AI tools (such as Large Language Models and Agent Tools) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

REPRODUCIBILITY STATEMENT

In this work, we introduce a new dataset and perform detailed calibration of real camera devices to enable reliable sim-to-real transfer. The paper provides a comprehensive description of the dataset simulation pipeline and the model design. To ensure reproducibility and support future research, we will publicly release the dataset, the camera calibration parameters and procedures, the simulation pipeline, and the model implementation upon acceptance. We believe these resources will promote progress in the community.

REFERENCES

T Asatsuma, Y Sakano, S Iida, M Takami, I Yoshiba, N Ohba, H Mizuno, T Oka, K Yamaguchi, and Suzuki. Sub-pixel architecture of cmos image sensor achieving over 120 db dynamic range with less motion artifact characteristics. In *Proceedings of the 2019 International Image Sensor Workshop*, volume 1, 2019.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Francesco Banterle, Alessandro Artusi, Kurt Debattista, and Alan Chalmers. *Advanced high dynamic range imaging*. AK Peters/CRC Press, 2017.

Michael S Brown and SJ Kim. Understanding the in-camera image processing pipeline for computer vision. In *IEEE International Conference on Computer Vision (ICCV)-Tutorial*, volume 3, pp. 1–354, 2019.

Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10193–10202, 2019.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *arXiv*, 2021.

Juhyung Choi, Jinnyeong Kim, Seokjun Choi, Jinwoo Lee, Samuel Brucker, Mario Bijelic, Felix Heide, and Seung-Hwan Baek. Dual exposure stereo for extended dynamic range 3d imaging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6283–6293, 2025.

Arnaud Darmont. High dynamic range imaging: sensors and architectures, 2013.

Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 643–652. ACM New York, NY, USA, 2023.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16, 2017.

Yasuyuki Fujihara, Maasa Murata, Shota Nakayama, Rihito Kuroda, and Shigetoshi Sugawa. An over 120 db single exposure wide dynamic range cmos image sensor with two-stage lateral overflow integration capacitor. *IEEE Transactions on Electron Devices*, 68(1):152–157, 2021.

Orazio Gallo, Natasha Gelfandz, Wei-Chao Chen, Marius Tico, and Kari Pulli. Artifact-free high dynamic range imaging. In *2009 IEEE International conference on computational photography (ICCP)*, pp. 1–7. IEEE, 2009.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Miguel Granados, Kwang In Kim, James Tompkin, and Christian Theobalt. Automatic noise modeling for ghost-free hdr reconstruction. *ACM Transactions on Graphics (TOG)*, 32(6):1–10, 2013.

Mohit Gupta, Daisuke Iso, and Shree K Nayar. Fibonacci exposure bracketing for high dynamic range imaging. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1473–1480, 2013.

Samuel W Hasinoff, Frédo Durand, and William T Freeman. Noise-optimal capture for high dynamic range photography. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 553–560. IEEE, 2010.

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.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

SONY Inc. Imx490,sony's cmos sensor for automotive. `https://www.sony-semicon.com/files/62/pdf/p-15_IMX490.pdf`, 2020. Accessed: 2020-11-22.

Bernd Jähne. Release 4 of the emva 1288 standard: Adaption and extension to modern image sensors. *(No Title)*, 2020.

Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017.

Jaewon Kim, Sol Namkung, and Dongsuk Jeon. Splithdr: Saturation-aware hdr recovery and denoising for real-time detection. In *European Conference on Computer Vision*, pp. 337–355. Springer, 2026.

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*, pp. 26067–26076, 2024.

Michael Kriss. *Handbook of digital imaging*. John Wiley & Sons, 2015.

Jeffrey C Lagarias, James A Reeds, Margaret H Wright, and Paul E Wright. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on optimization*, 9(1): 112–147, 1998.

Baiang Li, Ruyu Yan, Ethan Tseng, Zhoutong Zhang, Adam Finkelstein, Jiawen Chen, and Felix Heide. Lucky high dynamic range imaging. *ACM Transactions on Graphics (TOG)*, 2026.

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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

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 Computer Vision and Pattern Recognition Conference*, pp. 8859–8868, 2025b.

Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1792–1800, 2022.

Zhenyi Liu, Minghao Shen, Jiaqi Zhang, Shuangting Liu, Henryk Blasinski, Trisha Lian, and Brian Wandell. A system for generating complex physically accurate sensor images for automotive applications. *arXiv preprint arXiv:1902.04258*, 2019.

Huimin Lu, Hui Zhang, Shaowu Yang, and Zhiqiang Zheng. Camera parameters auto-adjusting technique for robust robot vision. In *2010 IEEE International Conference on Robotics and Automation*, pp. 1518–1523. IEEE, 2010.

M Ronnier Luo, Guihua Cui, and Bryan Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 26(5):340–350, 2001.

Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pp. 382–390. IEEE, 2007.

Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1375–1385, 2020.

Ali Mosleh, Avinash Sharma, Emmanuel Onzon, Fahim Mannan, Nicolas Robidoux, and Felix Heide. Hardware-in-the-loop end-to-end optimization of camera image processing pipelines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7529–7538, 2020.

S.K. Nayar and T. Mitsunaga. High dynamic range imaging: spatially varying pixel exposures. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, volume 1, pp. 472–479 vol.1, 2000. doi: 10.1109/CVPR.2000.855857.

Jun Ohta. *Smart CMOS image sensors and applications*. CRC press, 2020.

Emmanuel Onzon, Fahim Mannan, and Felix Heide. Neural auto-exposure for high-dynamic range object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7710–7720, 2021.

Emmanuel Onzon, Maximilian Bömer, Fahim Mannan, and Felix Heide. Neural exposure fusion for high-dynamic range object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17564–17573, 2024.

Matt Pharr and Greg Humphreys. *Physically Based Rendering, Second Edition: From Theory To Implementation*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2010. ISBN 0123750792.

Reza Pourreza-Shahri and Nasser Kehtarnavaz. Exposure bracketing via automatic exposure selection. In *2015 IEEE international conference on image processing (ICIP)*, pp. 320–323. IEEE, 2015.

Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Trans. Graph.*, 21(3):267–276, 2002.

Inwook Shim, Tae-Hyun Oh, Joon-Young Lee, Jinwook Choi, Dong-Geol Choi, and In So Kweon. Gradient-based camera exposure control for outdoor mobile platforms. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(6):1569–1583, 2018.

Ukcheol Shin, Jinsun Park, Gyumin Shim, Francois Rameau, and In So Kweon. Camera exposure control for robust robot vision with noise-aware image quality assessment. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1165–1172. IEEE, 2019.

Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1386–1396, 2020.

Isao Takayanagi, Norio Yoshimura, Kazuya Mori, Shinichiro Matsuo, Shunsuke Tanaka, Hirofumi Abe, Naoto Yasuda, Kenichiro Ishikawa, Shunsuke Okura, Shinji Ohsawa, et al. An over 90 db intra-scene single-exposure dynamic range cmos image sensor using a 3.0 μm triple-gain pixel fabricated in a standard bsi process. *Sensors*, 18(1):203, 2018.

SaiKiran Tedla, Beixuan Yang, and Michael Brown. Examining autoexposure for challenging scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13076–13085, 2023.

Justin Tomasi, Brandon Wagstaff, Steven L Waslander, and Jonathan Kelly. Learned camera gain and exposure control for improved visual feature detection and matching. *IEEE Robotics and Automation Letters*, 6(2):2028–2035, 2021.

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.

Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020.

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*, pp. 2758–2767, 2020.

Trygve Willassen, Johannes Solhusvik, Robert Johansson, Sohrab Yaghmai, Howard Rhodes, Sohei Manabe, Duli Mao, Zhiqiang Lin, Dajiang Yang, Orkun Cellek, et al. A 1280$\times$ 1080 4.2 μm split-diode pixel hdr sensor in 110 nm bsi cmos process. In *Proceedings of the International Image Sensor Workshop, Vaals, The Netherlands*, pp. 8–11, 2015.

Yiming Xie, Henglu Wei, Zhenyi Liu, Xiaoyu Wang, and Xiangyang Ji. Synfog: a photo-realistic synthetic fog dataset based on end-to-end imaging simulation for advancing real-world defogging in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21763–21772, 2024.

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*, pp. 13384–13393, 2023.

Tianyi Xu, Fan Zhang, Boxin Shi, Tianfan Xue, and Yujin Wang. Adaptiveae: An adaptive exposure strategy for hdr capturing in dynamic scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Shuyang Zhang, Jinhao He, Yilong Zhu, Jin Wu, and Jie Yuan. Efficient camera exposure control for visual odometry via deep reinforcement learning. *IEEE Robotics and Automation Letters*, 2024.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

Supplementary Material

This supplementary material provides additional implementation details for the proposed task-driven neural exposure controller:

- Details on HDR Radiance Simulation and Annotation.
- Details on Image Formation Model Parameter Calibration.
- Details on Neural Exposure Controller Architecture.
- Details on Additional Ablation Studies.
- Details on Real-World Evaluation.

A HDR Radiance Simulation and Annotation

A.1 Synthetic HDR Radiance Generation

Scene construction. The dataset contains 3,000 HDR scene radiance maps, with 2,500 scenes for training and 500 for validation, consistent with the split used in the main paper. The scenes depict rural roads with vegetation, vehicles, pedestrians, cyclists, and animals. We independently render four lighting components: headlights, streetlights, other active lights, and sky illumination. Other active lights include vehicle taillights and small emissive sources. Recombining these components produces different illumination conditions while retaining the same scene geometry and object annotations.

Light-group recombination. Driving scenes contain multiple incoherent light sources, such as sky illumination, vehicle headlights, streetlights, and taillights. We therefore combine radiant energy rather than optical wavefront amplitudes. For each color channel $c \in \{R, G, B\}$, the scene radiance map is generated as

$$\mathbf{\Phi} = \sum_{k \in \mathcal{K}} \omega_k \, \mathbf{\Phi}_k, \qquad \mathbf{\Phi}_k = [\mathbf{\Phi}_{k,R}, \mathbf{\Phi}_{k,G}, \mathbf{\Phi}_{k,B}], \qquad \mathcal{K} = \{\text{head}, \text{street}, \text{other}, \text{sky}\}, \tag{11}$$

where $\mathbf{\Phi}_{k,c}$ denotes the rendered radiance map of light group k for color channel c, and ω_k is the corresponding radiance-combination weight. We use representative wavelengths $\{460, 520, 630\}\,\text{nm}$ for the RGB channels, which avoids additional demosaicing during dataset synthesis and matches the simulator input $\mathbf{\Phi}$ used in the main paper.

Active-light intensities. For each active-light group, we match its maximum rendered luminance $L_k^{\max}$ to a sampled target $\widehat{L}_k$:

$$\omega_k = \frac{\widehat{L}_k}{L_k^{\max} + \epsilon}, \qquad k \in \{\text{head}, \text{street}, \text{other}\}, \tag{12}$$

where $\epsilon > 0$ prevents division by zero. The targets are sampled as

$$\begin{aligned} &\widehat{L}_{\text{head}} = 10^4 \max(20u_{\text{h}}, 2), && \widehat{L}_{\text{street}} = 10^4 \max(5u_{\text{s}}, 1), \\ &\widehat{L}_{\text{other}} = 10^{-4} \max(100u_{\text{o}}, 4)\widehat{L}_{\text{head}}, && u_{\text{h}}, u_{\text{s}}, u_{\text{o}} \sim \mathcal{N}(0, 1). \end{aligned} \tag{13}$$

The headlight and streetlight targets have lower bounds of 2×10^4 and $10^4\,\text{cd/m}^2$, respectively. The other-light target is at least 0.04% of the headlight target. These distributions are lower-clamped Gaussian samples and do not impose finite upper bounds. Peak-based calibration assumes that the relevant active source is visible in the rendered scene.

Ambient illumination and dynamic range. Sky illumination controls the ambient level. We denote the linear contrast control parameter by D_{scene}; under the convention $\text{DR}_{\text{dB}} = 20 \log_{10} D_{\text{scene}}$, a setting of $140\,\text{dB}$ corresponds to $D_{\text{scene}} = 10^7$. The sky target and weight are:

$$\widehat{L}_{\text{sky}} = \frac{2 \times 10^5}{D_{\text{scene}}}, \qquad \omega_{\text{sky}} = \frac{\widehat{L}_{\text{sky}}}{P_{30}(\mathbf{\Phi}_{\text{sky}}) + \epsilon}, \tag{14}$$

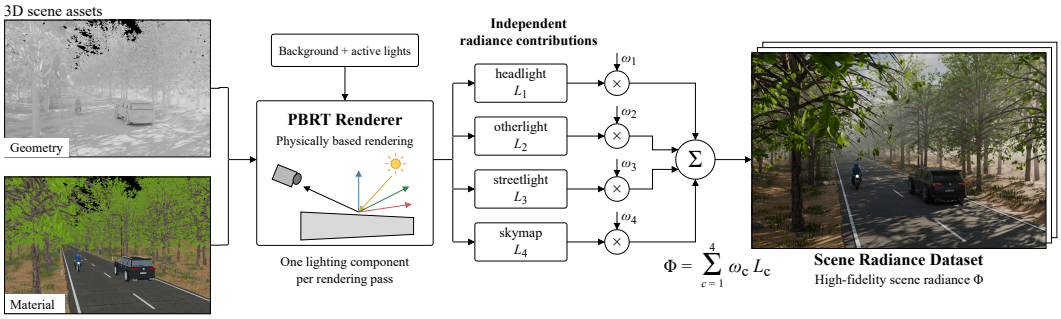

Figure 7: **HDR scene radiance generation.** Physically based rendering produces separate contributions from four light groups. Weighted recombination controls their relative intensities and the ambient illumination of the resulting scene radiance map.



Figure 8: **Light-group decomposition.** The combined scene radiance is shown alongside the independently rendered headlight, other-light, streetlight, and sky contributions. Tone mapping is used here only to visualize their large intensity differences.

where $P_q(\mathbf{\Phi})$ denotes the qth percentile of the luminance of $\mathbf{\Phi}$. Using the 30th percentile reduces sensitivity to isolated bright pixels. A larger D_{scene} suppresses ambient illumination relative to active sources. This parameter controls the lighting mixture; the realized pixelwise dynamic range also depends on scene content.

Global radiance normalization. We rescale the combined map to a reference median luminance $L_{\text{ref}} = 10$ in the simulator's radiance scale:

$$\mathbf{\Phi} = \frac{L_{\text{ref}}}{P_{50}(\mathbf{\Phi}_{\text{mix}}) + \epsilon} \mathbf{\Phi}_{\text{mix}}. \tag{15}$$

This is a global multiplicative normalization, rather than a lower bound on pixel luminance. It preserves within-scene radiance ratios; the physical luminance targets above specify the relative light-group scaling before this normalization. Fig. 8 illustrates the individual contributions and their combination.

A.2 Annotation Protocol

The dataset contains bounding-box and instance-mask annotations for seven classes: **Person**, **Animal**, **Car**, **Bus**, **Truck**, **Bicycle**, and **Motorcycle**. The Car class covers light and medium-sized vehicles, while Bus and Truck cover larger road vehicles and construction vehicles. Bicycle includes bicycles and similar light transportation objects. Person includes pedestrians and riders, with each individual labeled separately in crowded scenes. Animal includes medium-sized animals such as deer, cows, and dogs. Bounding boxes follow the visible object extents, and objects smaller than 25×25 pixels are excluded.

For the real-world evaluation sequences, tone-mapped views (Reinhard et al., 2002) are used to improve visibility during annotation. A vision-language model provides coarse initial labels, which human annotators verify and correct for spatial and temporal consistency. These annotation views serve a different purpose from the simple ISP used for exposure evaluation, described in Appendix Sec. A.4.

A.3 Dataset Statistics

Approximately 50% of the training scenes depict nighttime conditions, 30% daytime conditions, and 20% dusk conditions. Front-view and rear-view driving scenarios include vehicles viewed from

Table 6: **Dataset Statistics.** Category statistics on the number and proportion of samples in each category.

Split	Person	Animal	Car	Bus	Truck	Bicycle	Motorcycle	Total
Train	8,650 (32.52%)	2,816 (10.59%)	3,690 (13.87%)	1,650 (6.20%)	2,299 (8.64%)	3,519 (13.23%)	3,976 (14.95%)	26,600
Val	1,808 (33.48%)	564 (10.44%)	948 (17.56%)	160 (2.96%)	408 (7.56%)	512 (9.48%)	1,000 (18.52%)	5,400

different directions, exposing both headlights and taillights. The dataset contains 32,000 annotated instances: 26,600 in the training split and 5,400 in the validation split. The class distribution is given in Tab. 6.

A.4 Image Formation and ISP Pipeline

Image generation pipeline. Scene radiance, sensor formation, and image processing are separate stages. First, the independently rendered light groups are combined into $\mathbf{\Phi}$, as described in Sec. A.1. The split-pixel image formation model then converts this radiance into four sensor subframes $\mathbf{R}_{i,j}$ under the selected exposure times and analog gains (Eqs. 2 and 3 in the main paper). A simple ISP produces the corresponding RGB images:

$$\mathbf{\Phi} \xrightarrow{\text{photoelectric conversion}} \mathbf{V}_i \xrightarrow{\text{gain, readout noise, and quantization}} \mathbf{R}_{i,j} \xrightarrow[\gamma=2.2]{\text{gray-world AWB}} \mathbf{I}_{i,j} \xrightarrow{\text{lossless storage}} \text{PNG}. \tag{16}$$

Here, $\mathbf{R}_{i,j}$ denotes the quantized sensor subframe defined in Eq. 3 and $\mathbf{I}_{i,j}$ denotes the RGB image after white balance and gamma correction. The sensor model accounts for exposure-dependent photon collection, motion blur, full-well saturation, analog amplification, noise, and quantization. Its calibration is detailed in Sec. B.

Simple ISP. The ISP consists of gray-world automatic white balance (AWB), followed by a fixed gamma encoding. Gray-world AWB adjusts the relative RGB channel gains using the assumption of a neutral average scene color. For a normalized, white-balanced channel value x, gamma encoding is $\Gamma(x) = \text{clip}(x, 0, 1)^{1/2.2}$. The output retains 12 bits of effective sample precision and is stored losslessly in PNG files. Here, 12 bits refers to signal precision rather than a native PNG sample depth; a 16-bit PNG container can preserve these samples. PNG compression is a storage operation and does not introduce an additional tone curve.

Adaptive Tone Mapping. Auto-exposure control and tone mapping partly overlap in their effect on image brightness and the visibility of dark or bright regions. An adaptive tone mapper can brighten shadows or compress highlights after acquisition, thereby changing the apparent benefit of an exposure setting. However, the two operate at different stages: exposure time changes photon collection and motion blur, and analog gain changes signal amplification before digitization; tone mapping remaps already acquired values and cannot recover information lost to sensor saturation or insufficient signal. To make the contribution of exposure control easier to assess, we use the simple ISP above, with a fixed gamma curve, rather than an additional adaptive global or local tone-mapping stage. The AWB procedure and gamma setting are kept consistent across the AE methods being compared. Tone mapping used for annotations and illustrative radiance figures is separate from this evaluation pipeline.

B IMAGE FORMATION MODEL PARAMETER CALIBRATION

We calibrate the noise and response parameters of the SPD/LPD photodiodes and HCG/LCG readout circuits. The same parameterization is used for simulated training and synthetic evaluation. For photodiode $i \in \{\mathrm{SPD}, \mathrm{LPD}\}$ and circuit $j \in \{\mathrm{HCG}, \mathrm{LCG}\}$, the model is

$$\begin{aligned} \overline{\mathbf{\Phi}}_i &= \mathbf{\Phi} * \mathbf{K}_{\mathrm{motion}}(e_i; m), \\ \mathbf{V}_i &= \min\left\{\mathcal{P}_{\mathrm{shot}}(\overline{\mathbf{\Phi}}_i e_i) + \mathcal{P}_{\mathrm{dark},i}(\boldsymbol{\mu}_i e_i),\ M_i\right\}, \\ \mathbf{R}_{i,j} &= \mathrm{round}(A_j(\mathbf{V}_i; g_j) + \boldsymbol{\eta}_{\mathrm{read},j}), \qquad \boldsymbol{\eta}_{\mathrm{read},j} \sim \mathcal{N}(\mathbf{0}, \sigma_j^2). \end{aligned} \tag{17}$$

Here, $\boldsymbol{\mu}_i$ is a dark-current *rate*, M_i is the full-well capacity, and A_j is the analog response. As in the main paper, the scale of $\overline{\mathbf{\Phi}}_i$ includes the conversion required for $\overline{\mathbf{\Phi}}_i e_i$ to represent an expected photoelectron count. Shot noise and dark-current noise are modeled before amplification; readout noise is expressed in the output signal domain. We omit temperature dependence in this calibration model. The complete calibrated parameter set will be released upon acceptance.

B.1 SHOT NOISE AND DARK-CURRENT CALIBRATION

For a pixel receiving an expected photoelectron rate $\bar{\phi}_i$, the shot and dark electron counts satisfy

$$v_{\mathrm{shot}} \sim \mathcal{P}(\bar{\phi}_i e_i), \qquad v_{\mathrm{dark}} \sim \mathcal{P}(\mu_i e_i). \tag{18}$$

Both terms have variance equal to their mean. In particular, μ_i does not grow with exposure time; the expected accumulated dark charge $\mu_i e_i$ does. We capture dark frames at multiple exposure times and fit their mean and variance for a fixed readout setting. In electron-equivalent units, these relations are

$$m_{\mathrm{dark},i}(e_i) = m_{0,i} + \mu_i e_i, \tag{19}$$

$$v_{\mathrm{dark},i}(e_i) = v_{0,i} + \mu_i e_i, \tag{20}$$

where $m_{0,i}$ and $v_{0,i}$ are the exposure-independent offset and variance, respectively. The latter describes the baseline of the measured dark frames, rather than an additional Poisson dark-current term in Eq. 17. Conversion from digital measurements to electron-equivalent units uses the calibrated system gain below.

B.2 READOUT NOISE AND SYSTEM GAIN

Readout noise is modeled as a signal-independent Gaussian term,

$$\boldsymbol{\eta}_{\mathrm{read},j} \sim \mathcal{N}(\mathbf{0}, \sigma_j^2), \qquad j \in \{\mathrm{HCG}, \mathrm{LCG}\}. \tag{21}$$

Subtracting the mean dark signal removes its offset but does not remove its shot-noise variance. Readout-noise estimation therefore accounts for the exposure-dependent dark-current contribution, using the exposure-independent variance component of the calibration model. The resulting standard deviations form the prior vector $\boldsymbol{\sigma}$ used by the controller.

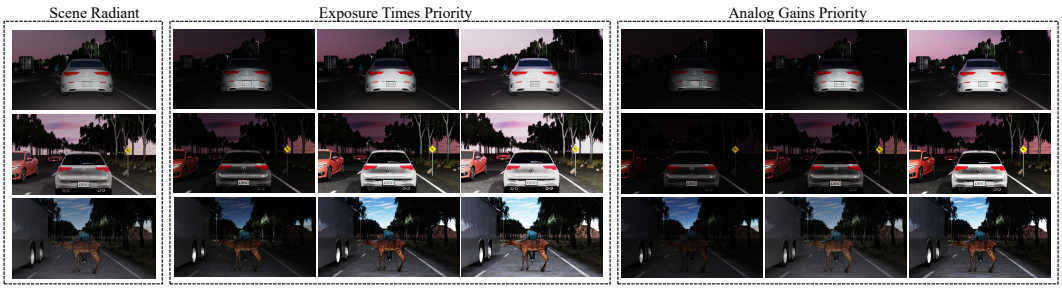

Figure 9: **Effect of exposure settings in the simulator.** Exposure-time sweeps use $\{5, 10, 20\}$ ms at a fixed gain of 3 dB; analog-gain sweeps use $\{0, 5, 14\}$ dB at a fixed exposure time of 5 ms. The scene radiance and image formation model are held fixed while the indicated capture parameter is varied.

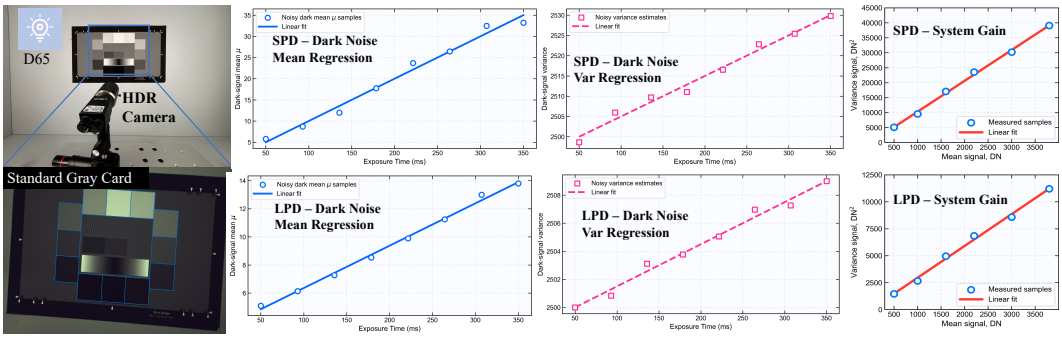

Figure 10: **Sensor calibration.** Left: standard-chart acquisition setup under controlled illumination. Right: dark-signal mean, dark-signal variance, and system-gain fits for the SPD and LPD branches. Captures at multiple exposure times and gain settings characterize the branch responses.

The system gain $K_{i,j}$ maps electrons to digital numbers (DN) in the linear, unsaturated response region. It is distinct from photoelectric conversion. For a fixed gain setting, let $m_{i,j}$ and $v_{i,j}$ denote the measured mean and temporal variance of a gray patch. The photon-transfer relation is

$$v_{i,j} - v_{\mathrm{dark},i,j} \simeq K_{i,j}\big(m_{i,j} - m_{\mathrm{dark},i,j}\big), \tag{22}$$

where the dark statistics are matched to the exposure and gain setting. Thus, the slope of dark-corrected variance against dark-corrected mean estimates $K_{i,j}$. Gray-patch captures at different exposures and gain settings characterize the response curves, while the dark measurements constrain the noise terms. Fig. 10 shows the calibration setup and fits.

Table 7: **Architecture of the neural exposure controller.** Four subframes are downsampled to $H \times W$ and encoded into two C-dimensional conditioned feature streams to predict exposure parameters for the next frame.

Name	Layer setting	Output dimension
Input	–	$4C_{\text{in}} \times H \times W$
Tokenization	Conv2d $[3 \times 3, 2, C/4]$ + GELU	$\frac{C}{4} \times \frac{H}{2} \times \frac{W}{2}$
	Conv2d $[3 \times 3, 2, C/2]$ + GELU	$\frac{C}{2} \times \frac{H}{4} \times \frac{W}{4}$
	Conv2d $[3 \times 3, 2, C]$ + GELU	$C \times \frac{H}{8} \times \frac{W}{8}$
	Conv2d $[3 \times 3, 2, C]$	$C \times \frac{H}{16} \times \frac{W}{16}$
	Flatten $[H/16, W/16]$	$N \times C$
	2D Sinusoidal PE $[C]$	$N \times C$
Prior Encoding	Conv2d $[1 \times 1, 1, 2C]$ + Tanh	$\boldsymbol{\alpha} : 2C$
	Conv2d $[1 \times 1, 1, 2C]$	$\boldsymbol{\beta} : 2C$
Encoder1	Transformer Block $[C, 4, 4]$	$N \times C$
Encoder2	Transformer Block $[C, 4, 4]$	$N \times C$
Global Pooling	Global AvgPool $[N]$	C
Exposure Head	Linear $[C, E]$ + Sigmoid	E
Gain Head	MLP $[C, C, K]$	K
	Softmax $[\tau = 100]$	$K_{\text{HCG}},\ K_{\text{LCG}}$

C Neural Exposure Controller Architecture

The neural exposure controller predicts the next capture settings from the current four subframes and sensor priors. A separate fusion module processes the resulting subframes for downstream perception. The two networks are summarized in Tab. 7 and Tab. 8. Throughout this section, the fixed subframe order is SPD–HCG, SPD–LCG, LPD–HCG, and LPD–LCG; tensor shapes omit the batch dimension.

C.1 Neural Exposure Controller

Input tokenization. The four RGB subframes are resized to 512×512 and concatenated along the channel axis, giving a $12 \times 512 \times 512$ input in channel-first layout. Four 3×3 convolutions use stride 2, padding 1, and output widths of 32, 64, 128, and 128 channels. The first three use GELU; the last has no activation. The resulting $128 \times 32 \times 32$ feature map is flattened into 1024 tokens with 128 features each. Fixed 2D sinusoidal position encodings are added before the token sequence is copied into the exposure and gain streams.

Prior encoding and adaptive normalization. The normalized exposure prior contains two exposure times and two dark-current parameters; the normalized gain prior contains two analog gains and two readout-noise parameters. Separate 1×1 convolutions embed these $4 \times 1 \times 1$ inputs into 256-dimensional vectors. Each vector is split into two 128-dimensional parts, one for each stream. The exposure embeddings pass through tanh and receive a unit offset to define the normalization scales; the gain embeddings define the shifts. Each token is normalized across its features without learned affine parameters, then scaled and shifted using these vectors. Both streams thus receive both prior types, with conditioning shared across tokens.

Conditioned Transformer blocks. Each stream passes through two Transformer blocks. Each block applies adaptive normalization before global self-attention and before a token-wise MLP, with a residual connection around each sublayer. Attention uses four heads with 32 channels each. The MLP expands the feature width from 128 to 512, applies GELU, and projects back to 128. The streams share attention and MLP weights at each depth; the two depths use distinct weights. Each stream reuses its conditioning vectors across both blocks and retains a 1024×128 token sequence. Averaging over tokens yields a 128-dimensional feature for each stream.

Table 8: **Architecture of cross-attention exposure fusion.** The four-subframe hierarchy produces multiscale features through an FPN output adapter.

Name	Layer setting	Output dimension
Input	–	$S \times C_{\text{in}} \times H \times W$
Shared Encoder	Patchify $[p \times p]$	$S \times N_1 \times p^2 C_{\text{in}}$
	MLP $[p^2 C_{\text{in}}, C, C]$	$S \times N_1 \times C$
	LayerNorm $[C]$	$S \times N_1 \times C$
	2D Sinusoidal PE $[C]$	$S \times N_1 \times C$
Encoder1	ST Transformer $[2, 2, 2]$	$S \times N_1 \times C$
Readout1	MLP $[SC, C, C]$	$C \times H_1 \times W_1$
Patch Merging1	Linear $[4C, 2C]$	$S \times N_2 \times 2C$
	LayerNorm $[2C]$	$S \times N_2 \times 2C$
Encoder2	ST Transformer $[2, 4, 2]$	$S \times N_2 \times 2C$
Readout2	MLP $[2SC, 2C, 2C]$	$2C \times H_2 \times W_2$
Patch Merging2	Linear $[8C, 4C]$	$S \times N_3 \times 4C$
	LayerNorm $[4C]$	$S \times N_3 \times 4C$
Encoder3	ST Transformer $[2, 8, 2]$	$S \times N_3 \times 4C$
Readout3	MLP $[4SC, 4C, 4C]$	$4C \times H_3 \times W_3$
Patch Merging3	Linear $[16C, 8C]$	$S \times N_4 \times 8C$
	LayerNorm $[8C]$	$S \times N_4 \times 8C$
Encoder4	ST Transformer $[2, 16, 2]$	$S \times N_4 \times 8C$
Readout4	MLP $[8SC, 8C, 8C]$	$8C \times H_4 \times W_4$
Feature Pyramid	Conv2d $[1 \times 1, 1, C_{\text{P}}]$	$C_{\text{P}} \times H_\ell \times W_\ell$
	Upsample $[2]$ + Add	$C_{\text{P}} \times H_\ell \times W_\ell$
	Conv2d $[3 \times 3, 1, C_{\text{P}}]$	$C_{\text{P}} \times H_\ell \times W_\ell$

Exposure and gain decoding. A linear exposure head produces two logits, one per photodiode. A sigmoid followed by log-space interpolation maps each logit to an exposure time between the configured positive minimum and 20 ms. The gain head uses a 128-unit hidden layer with GELU and a 36-unit output layer. The logits cover 21 HCG gain candidates from 0 to 20 dB and 15 LCG candidates from 0 to 14 dB, both in 1 dB steps. For each set, the logits are multiplied by 100 before softmax. Training uses the probability-weighted mean of each candidate set; camera control selects the most probable candidate. The simulator converts gains from dB to linear amplification factors.

C.2 Cross-Attention Exposure Fusion

Architecture overview. As summarized in Table 8, the fusion module comprises a shared patch encoder, four spatio-temporal stages with fusion readouts, and a feature pyramid network (FPN) output adapter. It processes four RGB subframes at the perception resolution, padded to multiples of 32. The shared encoder maps non-overlapping 4×4 patches to 64-dimensional tokens through an MLP, layer normalization, and fixed 2D sinusoidal position encoding. The four stages use 64, 128, 256, and 512 channels, with two blocks per stage. Between stages, 2×2 patch merging halves the spatial resolution and doubles the channel width while preserving the four streams.

Cross-subframe and spatial attention. Each block averages the four streams to form a common query base. Queries attend separately to keys and values within a 5×5 neighborhood in each source stream, allowing local spatial alignment across exposures. Attention uses normalized queries and keys, a learned scale, and relative-position bias. Each stream then undergoes self-attention within 8×8 windows; the second block shifts the windows by four tokens along both axes. Residual connections and layer normalization follow both attention steps. A token-wise MLP with GELU doubles the channel width and projects it back. Weights are shared across subframes, with distinct parameters at each depth.

Multiscale fusion readout. At each stage, a two-layer MLP with GELU projects the channel-concatenated streams back to the stage width. Adding the stream mean produces a fused feature map, while the four streams continue to the next stage. The FPN adapter applies 1×1 lateral projections, top-down upsampling and addition, and 3×3 output convolutions. The resulting P2–P5 maps have 256 channels at strides 4, 8, 16, and 32. This pyramid replaces the original backbone and neck outputs at the input to the downstream prediction heads. The following standard interfaces illustrate this connection.

Connection to YOLOX. For YOLOX (Ge et al., 2021), P3–P5 feed directly into the decoupled detection head at strides 8, 16, and 32. Each input stem is configured to accept 256 channels and feeds the standard classification, box-regression, and objectness branches.

Connection to Mask R-CNN. For FPN-based Mask R-CNN (He et al., 2017), P2–P5 supply the region proposal network (RPN) and the RoIAlign inputs for the box and mask heads. Max-pooling P5 with stride 2 adds a stride-64 P6 map for the RPN. Thus, the connection is at the pyramid feature inputs, before proposal generation and RoIAlign.

D Additional Ablation Studies

D.1 Evaluation with Additional Perception Models

We evaluate the exposure controller with DINO-Det (Zhang et al., 2022) and Deformable DETR (Zhu et al., 2021) for object detection, and SOLOv2 (Wang et al., 2020) and Mask2Former (Cheng et al., 2021) for instance segmentation. The comparison in Tab. 10 evaluates our method against NeuralAE21 (Onzon et al., 2021) under each perception architecture. For object detection, box AP increases from 43.4 to 48.9 with DINO-Det and from 50.2 to 53.0 with Deformable DETR, yielding gains of 5.5 and 2.8 points, respectively. For instance segmentation, mask AP increases from 47.2 to 51.3 with SOLOv2 and from 49.8 to 52.2 with Mask2Former, corresponding to gains of 4.1 and 2.4 points. Improvements are also consistent across AP_{50}, AP_{75}, and AR for all four models. These results support the effectiveness of the proposed method across the evaluated detection and segmentation architectures.

E Real-World Evaluation

Two LUCID Triton TRI054S-CC HDR cameras are mounted side by side behind the vehicle windshield. Each camera supplies a separate image stream to an NVIDIA Jetson AGX Orin platform running YOLO-X (Ge et al., 2021), as described in the main paper. The two streams use the same detector architecture and post-processing pipeline; the exposure-control method differs. Tab. 9 lists the acquisition hardware.

The camera interface supports frame-by-frame exposure updates. Changes in exposure can therefore produce visible brightness fluctuations in the recorded sequences. The attached video demo, `video_demo1.mp4`, illustrates the capture setup and exposure behavior, complementing the frame-level perception results in the main paper with a qualitative view of temporal exposure changes.

Table 9: **Hardware configuration of the acquisition platform.**

Item #	Component	Quantity	Model / specification
1	HDR camera	2	LUCID Triton TRI054S-CC
2	Objective lens	2	Edmund Optics, $25\,\mathrm{mm}$, $f/2.5$
3	Computing platform	2	NVIDIA Jetson AGX Platform

Table 10: **Performance comparison on advanced perception models.** Metrics are computed on bounding boxes for object detection and instance masks for segmentation. The better result within each model pair is shown in bold.

Perception model	Exposure controller	AP	AP_{50}	AP_{75}	AR
Object detection					
DINO-Det (2022)	NeuralAE21 (2021)	43.40	66.70	47.40	58.80
	Ours	**48.90**	**72.90**	**53.00**	**64.30**
Deform. DETR (2021)	NeuralAE21 (2021)	50.20	73.60	57.50	64.70
	Ours	**53.00**	**77.10**	**61.00**	**67.60**
Instance segmentation					
SOLOv2 (2020)	NeuralAE21 (2021)	47.20	71.60	48.90	57.30
	Ours	**51.30**	**76.90**	**54.20**	**61.20**
Mask2Former (2021)	NeuralAE21 (2021)	49.80	77.10	51.70	59.30
	Ours	**52.20**	**78.50**	**56.40**	**62.70**