SplitHDR: Saturation-Aware HDR Recovery and Denoising for Real-Time Detection

Jaewon Kim*, Sol Namkung*, and Dongsuk Jeon

Seoul National University, Seoul, South Korea
{jaewon.j.kim,nksol0405,djeon1}@snu.ac.kr

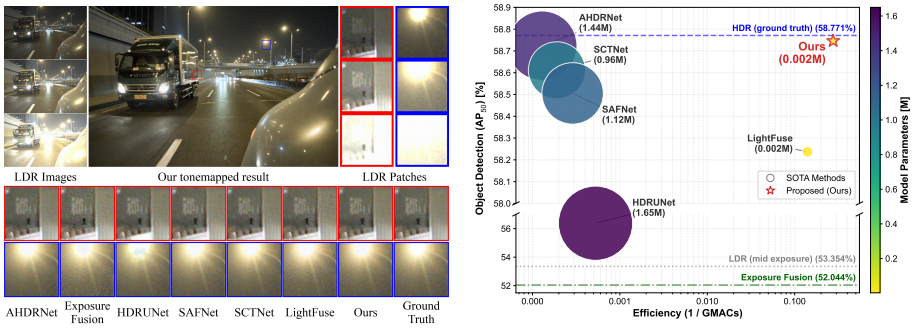
Abstract. High Dynamic Range (HDR) imaging is indispensable for autonomous driving to ensure robust perception under changing lighting conditions, such as tunnel exits or night driving. However, existing learning-based HDR methods suffer from excessive computational costs and latency, making them impractical for real-time edge deployment. In this paper, we propose **SplitHDR**, an extremely lightweight HDR framework tailored for machine vision. Leveraging advanced sensor technologies designed to minimize temporal displacement, our approach avoids relying on heavy alignment modules and introduces a luminance-aware dynamic inference mechanism. Specifically, the model constructs a saturation mask based on input intensity, efficiently routing saturated regions to a multi-frame fusion mode and unsaturated regions to a lightweight denoising mode. This task decomposition ensures that computational resources are allocated only where necessary. Experimental results demonstrate that our method achieves state-of-the-art efficiency with only 0.002M parameters and 3.63 GMACs, reducing the parameter count by over 99% compared to existing multi-exposure models. Despite the extreme compression, the HDR images reconstructed by our approach effectively preserve critical structural details, yielding highly competitive object detection performance crucial for safety-critical autonomous systems.

Keywords: HDR imaging · Object detection · Autonomous driving · Vision for edge devices

1 Introduction

Autonomous driving relies on real-time visual perception, specifically robust object detection and recognition of vehicles, pedestrians, and traffic-control elements, to ensure safe planning and control [20, 38]. However, perception must remain reliable under extreme illumination changes, such as tunnel exits, nighttime streets with strong headlight glare, or scenes that simultaneously contain very bright highlights and dark shadows. In such conditions, conventional

* Equal contribution.



(a) Comparison on ROD dataset [39] **(b)** Performance vs. efficiency

Fig. 1: Representative results and performance-efficiency comparisons. (a) Visual comparison on the ROD dataset [39]. Our method effectively suppresses noise in low-light regions and preserves highlight details across high dynamic range areas (see ROI crops). (b) Ours achieves state-of-the-art object detection (AP_{50}) with significantly lower computational cost (2.47k params, 3.63 GMACs) than prior HDR baselines.

low-dynamic-range (LDR) imaging frequently produces saturated highlights and under-exposed regions, discarding radiance cues that are critical for downstream detection and tracking [2].

HDR imaging is a natural remedy, traditionally achieved by recovering an HDR radiance map from LDR images. Recently, deep learning-based approaches have achieved impressive visual quality by learning to align and fuse exposure brackets in dynamic scenes [40], and even by leveraging generative priors such as diffusion for extreme exposure gaps [7]. However, these improvements often come with substantial computational overhead, prohibitive under the strict latency constraints and limited computing budgets of automotive edge platforms [24,38].

In parallel, sensor-side HDR has advanced rapidly in automotive imaging. Modern automotive CMOS sensors provide on-sensor HDR using specialized readout and split-pixel architectures, eliminating motion artifacts compared to conventional multi-frame HDR [18,33]. However, these sensor-level solutions still face inherent limitations, such as SNR drops at gain transition and amplified noise in under-exposed regions.

These limitations underscore a critical requirement: HDR for autonomous driving must be optimized for strict latency constraints while prioritizing robust downstream perception over photorealistic reconstruction. Moreover, not all pixels require the same treatment. In typical driving frames, large areas are reasonably exposed, while failures are concentrated in (i) **over-exposed saturated highlights** where texture details are lost due to clipping, and (ii) **noisy low- and mid-exposed regions** where discriminative cues are weakened. This motivates a region-adaptive strategy that allocates computation where it most benefits perception.

In this work, we propose an **ultra-lightweight HDR pipeline employing a divide-and-conquer strategy**. Our approach decomposes the problem into

two complementary sub-tasks: (1) **HDR recovery for saturated regions** and (2) **denoising and enhancement for unsaturated regions**. Then, they are applied selectively via a saturation-aware guidance map. We evaluate the proposed method on driving-oriented HDR data and report performance metrics for both reconstruction and detection accuracy under a fixed tone mapping and image signal processing (ISP) pipeline. As shown in Fig. 1b, our model is significantly more efficient (e.g., $\sim$668$\times$ fewer parameters and $\sim$2,130$\times$ fewer computations than other state-of-the-art multi-exposure HDR networks) while maintaining competitive reconstruction quality and detection robustness (Fig. 1a).

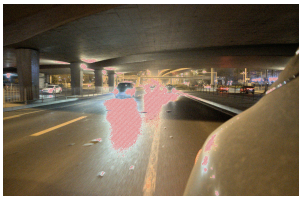
2 Related Work

HDR Image Reconstruction. Recovering the high dynamic range typically involves merging bracketed exposures. Classic HDR methods reconstruct a radiance map from these exposures using rule-based algorithms [8], typically followed by a tone mapping step for display. Meanwhile, exposure fusion [29] provides a practical alternative by blending bracketed LDR frames into a tone-mapped HDR image without explicit radiance map reconstruction. However, these conventional approaches often fail to reconstruct missing details in saturated or noisy regions. To address these limitations, recent studies have extensively explored deep learning-based approaches. Single-image methods [6, 9, 26] attempt to suppress noise in under-exposed regions and recover clipped highlights, but often lack radiometric accuracy as they rely on estimating missing information without physical reference. For high-fidelity HDR reconstruction, multi-exposure methods employ complex modules ranging from attention [40] to generative priors [7], yet these incur immense computational overhead.

HDR Sensors for Automotive Applications. To avoid the massive computational cost of compensating for misalignment between multi-exposure frames, modern automotive CMOS image sensors have shifted towards advanced architectures, such as multiple pixel gain [23, 34], multi-tap charge-splitting [11], lateral overflow integration capacitor (LOFIC) [1, 13], and split photodiodes [3, 17, 35]. These technologies enable the simultaneous capture of varying exposures within an identical integration time, physically eliminating macroscopic motion artifacts. However, simple blending after sensor acquisition introduces new bottlenecks such as amplified shot noise in under-exposed regions and SNR drops at hard-switching gain transitions [36]. Aligning with this trend, our framework assumes a synchronized multi-exposure input stack, strategically allocating the computational budget to resolve these inherent limitations.

Lightweight HDR Models for Edge Device Deployment. Processing high-resolution, high-frame-rate image streams in real-time embedded vision systems demands highly optimized models for hardware accelerators (e.g., FPGAs or ASICs). While efficient object detection algorithms [14, 31] have been

(a) Day (mid-exposure LDR)



(b) Night (mid-exposure LDR)

	Unsaturated	Saturated
All	83.94% (±7.44%)	16.06% (±7.44%)
Day	79.80% (±7.83%)	20.20% (±7.83%)
Night	88.08% (±3.88%)	11.92% (±3.88%)

(c) Saturation ratio

Fig. 2: Saturation in mid-exposure LDR frames. Saturated regions are visualized with red hatching, showing that saturation occupies only a fraction of the image and varies by scene (day vs. night).

accelerated on FPGAs with reported throughput exceeding 100 FPS [10, 25], lightweight models for the upstream HDR fusion remain scarce. Although a few hardware-acceleration studies for HDR exist [4,32], they are confined to conventional rule-based methods. For learning-based approaches, previous work such as LightFuse [27] proposed a dual-exposure fusion network for FPGA implementation, reporting 33 FPS on an NVIDIA Tesla V100 GPU. However, the absence of a less-noisy mid-exposure frame limits denoising effectiveness, which degrades reconstruction quality and yields weaker downstream object detection performance. Our approach preserves efficiency comparable to LightFuse, while providing superior HDR fusion performance. Moreover, our method achieves or exceeds the detection accuracy of heavier state-of-the-art approaches.

3 SplitHDR

3.1 Motivation: Saturation Is Sparse, but Fusion Is Expensive

HDR reconstruction is primarily needed in clipped regions: once the sensor saturates, pixel values no longer increase with scene radiance and must be recovered by fusing multiple exposures. In practice, however, saturation is typically bounded and localized rather than pervasive. This is because exposure-control methods [16, 19, 30] explicitly penalize saturation, choosing exposure by balancing highlight clipping against noise and reconstruction error in the remaining well-exposed regions. Moreover, HDR bracketing strategies commonly include a user- or system-defined tolerance on the maximum percentage of underexposed or saturated pixels [28].

In driving scenes, extreme highlights (e.g., sun glints, headlights) can make clipped pixels non-negligible. Nevertheless, unsaturated pixels still occupy the majority of the image in our driving HDR dataset (mean 83.9% with std. 7.4%, Fig. 2). This leads to a clear inefficiency: conventional fusion networks spend most of their compute on regions that are already well-exposed. Furthermore, exposure importance is highly region-dependent (Fig. 3). In unsaturated areas, the long exposure provides the best SNR and the highest PSNR to the HDR ground truth, while in saturated areas the long exposure is clipped and the

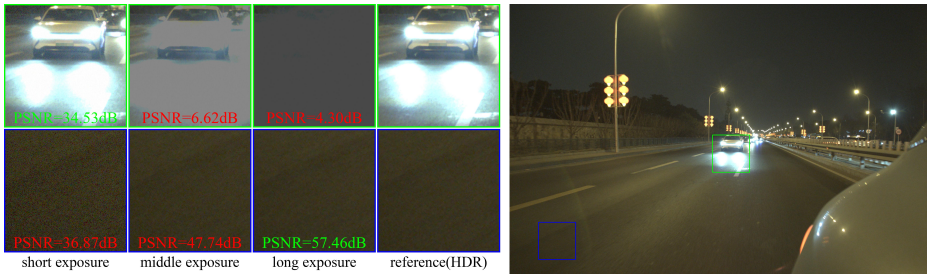


Fig. 3: Exposure-dependent fidelity across regions. Right: Saturated (green box) and unsaturated (blue box) ROIs. **Left:** Linear PSNR comparison across exposures. The short exposure yields the best PSNR for saturated areas, whereas the long exposure is optimal for unsaturated areas.

short exposure retains the most useful details. These observations motivate a region-aware design that (i) partitions processing based on local characteristics and (ii) prioritizes efficiency for the unsaturated majority and focuses compute on saturated regions.

3.2 Backbone Architecture and Dual-Mode Operation

Three-Exposure Interface and Residual Learning. Our lightweight backbone is inspired by LightFuse [27], which originally performs dual-exposure fusion using two sub-networks: GlobalNet utilizing a U-Net-like structure constructed with depthwise and depthwise-separable convolutions, and DetailNet composed of a sequence of point-wise convolutions operating on full-resolution features. As shown in Fig. 4a, we extend the input interface from dual to three-exposure stack. Specifically, we utilize the mid-exposure frame as a reference via a global skip connection, as it is typically well-exposed for the majority of pixels (Fig. 2).

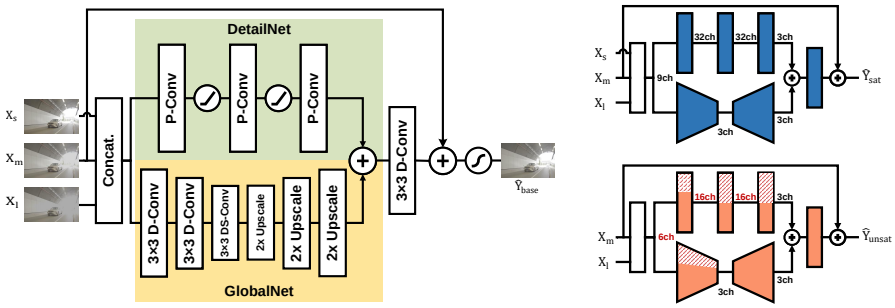
Let (s, m, l) denote short, mid, and long LDR frames in linear space and let $\mathcal{T}(\cdot)$ be a differentiable range-compression operator (e.g., μ-tonemap) used for training stability:

$$X_s = \mathcal{T}(s), \quad X_m = \mathcal{T}(m), \quad X_l = \mathcal{T}(l). \tag{1}$$

The backbone predicts a residual correction around the mid-exposure frame:

$$\hat{Y}_{\text{base}} = X_m + \Delta(X_s, X_m, X_l), \tag{2}$$

where $\Delta(\cdot)$ is produced by combining the global and detail branches. This architecture preserves the scene structure of the mid-exposure and improves optimization stability under a limited parameter budget. Consequently, the network can focus on learning the necessary residual corrections, recovering clipped highlights and enhancing shadowed regions relative to the reference.



(a) Architecture of the lightweight fusion backbone **(b)** Region-aware mode configuration

Fig. 4: Backbone architecture and region-adaptive parameters. (a) The backbone architecture with a dual-branch structure and global skip connection for residual learning. **(b)** The model dynamically switches configurations: saturated mode fuses all inputs using the full backbone, while unsaturated mode bypasses X_s, halving intermediate channel widths (red hatching) to save compute.

Task-Specialized Dual-Mode Operation. Building on the lightweight backbone above, SplitHDR operates in two task-specialized configurations for **unsaturated** and **saturated** regions, respectively. For **unsaturated** regions, we remove the short-exposure dependency (i.e., the short frame is not fed to the network in this mode). Excluding this input not only prevents the lowest-SNR frame from propagating unnecessary noise, but also enables substantial channel reduction (e.g., $9 \rightarrow 6$ channels for GlobalNet and $32 \rightarrow 16$ channels for DetailNet) by disabling the hatched components in Fig. 4b. As a result, the network processes the majority unsaturated area at a drastically reduced cost. For **saturated** regions, the full backbone is activated, and the short exposure is used to recover lost details in over-exposed regions. This conditional design yields two compute profiles: the saturated mode corresponds to the full backbone of 1.8k parameters, whereas the unsaturated mode has substantially fewer parameters (0.67k) due to the reduced feature width. Since unsaturated pixels dominate in practice (Fig. 2), SplitHDR spends most of its budget on the cheap path and *allocates the heavier fusion path only to saturated regions where it is needed.*

3.3 Region-Aware Routing for Conditional Execution

To enable dual-mode operation, SplitHDR uses a soft saturation mask as guidance. The mask is used during training to learn two task-specialized parameter sets, and during inference to realize conditional execution by region. Based on the mask value α, the image is categorized into three distinct regions: unsaturated ($\alpha = 0$), saturated ($\alpha = 1$), and a soft transition region ($0 < \alpha < 1$). The unsaturated region, which constitutes the majority of the image, is processed using configuration with reduced channels. Conversely, the saturated region activates the full backbone configuration to recover clipped details. To prevent boundary artifacts and ensure a visually smooth transition, regions with in-

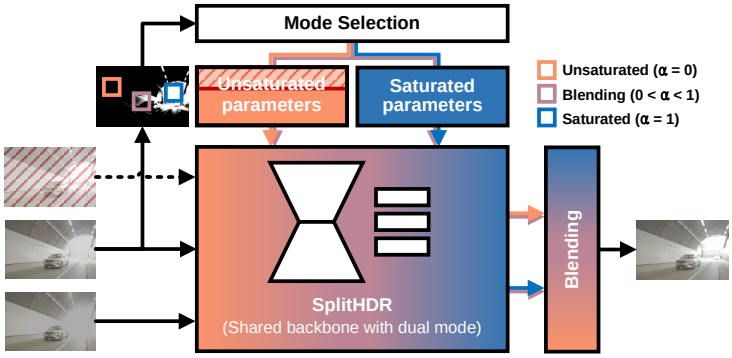


Fig. 5: Overview of region-aware dual-mode execution. SplitHDR dynamically switches parameter sets based on the saturation mask α. By selectively executing computations across saturated, unsaturated, and blending regions, the network drastically reduces redundant operations while preserving high-fidelity reconstruction.

termediate mask values are evaluated by both configurations and subsequently blended according to α. This strategy avoids executing the expensive saturated configuration across the entire frame, *focusing the processing only on tiles that contain saturation* (Fig. 5).

Saturation Mask Generation. In our preprocessing, the mid exposure is represented in a *linear* domain but normalized to the *short-exposure reference scale*. As a result, even if the original mid frame clips at value 1 in its own sensor scale, its maximum representable value in our mid input is bounded by the exposure ratio r, which is defined as

$$r = \frac{t_s}{t_m}, \tag{3}$$

since intensities are scaled down to match the short reference. To ensure consistency with the network input, we apply the same range compression operator $\mathcal{T}(\cdot)$ used for training when comparing luminance values to thresholds.

We first compute a luminance proxy from the mid frame in the network-input domain:

$$Y_m = \frac{R_m + 2G_m + B_m}{4}, \tag{4}$$

where (R_m, G_m, B_m) are the channels of the range-compressed mid input. We then define upper and lower thresholds around the saturation boundary:

$$\tau_{upper} = \mathcal{T}(r), \qquad \tau_{lower} = 0.9\,\mathcal{T}(r), \tag{5}$$

where τ_{upper} corresponds to the *maximum* mid value under short-reference normalization (i.e., confidently saturated), and τ_{lower} marks the onset of saturation.

Finally, we form a soft transition mask $\alpha \in [0, 1]$ via a linear ramp with clamping:

$$\alpha = \text{clamp}\left(\frac{Y_m - \tau_{lower}}{\tau_{upper} - \tau_{lower} + \varepsilon}, 0, 1 \right). \tag{6}$$

The transition band $[\tau_{lower}, \tau_{upper}]$ avoids a hard binary decision: boundary pixels receive gradients from both modes during training and are alpha-blended during inference, yielding smoother transition around saturation edges.

Exclusive Masked Supervision. Instead of relying on a single network to address the conflicting objectives of noise suppression in shadows and texture recovery in highlights, a divide-and-conquer strategy is adopted in the training stage. As shown in Fig. 5, we enforce specialization of the two execution modes by applying an *exclusive* pixel-wise training signal. In other words, the unsaturated mode is trained primarily on unsaturated pixels, and the saturated mode is trained primarily on saturated pixels, using the soft mask α.

We denote the two mode outputs as $\hat{Y}_{\text{unsat}}$ (unsaturated mode) and $\hat{Y}_{\text{sat}}$ (saturated mode). Let Y denote the supervision target in the same domain as the network output. We use a ℓ_1 reconstruction loss and weight it per pixel:

$$\mathcal{L}_{\text{unsat}} = \frac{\sum_p (1 - \alpha_p) \, \|\hat{Y}_{\text{unsat},p} - Y_p\|_1}{\sum_p (1 - \alpha_p) + \varepsilon}, \qquad \mathcal{L}_{\text{sat}} = \frac{\sum_p \alpha_p \, \|\hat{Y}_{\text{sat},p} - Y_p\|_1}{\sum_p \alpha_p + \varepsilon}. \tag{7}$$

Then, we optimize the total objective as

$$\mathcal{L} = \lambda_{\text{unsat}} \mathcal{L}_{\text{unsat}} + \lambda_{\text{sat}} \mathcal{L}_{\text{sat}}. \tag{8}$$

Here, pixels with $\alpha_p \approx 0$ (unsaturated) contribute almost exclusively to $\mathcal{L}_{\text{unsat}}$, while pixels with $\alpha_p \approx 1$ (saturated) contribute almost exclusively to $\mathcal{L}_{\text{sat}}$. The transition band of α softly shares supervision near the boundary, which reduces artifacts around saturation edges.

Conditional Inference Execution. At inference time, we synthesize a unified HDR prediction by combining the two mode outputs using the soft mask. First, we partition the image into tiles and determine the required execution modes based on its average saturation ratio within each tile (e.g., saturated tiles whose mean mask value satisfies $\frac{1}{|T|} \sum_{p \in T} \alpha_p \geq \gamma$, where we set $\gamma = 0.01$ (1%)). Next, tiles are processed using their corresponding execution modes: the unsaturated configuration is invoked for unsaturated tiles, the saturated configuration for saturated tiles, and both for transition tiles to enable seamless blending. Finally, the outputs are blended pixel-wise according to the soft mask α to produce the final HDR prediction $\hat{Y}$:

$$\hat{Y} = (1 - \alpha) \odot \hat{Y}_{\text{unsat}} + \alpha \odot \hat{Y}_{\text{sat}}, \tag{9}$$

where $\odot$ denotes element-wise multiplication. This tile-wise conditional execution avoids running the expensive saturated configuration on unsaturated tiles, yielding substantial compute savings in practice.

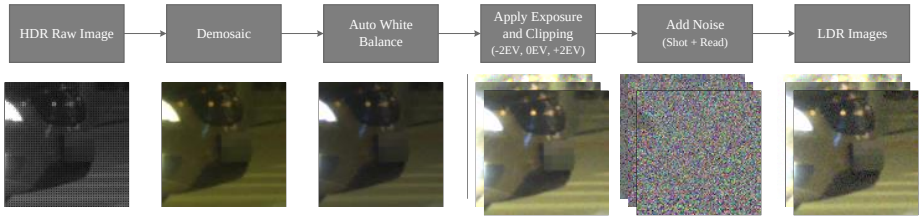


Fig. 6: ISP pipeline for synthetic multi-exposure LDR generation. We process 24-bit HDR Bayer raw images into linear RGB HDR images via demosaicing and auto white balance. We then generate three LDR brackets through exposure scaling, and inject noise to simulate realistic sensor measurements.

4 Experiments

4.1 Dataset and Implementation Details

Dataset. We validate our method using a multi-exposure fusion dataset synthesized from the ROD benchmark [39], comprising 16,089 training and 4,000 testing samples. It provides 24-bit Bayer raw sensor data captured in day and night driving scenes, annotated with six traffic-related categories. As illustrated in Fig. 6, we simulate the physical acquisition process by generating three-exposure stacks with 2-stop intervals (i.e., ΔEV$=2$, corresponding to a $4\times$ exposure ratio between adjacent frames) from the Bayer raw ground truth. To ensure realistic sensor degradation, we inject heteroscedastic Gaussian noise calibrated to account for both signal-dependent shot noise and signal-independent read noise, following standard Poisson-Gaussian noise modeling [5, 12]. Full details on preprocessing can be found in our supplementary material.

Implementation Details. The proposed method is implemented using PyTorch. The model is trained on the entire training set, which consists of 16,089 exposure stacks with 2880×1856 resolution. The network weights are initialized using the He initialization [15] and optimized via the Adam optimizer [21] with parameters $\beta_1 = 0.9, \beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate is fixed at 1×10^{-3} throughout training. We adopt exhaustive tiling: each stack is partitioned into non-overlapping 256×256 tiles (77 tiles per stack), and all tiles are used as training samples, resulting in 1,238,853 patch samples in total (309,714 optimization steps with batch size 4). Geometric augmentations, including random horizontal flips, vertical flips, and rotations, are applied during training.

Model for Downstream Detector. Assuming real-time processing for autonomous driving scenarios, we employ YOLOX-Tiny [14] to evaluate the downstream performance. The detection head is reconfigured to predict the six traffic-related categories defined in the ROD benchmark. To ensure the benchmark is not biased toward any specific HDR method, the model is trained from scratch

Table 1: Comparison of different methods in terms of parameters, MACs, HDR reconstruction, and object detection performance

Method	Params(k)	GMACs	HDR Reconstruction				Object Detection			
			PSNR-μ	PSNR-L	SSIM-μ	SSIM-L	AP	AP50	AP75	AR
HDR (ground truth)	-	-	-	-	-	-	38.34	58.77	40.18	47.83
LDR (mid)	-	-	30.41	25.89	0.865	0.962	34.24	53.35	36.13	43.67
Exposure Fusion [29]	-	-	25.01	37.47	0.631	0.930	33.94	52.04	38.82	43.27
AHDRNet [40]	1441.28	7733.12	47.37	50.98	0.994	0.997	**38.38**	58.72	40.19	47.84
HDRUNet [6]	1651.49	1898.31	37.19	28.55	0.961	0.968	36.36	56.38	38.14	45.78
SCTNet [37]	961.48	5256.57	**48.28**	**51.21**	**0.995**	**0.998**	38.33	58.61	40.12	47.78
SAFNet [22]	1120.85	3473.70	41.23	49.21	0.950	0.991	38.16	58.51	39.98	47.63
LightFuse [27]	**1.57**	7.13	40.67	38.82	0.958	0.975	37.88	58.24	39.75	47.20
Ours	2.47	**3.63**	44.96	46.12	0.985	0.994	**38.38**	**58.75**	**40.31**	**47.87**

(a) Day scene. **(b)** Night scene.

Fig. 7: Qualitative comparison results (day scene and night scene)

for 300 epochs on the tone-mapped HDR ground truth rather than the outputs of any particular method. All inputs are resized to 1280×1280 during both training and testing to preserve fine-grained spatial details for small object detection.

Evaluation Metrics. For quantitative comparison, we calculate two distinct sets of metrics. For HDR reconstruction quality, we report PSNR and SSIM on both the linear HDR domain and the μ-tonemapped domain. For downstream detection performance, we adopt Average Precision (AP) and Average Recall (AR) over all Intersection-over-Union (IoU) thresholds. Furthermore, AP values at specific IoU thresholds of 0.5 (AP_{50}) and 0.75 (AP_{75}) are provided to analyze detection accuracy at varying degrees of strictness. For model complexity, we report the total number of parameters and MACs calculated at a fixed input resolution of our dataset.

4.2 Comparisons with Prior Work

We compare SplitHDR with state-of-the-art HDR fusion methods in terms of HDR quality and downstream object detection, as summarized in Table 1 and Fig. 7. Our benchmarks include Exposure Fusion [29] as a non-learning baseline,

Table 2: Object detection performance comparison with state-of-the-art models in terms of model size and scene-wise detection performance

Input image	Params(k)	GMACs	All		Day scene		Night scene	
			AP	AR	AP	AR	AP	AR
HDR (ground truth)	-	-	38.34	47.83	38.57	48.44	38.26	47.10
LDR (mid)	-	-	34.24	43.67	37.92	47.58	31.96	40.80
Exposure Fusion [29]	-	-	33.94	43.27	35.44	45.09	32.85	41.60
AHDRNet [40]	1441.28	7733.12	**38.38**	47.84	38.63	48.52	**38.28**	47.06
HDRUNet [6]	1651.49	1898.31	36.36	45.78	38.33	48.21	35.26	43.99
SCTNet [37]	961.48	5256.57	38.33	47.78	38.59	48.39	38.20	47.05
SAFNet [22]	1120.85	3473.70	38.16	47.63	38.57	48.42	37.94	46.77
LightFuse [27]	**1.57**	7.13	37.88	47.20	38.46	48.12	37.49	46.19
Ours	2.47	**3.63**	**38.38**	**47.87**	**38.67**	**48.62**	38.27	**47.07**

alongside deep learning models with varied input configurations: HDRUNet [6] for single-image, LightFuse [27] for lightweight dual-exposure fusion, and AHDRNet [40], SCTNet [37], and SAFNet [22] for multi-exposure methods. For fair comparison, all methods were trained from scratch on our synthetic dataset under the same training budget, adhering to the hyper-parameters recommended in their original papers.

Quantitative Analysis. We evaluate all methods on the same test split using identical detection back ends and inference settings. In addition to standard image-quality metrics (PSNR and SSIM), we report downstream object detection accuracy to directly measure how HDR reconstruction affects perception under challenging illumination. As shown in Table 2, our method achieves state-of-the-art performance with a minimal computational footprint—utilizing only 2.47k parameters and 3.63 GMACs, a reduction of several orders of magnitude compared to prior models. Despite this efficiency, our method *matches or exceeds* prior work on the overall metrics (tied-best All-AP and the best All-AR) and achieves the best performance in day scenes while remaining highly competitive at night (within 0.01 AP of the best result).

Qualitative Analysis. Fig. 8 illustrates a representative night case with strong headlight glare. HDRUNet, constrained by its single-image input, struggles to reconstruct missing details in saturated regions due to the lack of physical references for fusion, resulting in false detections. LightFuse recovers details in saturated regions through the short-exposure frame, but the absence of a less-noisy mid-exposure frame makes it difficult to adequately denoise the short-exposure signal, causing noise artifacts to propagate to the final output. Multi-exposure methods such as Exposure Fusion and SAFNet provide robust results in saturated areas but fail to effectively suppress noise in unsaturated regions, leading to visible noise in the output. In contrast, our approach balances high-fidelity detail reconstruction in saturated regions with robust noise suppression in unsaturated areas. This advantage is reflected in the object detection performance, where our model achieves accurate localization and higher confidence scores across both

Fig. 8: Detection results comparison under headlight glare (night). Top: full view with ROI1 (red) and ROI2 (green). Bottom: per-method ROI crops with detector boxes and confidence scores. Our method improves highlight recovery and noise suppression, yielding higher-confidence detections.

ROIs. Additional qualitative results across diverse scenes are provided in the supplementary material.

Computational Efficiency. As summarized in Table 3, we evaluate the computational efficiency of the proposed method. Performance gain denotes the metric improvement achieved by the fused image relative to the mid-exposed LDR input. Computational efficiency is defined as the performance gain divided by the total GMACs, representing the benefit-to-cost ratio of each model. This analysis reveals that our approach achieves significantly higher efficiency than state-of-the-art algorithms while maintaining or surpassing their detection performance. While LightFuse, a lightweight dual-exposure fusion model, offers improved efficiency, our model further optimizes the performance-complexity trade-off by conditional execution. Specifically, it achieves a ΔPSNR-μ/GMACs of 4.008,

Table 3: Computational efficiency comparison with state-of-the-art models

Method	Performance(↑)		Perf. gain(↑)		Efficiency(↑)		Cost(↓)	
	PSNR-μ	AP	ΔPSNR-μ	ΔAP	$\frac{\Delta\text{PSNR-}\mu}{\text{GMACs}}$	$\frac{\Delta\text{AP}}{\text{GMACs}}$	Latency(ms)	Mem.(MB)
LDR (mid)	30.41	34.24	-	-	-	-	-	-
AHDRNet [40]	47.37	**38.38**	16.96	**4.14**	0.002	0.001	5570.4	4905.3
HDRUNet [6]	37.19	36.36	6.78	2.12	0.004	0.001	2226.5	1404.4
SCTNet [37]	**48.28**	38.33	**17.87**	4.09	0.003	0.001	27905.2	5162.76
SAFNet [22]	41.23	38.16	10.82	3.92	0.003	0.001	1800.4	2631.6
LightFuse [27]	40.67	37.88	10.26	3.65	1.439	0.511	15.5	1329.6
Ours	44.96	**38.38**	14.55	**4.14**	**4.008**	**1.140**	**15.0**	**1233.2**

approximately 2.8 times higher than LightFuse (1.439), and the detection efficiency (ΔAP/GMACs) reaches 1.140, significantly exceeding other models.

Furthermore, to validate the practical impact of region-wise conditional execution, we measure inference latency and peak memory consumption. To ensure a fair comparison, all models are evaluated under identical conditions and settings on an NVIDIA RTX 3090 GPU. The input frames are divided into 128×128 tiles with sufficient overlap to avoid tile-boundary artifacts. Due to the saturation-aware routing strategy, our method achieves both the shortest latency and the lowest peak memory usage among all methods. These results demonstrate that our method provides substantial performance gains with minimal computational overhead, making it highly suitable for resource-constrained environments.

4.3 Ablation Study

Component-wise Ablation. As shown in Table 4, we perform an ablation study to clarify the contribution of each component to the performance. Starting from a single two-exposure backbone as a baseline (C1), we first extend the input stack to three exposures (C2), then evaluate a naïve capacity-scaled variant that employs a dual-mode configuration and always executes the full model (C3). We then add masked supervision (C4) to specialize each mode for its target region. Finally, we test our saturation-aware conditional execution (C5).

Impact on Task Performance. Increasing model capacity with a dual configuration (C3) contributes most to HDR reconstruction quality (40.93 $\rightarrow$ 44.01 PSNR-μ) at the expense of additional computational cost (7.84 $\rightarrow$ 10.23 GMACs); however, this reconstruction gain does not directly translate into object detection (37.68 $\rightarrow$ 37.80 AP). Adding masked supervision to this dual configuration (C4) leads to substantial detection gains (37.80 $\rightarrow$ 38.38 AP) with a further modest reconstruction improvement (44.01 $\rightarrow$ 44.96 PSNR-μ) at the same compute. These performance gains arise because gradients from saturated pixels predominantly update the fusion configuration, while gradients from unsaturated pixels update the denoising configuration. This naturally encourages specialization for each region instead of forcing a single computation path to handle both regimes.

Table 4: Component-wise ablation study: configurations (top) and performance (bottom). The **blue-bold** denotes the design choice introduced at that step, and the underlining indicates the most improved metric from that addition.

Feature	C1	C2	C3	C4	**C5 (Ours)**
Input stack	2	**3**	3	3	3
# of modes	1	1	**2**	2	2
Masked supervision	–	–	×	✓	✓
Tile-wise cond. exec.	–	–	×	×	✓
Params (k)	1.57	1.80	2.47	2.47	2.47
GMACs	7.13	7.84	10.23	10.23	**3.63**
PSNR-μ	40.67	40.93	44.01	**44.96**	**44.96**
AP	37.88	37.68	37.80	**38.38**	**38.38**

Impact on Computational Efficiency. Starting from the baseline (C1), adding an input stack (C2) and a dual-mode configuration (C3) progressively increases both the parameter count (1.57k → 2.47k) and computational cost (7.13 → 10.23 GMACs). Our conditional execution reduces this overhead without compromising task performance. The expensive fusion configuration runs only on saturated regions, while a lightweight denoising configuration (37% of the full backbone channels) handles the remaining unsaturated regions. As unsaturated areas dominate most frames (83.94% on average in our dataset), this selective execution significantly reduces the total compute to 3.63 GMACs, a 54% reduction over C2 and 65% over C3 and C4. Remarkably, this is even 49% below the baseline (7.13 GMACs) while achieving higher PSNR-μ and AP, demonstrating that conditional execution decouples model capacity from compute.

Tile Size and Latency Analysis. To validate our choice of the tile size for conditional execution, we profile latencies and routing statistics across tile sizes in Table 5. Compute savings from conditional execution outweigh the scheduling overhead, and 128×128 provides the optimal latency–overhead trade-off.

Table 5: Tile size analysis and component latency breakdown.

Tile size	32	64	**128**	256	128
Conditional execution	✓	✓	✓	✓	✗
Total # of tiles	5220	1305	**345**	96	345
- Unsaturated only	3848	893	**214**	52	-
- Saturated only	27	5	**1**	0	-
- Contains both	1345	407	**130**	44	345
Total latency (ms)	21.24	16.34	**15.24**	17.10	17.70
- Mask generation & route	0.53	1.18	**1.46**	2.70	1.32
- Inference	20.71	15.16	**13.78**	14.40	16.38

The **blue-bold** denotes our design choice.

Saturation Threshold and Boundary Consistency. Table 6 presents a sensitivity analysis on the mask type and saturation threshold, showing that the 90% soft mask achieves the best performance. Fig. 9 shows smoother bound-

Table 6: Sensitivity analysis on saturation threshold.

Mask	Thres. (%)	PSNR-μ	SSIM-μ	AP	AR
Hard (Binary)	90	44.87	0.985	38.35	47.72
Soft	85	44.93	0.985	38.35	47.84
	90	**44.96**	**0.985**	**38.38**	**47.87**
	95	44.75	0.985	38.29	47.67
	99	42.40	0.981	37.75	47.12

The **blue-bold** denotes our design choice.

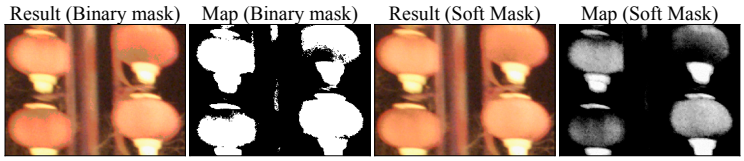


Fig. 9: Binary vs. soft saturation routing.

aries than hard routing, mitigating the spatial inconsistency risk from running different parameter sets across regions.

Robustness under Sensor Perturbations. To verify that SplitHDR remains robust against the diverse perturbations encountered in real sensor pipelines, we evaluate four perturbation types covering sensor noise, exposure-ratio drift, alternative ISP pipelines, sub-pixel optical offsets of advanced HDR sensors (e.g., photodiode-offset and microlens misalignment), and random frame misalignment. Across all variations, SplitHDR remains stable with a drop of at most 0.13 AP and 0.10 AR (approximately 0.3% relative) without retraining. Detailed experimental setup and results are provided in the supplementary material.

5 Conclusion

In this work, we present SplitHDR, a region-adaptive HDR framework tailored for real-time object detection in autonomous driving. Our approach addresses the computational redundancy in existing HDR pipelines by leveraging the observation that saturation is spatially sparse in typical driving scenarios. By introducing a luminance-aware dynamic routing mechanism, the model effectively decomposes the HDR task into lightweight denoising for the dominant unsaturated regions and targeted fusion for saturated highlights. Experimental results demonstrate that this divide-and-conquer strategy yields a highly efficient architecture, and the resulting model requires only 3.63 GMACs and 2.47k parameters, reducing the computational burden by orders of magnitude compared to state-of-the-art multi-exposure networks. Crucially, this extreme compression does not compromise downstream perception; the HDR images recovered by SplitHDR effectively maintain structural details, yielding competitive object detection performance. These findings validate that allocating computational resources based on pixel-wise information density is a viable path toward deploying robust, latency-critical perception systems on edge devices.

Acknowledgements. This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation under Grant RS-2025-02218733, Grant RS-2021-II211343, Grant IITP-2025-RS-2023-00256081, and Grant RS-2024-00347394, and in part by the National Research Foundation of Korea under Grant RS-2024-00408040 and Grant RS-2022-00144419.

References

1. Akahane, N., Sugawa, S., Adachi, S., Mori, K., Ishiuchi, T., Mizobuchi, K.: A sensitivity and linearity improvement of a 100-db dynamic range cmos image sensor using a lateral overflow integration capacitor. IEEE Journal of Solid-State Circuits **41**(4), 851–858 (2006)
2. Alam, M.Z., Kaleem, Z., Kelouwani, S.: How to deal with glare for improved perception of autonomous vehicles. arXiv preprint arXiv:2404.10992 (2024)
3. Asatsuma, T., Sakano, Y., Iida, S., Takami, M., Yoshiba, I., Ohba, N., Mizuno, H., Oka, T., Yamaguchi, K., Suzuki, A., et al.: Sub-pixel architecture of cmos image sensor achieving over 120 db dynamic range with less motion artifact characteristics. In: Proceedings of the International Image Sensor Workshop (IISW). vol. 1 (2019)
4. Bouderbane, M., Lapray, P.J., Dubois, J., Heyrman, B., Ginhac, D.: Real-time ghost free HDR video stream generation using weight adaptation based method. In: Proceedings of the International Conference on Distributed Smart Camera. pp. 116–120 (2016)
5. Brooks, T., Mildenhall, B., Xue, T., Chen, J., Sharlet, D., Barron, J.T.: Unprocessing images for learned raw denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
6. Chen, X., Liu, Y., Zhang, Z., Qiao, Y., Dong, C.: HDRUNet: Single image hdr reconstruction with denoising and dequantization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 354–363 (2021)
7. Chen, Z., Wang, Y., Cai, X., You, Z., Lu, Z., Zhang, F., Guo, S., Xue, T.: UltraFusion: Ultra high dynamic imaging using exposure fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16111–16121 (2025)
8. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 643–652. Association for Computing Machinery (2023)
9. Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R.K., Unger, J.: HDR image reconstruction from a single exposure using deep CNNs. ACM Transactions on Graphics (TOG) **36**(6), 1–15 (2017)
10. Elhence, A., Panda, A., Chamola, V., Sikdar, B.: FPGA-accelerated YOLOX with enhanced attention mechanisms for real-time wildfire detection on UAVs. IEEE Transactions on Instrumentation and Measurement (2025)
11. Feng, Y., Kagawa, K., Mars, K., Yasutomi, K., Kawahito, S.: Programmable dynamic range HDR imaging with LED-flicker and motion artifact mitigation using a 4-tap CMOS image sensor. IEEE Sensors Journal (2025)
12. Foi, A., Trimeche, M., Katkovnik, V., Egiazarian, K.: Practical poissonian–gaussian noise modeling and fitting for single-image raw-data. IEEE Transactions on Image Processing (TIP) (2008)
13. Fujihara, Y., Murata, M., Nakayama, S., Kuroda, R., Sugawa, S.: 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 (2020)
14. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: Exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430 (2021)

15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1026–1034 (2015)
16. Hirakawa, K., Wolfe, P.J.: Optimal exposure control for high dynamic range imaging. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). pp. 3137–3140. IEEE (2010)
17. Iida, S., Sakano, Y., Asatsuma, T., Takami, M., Yoshiba, I., Ohba, N., Mizuno, H., Oka, T., Yamaguchi, K., Suzuki, A., et al.: A 0.68 e-rms random-noise 121dB dynamic-range sub-pixel architecture CMOS image sensor with LED flicker mitigation. In: Proceedings of the IEEE International Electron Devices Meeting (IEDM). pp. 10–2. IEEE (2018)
18. Iida, S., Kawamata, D., Sakano, Y., Yamanaka, T., Nabeyoshi, S., Matsuura, T., Toshida, M., Baba, M., Fujimori, N., Basavalingappa, A., et al.: A 3.0 μm pixels and 1.5 μm pixels combined complementary metal-oxide semiconductor image sensor for high dynamic range vision beyond 106 dB. Sensors **23**(21), 8998 (2023)
19. Ilstrup, D., Manduchi, R.: One-shot optimal exposure control. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 200–213. Springer (2010)
20. Janai, J., Güney, F., Behl, A., Geiger, A.: Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends in Computer Graphics and Vision **12**(1-3), 1–308 (2020)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Kong, L., Li, B., Xiong, Y., Zhang, H., Gu, H., Chen, J.: SAFNet: Selective alignment fusion network for efficient HDR imaging. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 256–273. Springer (2024)
23. Lee, D.Y., Eom, T.H., Kim, H.J.: Adaptive column-wise multi-gain HDR CMOS image sensor for single-shot HDR imaging. IEEE Transactions on Circuits and Systems I: Regular Papers (2025)
24. Li, G., Qiu, L., Cao, P., Xie, F., Ji, X., Sun, Q.: Real-time scene-adaptive tone mapping for high-dynamic range object detection. In: Proceedings of Conference on Neural Information Processing Systems (NeurIPS) (2025)
25. Li, S., Zhu, Z., Sun, H., Ning, X., Dai, G., Hu, Y., Yang, H., Wang, Y.: Toward high-accuracy and real-time two-stage small object detection on FPGA. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) **34**(9), 8053–8066 (2024)
26. Liu, Y.L., Lai, W.S., Chen, Y.S., Kao, Y.L., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Single-image HDR reconstruction by learning to reverse the camera pipeline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1651–1660 (2020)
27. Liu, Z., Yang, J., Yanushkevich, S., Yadid-Pecht, O.: LightFuse: Lightweight CNN based dual-exposure fusion. arXiv preprint arXiv:2107.02299 (2021)
28. Martínez, M.A., Valero, E.M., Hernández-Andrés, J.: Adaptive exposure estimation for high dynamic range imaging applied to natural scenes and daylight skies. Applied optics **54**(4), B241–B250 (2015)
29. Mertens, T., Kautz, J., Van Reeth, F.: Exposure fusion. In: Proceedings of the Pacific Conference on Computer Graphics and Applications (PG). pp. 382–390. IEEE (2007)
30. Pourreza-Shahri, R., Kehtarnavaz, N.: Automatic exposure selection for high dynamic range photography. In: Proceedings of the IEEE International Conference on Consumer Electronics (ICCE). pp. 471–472. IEEE (2015)

31. RangiLyu: NanoDet-Plus: Super fast and high accuracy lightweight anchor-free object detection model. `https://github.com/RangiLyu/nanodet` (2021)
32. Sangiovanni, M.A., Spagnolo, F., Corsonello, P.: Hardware-oriented multi-exposure fusion approach for real-time video processing on FPGA. In: Proceedings of the Conference on Ph. D Research in Microelectronics and Electronics (PRIME). pp. 129–132. IEEE (2022)
33. Sim, S., Jun, J.: Advancements in active-pixel-type cmos image sensor design techniques and architectures for wide dynamic range. Sensors **26**(2), 489 (2026)
34. Solhusvik, J.: A 1280×960 3.75 μm pixel CMOS imager with triple exposure HDR. Proceedings of the International Image Sensor Workshop (IISW) pp. 344–347 (2009)
35. Solhusvik, J., Willassent, T., Mikkelsen, S., Wilhelmsen, M., Manabe, S., Mao, D., He, Z., Mabuchi, K., Hasegawa, T.: 1280×960 2.8 μm HDR CIS with DCG and split-pixel combined. In: Proceedings of the International Image Sensor Workshop (IISW). pp. 23–27 (2019)
36. Takayanagi, I., Kuroda, R.: HDR CMOS image sensors for automotive applications. IEEE Transactions on Electron Devices **69**(6), 2815–2823 (2022)
37. Tel, S., Wu, Z., Zhang, Y., Heyrman, B., Demonceaux, C., Timofte, R., Ginhac, D.: Alignment-free hdr deghosting with semantics consistent transformer. arXiv preprint arXiv:2305.18135 (2023)
38. Wu, B., Iandola, F., Jin, P.H., Keutzer, K.: SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 129–137 (2017)
39. Xu, R., Chen, C., Peng, J., Li, C., Huang, Y., Song, F., Yan, Y., Xiong, Z.: Toward RAW object detection: A new benchmark and a new model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13384–13393 (2023)
40. Yan, Q., Gong, D., Shi, Q., Hengel, A.v.d., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1751–1760 (2019)