HDRNET: HIGHLIGHT-CORRECTED DUAL-DOMAIN RATIO NETWORK FOR LOW-LIGHT SCENE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-light vision remains a fundamental challenge, as illumination degradation impairs downstream tasks. Recent methods build illumination-invariant color ratios in the log domain under a Lambertian model, and Phong-inspired extensions suppress highlight residuals after ratio formation. However, local highlights introduce channel-dependent biases into the log image before ratio construction. Such biases survive cross-channel subtraction and corrupt the ratio maps in the object regions. We propose HDRNet, a compact plug-and-play highlight-corrected dual-domain ratio network for low-light scene understanding. Local Highlight Attenuation module corrects for highlight-induced log-domain bias by predicting a bounded, channel-wise attenuation map from RGB cues and subtracting it before ratio construction, reducing contamination in the resulting ratio features. The corrected signal feeds a dual-branch module for spatial and spectral feature extraction, and a residual gated fusion integrates both branches with RGB appearance cues. Highlight Prior Regularization supervises the front-end during training. HDRNet delivers mAP gains of **2.2** points on ExDark and **2.6** points on DarkFace, and achieves **42.0% mask AP** on LIS and **62.1% mIoU** on ACDC-Night. Controlled ratio-recovery experiments support the correction mechanism, while highlight-stratified evaluation shows the largest detection gains under strong background highlights. Code will be released upon acceptance.

1 INTRODUCTION

Low-light scene understanding is a fundamental problem in computer vision and underpins nighttime autonomous driving and security surveillance (Zhang et al., 2016; Hong et al., 2022; Li et al., 2023b; Sakaridis et al., 2021). Modern detectors and segmenters perform well under normal illumination (Redmon & Farhadi, 2018; Kirillov et al., 2023), while their accuracy drops sharply in low light. Illumination degradation, underexposure, and color distortion corrupt the input and threaten the safety and reliability of these systems.

Existing solutions fall into four paradigms: (1) Enhancement preprocessing, (2) Synthetic data training, (3) Multi-task learning, and (4) Plug-and-play front-end, as illustrated in Figure 1(a). Enhancement preprocessing restores visibility before recognition, such as Retinex-based decomposition (Wei et al., 2018; Zhang et al., 2019; Brateanu et al., 2026) and curve estimation (Guo et al., 2020). These methods optimize visual quality rather than task accuracy and add a full enhancement network at inference. Synthetic data training, such as DAI-Net (Du et al., 2024) and WARLearn (Agarwal et al., 2025), avoids this overhead by exposing the recognition model directly to darkened images. However, synthetic degradation still differs from real nighttime imaging, so a domain gap remains. Multi-task learning (Cui et al., 2022; 2021) closes this gap by adapting to real low-light data jointly with recognition. The adaptation is tied to one architecture and must be retrained for every new task. Plug-and-play front-end removes this coupling. They insert a lightweight module before an unchanged downstream model and transfer across architectures and tasks.

Among plug-and-play front-ends, FeatEnHancer (Hashmi et al., 2023) learns task-oriented features, while YOLA (Hong et al., 2024) and FRBNet (Sun et al., 2025) use log-domain color ratios to cancel illumination shared across channels. However, local highlights introduce channel-dependent biases

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.

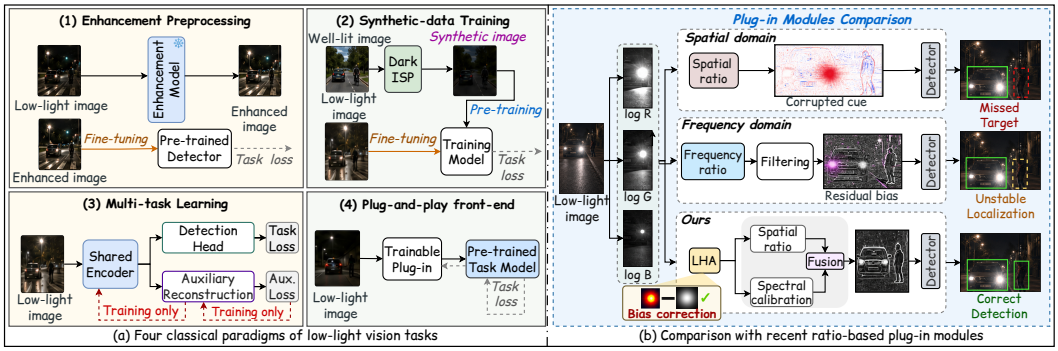


Figure 1: (a) Four classical paradigms of Low-light scene understanding, and (b) Comparison with recent ratio-based plug-and-play front-ends. Highlight-induced bias can cause missed targets and unstable localization. HDRNet attenuates this bias before ratio construction and fuses both spatial and spectral cues for reliable detection.

that survive ratio construction in both methods. YOLA uses finite-support spatial filters with limited capacity to capture long-range illumination variation, while FRBNet applies sign-unconstrained spectral modulation without guaranteeing phase preservation. As illustrated in Figure 1(b), to address these limitations, our method attenuates highlight-induced biases prior to ratio construction. Furthermore, rather than relying on a single ratio representation, we construct robust dual ratios across both spatial and spectral domains.

We propose HDRNet, a plug-and-play front-end centered on pre-ratio highlight correction. To remove the highlight bias, Local Highlight Attenuation (LHA) draws on dichromatic reflection (Shafer, 1985) and estimates a bounded, localized log-domain attenuation map. Subtracting this map from the log image reduces the highlight-induced bias passed to the ratio branches. To capture illumination variation beyond a finite support, Spatial Log-Ratio Extraction (SLRE) and Spectral Suppression and Recalibration (SSR) extract local invariant cues and non-local calibrations as parallel dual-domain branches. The SSR response is real and non-negative, so the ratio-spectrum phase is preserved by construction. To keep correction selective, Highlight Prior Regularization (HPR) supervises the front-end during training, and a residual gated fusion integrates both branches with RGB appearance cues. Extensive experiments on six low-light tasks, including object detection, face detection, semantic segmentation, instance segmentation, image classification, and video action recognition, show that HDRNet consistently outperforms strong baselines.

Our contributions are summarized as follows.

- We propose LHA to attenuate channel-dependent highlight bias that survives cross-channel subtraction. Guided by RGB highlight cues, it subtracts a bounded, spatially localized attenuation map from the log image before ratio construction, reducing highlight contamination in the resulting features.
- We design a dual-domain ratio extractor. The spatial branch SLRE captures fine-grained ratio structure, and the spectral branch SSR captures long-range illumination variation.
- Experiments on six low-light tasks show consistent gains over strong baselines, with the largest margins under strong background highlights.

2 Related Work

Low-light scene understanding. Low-light scene understanding remains a significant challenge for downstream recognition. One common line of research applies image enhancement (Zhang et al., 2019; Guo et al., 2020; Feijoo et al., 2025) to restore visibility before recognition, yet its objective differs from that of recognition. Synthetic data (Du et al., 2024; Agarwal et al., 2025) and multi-task learning (Cui et al., 2022; 2021) sidestep this mismatch, but the former hinges on simulation realism and the latter binds the solution to one architecture. Plug-and-play front-ends (Qin et al., 2022; Hashmi et al., 2023) avoid such intrusion by inserting a lightweight module before an unchanged

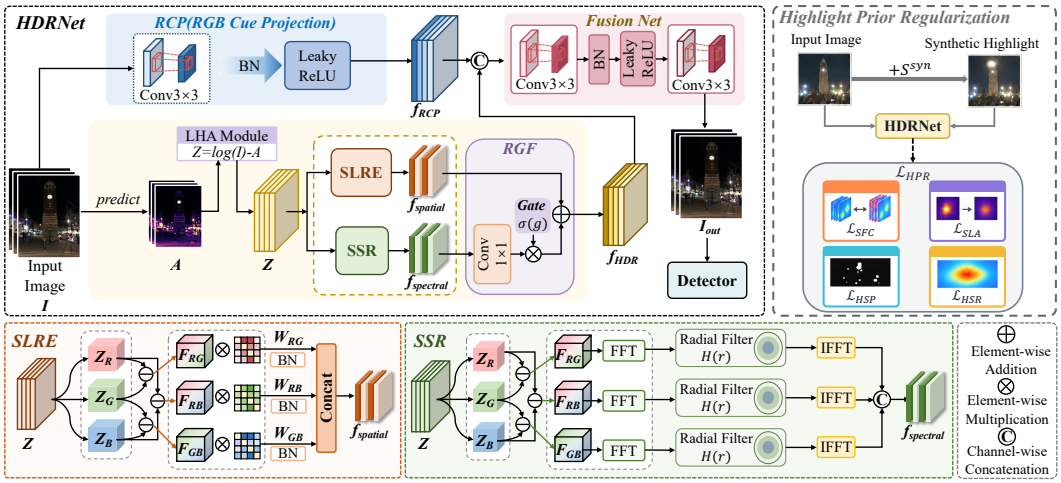


Figure 2: Overview of HDRNet. LHA attenuates highlight-induced log-domain bias before channel-ratio construction. The dual-domain design comprises SLRE and SSR, which extract complementary local spatial and phase-preserving non-local spectral ratio features, respectively. RGF combines these features through gated residual fusion. The fused features are concatenated with RGB cues from RCP and mapped to the downstream model input by the fusion network. During training, HPR supervises LHA with synthetic highlights to keep the attenuation selective and physically plausible.

network. Among them, YOLA (Hong et al., 2024) cancels illumination through cross-channel subtraction and zero-mean filtering under a Lambertian model, and FRBNet (Sun et al., 2025) adds a Phong-inspired specular term with frequency-domain ratio features. Despite their differences, both construct ratios before correcting local highlights, so channel-dependent highlight bias survives subtraction. In contrast, In contrast, our method attenuates this bias before ratio construction..

Frequency-domain modeling in low-light scene understanding. Frequency-domain methods separate illumination from structural information in the spectral domain and have been widely adopted for low-light scene understanding. FourLLIE (Wang et al., 2023) exploits Fourier frequency information for low-light image enhancement, while UHDFour (Li et al., 2023a) processes amplitude and phase separately to enhance illumination and suppress noise. These methods mainly optimize pixel reconstruction. FRBNet (Sun et al., 2025) instead applies frequency-domain modeling to ratio features and captures non-local illumination variation through learnable spectral modulation. However, sign-unconstrained spectral modulation can alter the phase of the ratio spectrum and affect spatial structure. Our spectral response is real and non-negative, avoiding phase changes introduced by the learned frequency-domain multiplier.

3 Method

3.1 Overview

HDRNet is a lightweight plug-and-play front-end for low-light scene understanding, as shown in Figure 2. Local Highlight Attenuation (LHA) predicts a bounded, localized attenuation map to correct highlight-induced log-domain bias before ratio construction. Spatial Log-Ratio Extraction (SLRE) and Spectral Suppression and Recalibration (SSR) then extract complementary local spatial and phase-preserving non-local spectral cues. Residual-Gated Fusion (RGF) combines both branches, while RGB Cue Projection (RCP) restores absolute appearance information lost during ratio construction and DC suppression. A fusion network produces a three-channel image for the unchanged downstream model. During training, Highlight Prior Regularization (HPR) uses synthetic highlights to keep the correction selective and physically plausible.

3.2 Theoretical Analysis

Lambertian ratio invariance. Color-ratio features have been shown effective for low-light scene understanding (Hong et al., 2024; Sun et al., 2025). Let $I_c(\mathbf{x})$ denote the observed intensity of color channel $c \in \{R, G, B\}$ at pixel $\mathbf{x}$. Under the Lambertian assumption (Shafer, 1985), the observation factorizes as

$$I_c(\mathbf{x}) = \ell(\mathbf{x})\, e_c(\mathbf{x})\, \rho_c(\mathbf{x}) + n_c(\mathbf{x}), \tag{1}$$

where ℓ is an achromatic factor absorbing geometric shading and illumination strength, e_c is the illuminant chromaticity, ρ_c is the channel-wise surface reflectance, and n_c is sensor noise. The detailed derivation is given in **Appendix A.1**. Following the local illumination model of YOLA (Hong et al., 2024), we assume that e_c is approximately constant within a small local window, i.e., $e_c(\mathbf{x}+\mathbf{u}) \approx e_c(\mathbf{x})$ for $\mathbf{u} \in \Omega$, where $\Omega = \{-q, \ldots, q\}^2$ collects the integer offsets of a $(2q{+}1) \times (2q{+}1)$ window centered at $\mathbf{x}$. The logarithm converts Eq. 1 into an additive form, and subtracting two channels removes the shared achromatic factor ℓ,

$$\begin{aligned} F_{ij}^{\text{ideal}}(\mathbf{x}) &= \log I_i(\mathbf{x}) - \log I_j(\mathbf{x}) \\ &\approx \underbrace{\log \rho_i(\mathbf{x}) - \log \rho_j(\mathbf{x})}_{\text{material}} + \underbrace{\log e_i(\mathbf{x}) - \log e_j(\mathbf{x})}_{\text{illuminant chromaticity}}, \quad i, j \in \{R, G, B\}. \end{aligned} \tag{2}$$

The material term preserves the reflectance-dependent color differences that are informative for recognition. The chromaticity term is locally constant under the above assumption and is therefore suppressed by the zero-mean spatial filters of Eq. 11 and the zero-DC spectral profile of Eq. 16.

Highlight-induced ratio residual. The smooth illumination model becomes insufficient when local highlights are present. We adopt a dichromatic-inspired additive decomposition (Shafer, 1985)

$$I_c(\mathbf{x}) = \ell(\mathbf{x})\, e_c(\mathbf{x})\, \rho_c(\mathbf{x}) + h_c(\mathbf{x}) + n_c(\mathbf{x}), \tag{3}$$

where $h_c \geq 0$ is a spatially localized and potentially channel-dependent highlight term. Ignoring the small noise term, the log-difference ratio becomes

$$\Delta_c(\mathbf{x}) = \log\left(1 + \frac{h_c(\mathbf{x})}{\ell(\mathbf{x})\, e_c(\mathbf{x})\, \rho_c(\mathbf{x})}\right), \tag{4}$$

$$F_{ij}^{\text{obs}}(\mathbf{x}) \approx F_{ij}^{\text{ideal}}(\mathbf{x}) + \Delta_i(\mathbf{x}) - \Delta_j(\mathbf{x}). \tag{5}$$

Unlike the smooth chromaticity offset in Eq. 2, the residual $\Delta_i - \Delta_j$ is confined to the highlight support with sharp boundaries, so it is not generally cancelled by cross-channel subtraction or zero-mean filtering. The smooth illuminant color is thus handled by the filtering stages, whereas localized highlight residuals motivate the explicit, RGB-guided attenuation used in our design.

3.3 Local Highlight Attenuation

Highlight candidate prior. To handle the residual term Δ_c in Eq. 5, LHA predicts a highlight attenuation map $A \in [0, a_{\max}]^{3\times H\times W}$. The map approximates the channel-wise log-domain bias that needs to be removed before ratio construction.

Specular highlights are typically bright and weakly saturated. We therefore derive two RGB-based candidate cues. The first cue $v(\mathbf{x})$ is the HSV value of pixel $\mathbf{x}$. It is defined as the largest normalized RGB intensity and serves as a brightness proxy. The second cue $s(\mathbf{x})$ is the HSV saturation. It measures the relative spread among the three channel intensities:

$$v(\mathbf{x}) = \max_c I_c(\mathbf{x}), \qquad s(\mathbf{x}) = \frac{\max_c I_c(\mathbf{x}) - \min_c I_c(\mathbf{x})}{\max_c I_c(\mathbf{x}) + \epsilon_{\log}}, \tag{6}$$

where $\epsilon_{\log}$ is a small numerical stabilizer. Specular reflections typically exhibit elevated radiance with limited chromatic variation. Therefore, pixels with high v and low s are more likely to belong to highlight regions.

We combine the brightness and saturation cues into a soft highlight-confidence map:

$$P_{\text{hl}}(\mathbf{x}) = \sigma(\kappa(v(\mathbf{x}) - \tau_v))\, \sigma(\kappa(\tau_s - s(\mathbf{x}))), \tag{7}$$

where τ_v and τ_s control the brightness and saturation thresholds, and κ controls the transition sharpness. Next, we clip the input as $I_\epsilon = \mathrm{clip}(I, \epsilon_{\log}, 1)$ to stabilize the logarithm. The predictor f_θ contains four depthwise-separable convolutional blocks with 16 hidden channels and dilation factors $(1, 2, 4, 2)$, followed by a 1×1 three-channel prediction head. The attenuation map and corrected log image are

$$A(\mathbf{x}) = a_{\max}\, \sigma(f_\theta(I_\epsilon)(\mathbf{x})) \odot P_{\mathrm{hl}}(\mathbf{x}), \qquad Z_c(\mathbf{x}) = \log(I_{\epsilon,c}(\mathbf{x})) - A_c(\mathbf{x}), \tag{8}$$

where P_{hl} is broadcast across channels. The sigmoid and $a_{\max}$ match the non-negative, bounded role of Δ_c and limit over-correction. Simultaneously, multiplication by P_{hl} biases attenuation toward likely highlight regions and discourages spatially diffuse correction. We perform the subtraction after taking the logarithm because A represents a log-domain bias. In contrast, using $\log(I_\epsilon - A)$ would incorrectly treat A as a linear-domain quantity and would also be unstable near zero intensity. The corrected image Z then provides a cleaner input for subsequent ratio construction. The resulting corrected log-ratio can be written as

$$F_{ij}^{\mathrm{corr}} = Z_i - Z_j = F_{ij}^{\mathrm{obs}} - (A_i - A_j). \tag{9}$$

Thus, LHA can predict the non-zero channel difference $A_i - A_j$ required to approximate the residual $\Delta_i - \Delta_j$ without being restricted to an achromatic correction.

3.4 Spatial Log-Ratio Extraction

LHA reduces the highlight-induced bias before ratio construction. The corrected log image Z can therefore provide a more reliable basis for local invariant feature extraction. Following the log-difference ratio representation in Eq. 9, SLRE constructs three pairwise maps:

$$F(Z) = (F_{RG}, F_{RB}, F_{GB}) = (Z_R - Z_G,\ Z_R - Z_B,\ Z_G - Z_B). \tag{10}$$

Cross-channel subtraction removes the shared factor ℓ and preserves the channel-dependent color differences. The ratio maps still contain locally smooth offsets, including the illuminant chromaticity term of Eq. 2 and possible calibration bias. SLRE suppresses these low-frequency nuisances with the local ratio-filtering principle of YOLA (Hong et al., 2024). A bank of K learnable zero-mean filters $\{W_k\}_{k=1}^K$ is shared across the three channel pairs. Each filter is defined on the support Ω introduced in Section 3.2, with $|\Omega| = (2q+1)^2$ taps. Here we use $q = 2$ and $K = 8$. The zero-mean constraint is imposed by re-centering the raw weights in every forward pass,

$$\bar{W}_k(\mathbf{u}) = W_k(\mathbf{u}) - \frac{1}{|\Omega|} \sum_{\mathbf{v} \in \Omega} W_k(\mathbf{v}), \qquad \sum_{\mathbf{u} \in \Omega} \bar{W}_k(\mathbf{u}) = 0. \tag{11}$$

Here $\mathbf{u}, \mathbf{v} \in \Omega$ denote offsets within the window and $W_k(\mathbf{u})$ is the weight of the k-th filter at offset $\mathbf{u}$. A zero-mean filter has zero response to a locally constant offset. It therefore removes smooth ratio-domain bias and emphasizes local changes such as edges and textures. For each channel pair p and filter k, the response at pixel $\mathbf{x}$ is

$$O_{p,k}(\mathbf{x}) = \sum_{\mathbf{u} \in \Omega} \bar{W}_k(\mathbf{u})\, F_p(\mathbf{x} + \mathbf{u}), \qquad p \in \{RG, GB, RB\}. \tag{12}$$

The term $\mathbf{x} + \mathbf{u}$ enumerates the spatial neighbors of $\mathbf{x}$ within the local window. Out-of-image positions are zero-padded, and the resulting output $O_{p,k}$ retains the original input resolution. Clamped log values are excluded by a binary validity mask. Following YOLA (Hong et al., 2024), each channel mask identifies strictly positive sensor readings, $M_c(\mathbf{x}) = \mathbf{1}[I_c(\mathbf{x}) > 0]$ for $c \in \{R, G, B\}$. Each channel pair is anchored by a distinct channel,

$$M_{RG} = M_R, \qquad M_{GB} = M_G, \qquad M_{RB} = M_B. \tag{13}$$

The response bank of each pair is normalized, masked, and concatenated across the three pairs,

$$f_{\mathrm{spatial}} = \mathrm{Concat}_p \left[M_p \odot \mathrm{BN}_p\big(\{O_{p,k}\}_{k=1}^K\big) \right]. \tag{14}$$

Each mask is broadcast across the K responses of its pair. SLRE captures local invariant structure with finite-support filters and cannot model long-range illumination variation. This limitation motivates the spectral branch.

3.5 SPECTRAL SUPPRESSION AND RECALIBRATION

Phase-preserving frequency calibration. Specifically, SSR receives the LHA-corrected log image $Z = \log(I_\epsilon) - A$ and constructs the same three pairwise log-ratio maps as in Eq. 10. Each ratio map is then transformed into the frequency domain:

$$\widehat{F}_p(\mathbf{k}) = \mathcal{F}(F_p)(\mathbf{k}), \qquad p \in \{RG, RB, GB\}, \tag{15}$$

where $\mathbf{k}$ denotes a frequency coordinate and $r(\mathbf{k})$ is its normalized radial distance. Like in the spatial domain, the frequency response first uses a soft zero-DC band-pass profile:

$$H_0(r) = \left(1 - \exp\left[-(r/r_l)^2\right]\right) \exp\left[-(r/r_h)^2\right]. \tag{16}$$

The DC coefficient equals the spatial mean of a ratio map, so $H_0(0) = 0$ is the frequency-domain counterpart of the zero-mean filters in SLRE. The first factor removes the DC component and nearby low-frequency energy, suppressing global ratio offsets from illumination color, sensor bias, or mean scene chromaticity. The second factor suppresses high-frequency noise, so H_0 retains the band most informative for ratio-based recognition.

A fixed band-pass response cannot adapt to different degradation spectra. We therefore introduce a radial-basis residual:

$$\delta_\psi(r) = \sum_{i=1}^{K_f} \sigma(\beta_i) c_i \exp\left(-\frac{(r-\mu_i)^2}{2\tau_b^2}\right), \qquad H(r) = H_0(r)\left(1 + \eta \tanh(\delta_\psi(r))\right), \tag{17}$$

where μ_i is the fixed radial center of the i-th basis function, c_i is its learnable coefficient, and τ_b is a shared learnable bandwidth. The learnable gate $\sigma(\beta_i)$ controls the contribution of the i-th basis function. We set $0 < \eta < 1$. The adaptive factor then remains positive, and $H(r)$ remains non-negative with $H(0) = 0$.

Then we apply the learned spectral response to each ratio spectrum and transform the filtered result back to the spatial domain:

$$F_p^{\text{cal}} = \mathcal{F}^{-1}\Big(H(r(\mathbf{k}))\widehat{F}_p(\mathbf{k})\Big), \qquad f_{\text{spectral}} = [F_{RG}^{\text{cal}}, F_{RB}^{\text{cal}}, F_{GB}^{\text{cal}}]. \tag{18}$$

For every frequency coordinate with $H(r(\mathbf{k})) > 0$, multiplying by the real and non-negative response changes only the spectral magnitude:

$$\arg\Big(H(r(\mathbf{k}))\widehat{F}_p(\mathbf{k})\Big) = \arg\Big(\widehat{F}_p(\mathbf{k})\Big). \tag{19}$$

Unlike a complex or sign-unconstrained spectral response, H contains no learnable phase term and never flips the sign of a frequency component. The real, non-negative response therefore calibrates spectral magnitudes without introducing phase changes at retained nonzero frequency components on the working-resolution grid.

Residual-gated fusion. The calibrated spectral feature complements the local structure extracted by SLRE. We first project f_{spectral} and add it to f_{spatial} through a residual gate:

$$\tilde{f}_{\text{spectral}} = \phi_{\text{sp}}(f_{\text{spectral}}), \qquad f_{\text{HDR}} = f_{\text{spatial}} + \sigma(g)\tilde{f}_{\text{spectral}}. \tag{20}$$

We initialize the gate to $g_0 = -4.0$ (i.e., $g \leftarrow g_0$) so that the spectral branch initially makes a small contribution, which is subsequently learned during training. RCP separately projects the original RGB image to retain absolute appearance cues removed by ratio construction and DC suppression. Through the fusion net in Figure 2, the final output is

$$I_{\text{out}} = \phi_{\text{fuse}}([\phi_{\text{rgb}}(I), f_{\text{HDR}}]). \tag{21}$$

3.6 HIGHLIGHT PRIOR REGULARIZATION

HDRNet and the downstream model are optimized jointly. The task loss $\mathcal{L}_{\text{task}}$ is the native objective of the downstream model evaluated on I_{out}. Following YOLA (Hong et al., 2024), we use a tone consistency loss $\mathcal{L}_{\text{cons}}$ to align the RGF outputs of Eq. 20 between an image and its gamma-jittered view, encouraging invariance to camera tone. **Appendix A.2** details both terms and the HPR loss.

Table 1: Low-light object detection and face detection on ExDark (Loh & Chan, 2019) and DarkFace (Yang et al., 2020b). Best and second-best results are shown in **Bold** and underline. All reported values are in %.

Paradigm	Method	ExDark						DarkFace					
		YOLOv3		TOOD		VFNet		YOLOv3		TOOD		VFNet	
		Recall	mAP	Recall	mAP	Recall	mAP	Recall	mAP	Recall	mAP	Recall	mAP
	Baseline	84.3	70.9	91.5	72.6	79.7	68.3	74.2	55.0	81.3	57.1	76.8	68.1
Enhancement	SMG [CVPR 2023]	82.6	68.7	91.9	71.5	80.5	68.7	73.8	52.3	80.5	56.4	80.1	68.6
	NeRCo [ICCV 2023]	83.2	68.4	91.2	71.7	80.8	69.2	73.4	53.4	79.2	56.5	80.4	69.0
	LightenDiffusion [ECCV 2024]	84.0	71.4	92.5	73.2	80.2	68.6	75.3	56.7	81.5	59.0	81.3	70.4
	DarkIR [CVPR 2025]	82.6	68.2	91.3	72.1	81.1	69.9	74.2	55.6	81.6	59.1	80.9	70.2
	Multinex [CVPR 2026]	83.5	70.1	92.4	72.0	81.5	70.1	72.8	55.5	81.1	58.6	81.1	70.2
Synthetic Data	DAINet [CVPR 2024]	86.4	73.5	-	-	-	-	74.9	57.1	-	-	-	-
	WARLearn [WACV 2025]	85.1	72.3	92.2	73.6	81.2	69.7	74.7	56.6	80.7	58.5	80.6	70.4
Multi-task	MAET [ICCV 2021]	86.0	72.6	92.7	74.3	80.9	68.8	74.3	55.2	80.1	57.7	79.9	70.1
	IAT [BMVC 2022]	84.4	72.5	92.5	73.2	81.3	69.8	73.8	55.6	79.2	58.0	80.8	70.9
Plug-and-play	DENet [ACCV 2022]	84.7	71.2	92.8	73.5	80.8	68.6	72.0	52.4	74.2	49.8	80.4	68.5
	FeatEnHancer [ICCV 2023]	87.1	71.2	**94.5**	74.1	81.7	70.1	74.4	55.5	80.9	57.7	78.6	69.7
	YOLA [NeurIPS 2024]	86.7	72.9	92.7	74.3	82.0	70.2	75.3	58.6	81.2	58.8	80.4	70.8
	FRBNet [NeurIPS 2025]	87.5	73.0	93.1	74.6	82.1	70.7	75.6	59.1	**81.8**	59.3	82.7	72.1
	HDRNet (Ours)	**88.0**	**74.6**	93.3	**75.4**	**84.7**	**72.9**	**78.7**	**61.2**	81.7	**61.9**	**85.1**	**73.4**

Task supervision alone does not explicitly enforce spatially selective attenuation. Optimizing only the task loss may therefore permit spatially diffuse attenuation. Highlight Prior Regularization supplies the missing constraint. We inject an analytic chromatic Gaussian highlight into each training image, which yields a synthetic view I^{syn}, a known log domain perturbation Δ^{syn} and a validity mask V^{syn} that excludes saturated pixels. Let A and A^{syn} denote the LHA attenuation maps of the two views and let sg denote the stop gradient operator. Four terms then constrain the correction,

$$\begin{aligned}
\mathcal{L}_{\mathrm{SFC}} &= \mathrm{SmoothL1}(f_{\mathrm{HDR}}(I^{\mathrm{syn}}),\, \mathrm{sg}[f_{\mathrm{HDR}}(I)])\,, \\
\mathcal{L}_{\mathrm{SLA}} &= \frac{\|V^{\mathrm{syn}} \odot [A^{\mathrm{syn}} - \mathrm{sg}[A] - \Delta^{\mathrm{syn}}]\|_1}{\max(\|V^{\mathrm{syn}}\|_1,\, 1)}, \\
\mathcal{L}_{\mathrm{HSP}} &= \frac{1}{3HW} \sum_{\mathbf{x}} \sum_{c\in\{R,G,B\}} A_c(\mathbf{x}), \\
\mathcal{L}_{\mathrm{HSR}} &= \mathrm{TV}(A) + \tfrac{1}{2}\,\mathrm{TV}(A^{\mathrm{syn}}).
\end{aligned} \tag{22}$$

Each term encodes a physical property of highlight correction rather than a direct accuracy target. Feature consistency keeps the corrected ratio representation stable under injection. Log alignment ties the attenuation response to a known perturbation and therefore constrains its magnitude. The sparsity term discourages unnecessary attenuation. Smoothness suppresses isolated responses. Writing $\mathcal{K} = \{\mathrm{SFC}, \mathrm{SLA}, \mathrm{HSP}, \mathrm{HSR}\}$, the complete training objective is

$$\mathcal{L}_{\mathrm{HPR}} = \sum_{k\in\mathcal{K}} \lambda_k \mathcal{L}_k, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{cons}}\mathcal{L}_{\mathrm{cons}} + \mathcal{L}_{\mathrm{HPR}}. \tag{23}$$

All synthetic views and regularization terms are removed during inference.

4 Experiments

We evaluate HDRNet on four primary low-light benchmarks. ExDark (Loh & Chan, 2019) covers generic object detection, and DarkFace (Yang et al., 2020b) focuses on nighttime faces. ACDC-Night (Sakaridis et al., 2021) and LIS (Chen et al., 2023) test semantic and instance segmentation. **Appendix B.5** extends the evaluation to image classification and video recognition. Within each benchmark and downstream model, every method shares the same split, input configuration, backbone, and optimization schedule. Detailed training protocols are provided in **Appendices A.5–A.7**.

4.1 Experimental Results

Low-light Detection. We prepend HDRNet to YOLOv3 (Redmon & Farhadi, 2018), TOOD (Feng et al., 2021), and VFNet (Zhang et al., 2021) and evaluate them on ExDark (Loh & Chan, 2019) and DarkFace (Yang et al., 2020b). HDRNet leaves the detection heads and task losses unchanged. Each

Table 2: Low-light semantic segmentation on ACDC-Night (Sakaridis et al., 2021). The symbol set {RO, SI, BU, WA, FE, PO, TL, TS, VE, TE, SK, PE, CA, TR, BI} represents *road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, car, train, bicycle*.

Method	RO	SI	BU	WA	FE	PO	TL	TS	VE	TE	SK	PE	CA	TR	BI	mIoU
Baseline [ECCV 2018]	90.3	61.1	72.4	33.8	34.3	46.3	50.6	32.4	69.4	14.7	80.9	27.5	62.1	76.8	12.9	51.0
RetinexNet [BMVC 2018]	89.3	60.1	70.7	29.9	28.9	42.2	47.1	25.6	65.1	8.7	75.8	21.8	53.9	67.1	8.3	46.3
DRBN [CVPR 2020]	90.9	61.2	73.5	32.7	32.2	43.9	49.1	26.8	65.5	10.7	76.3	24.1	56.6	71.1	12.1	48.4
FIDE [CVPR 2020]	89.8	60.2	73.6	32.1	31.5	43.8	44.7	27.9	66.8	12.9	77.6	26.7	56.5	70.8	12.1	48.5
KinD [ACM MM 2019]	90.5	61.7	72.4	31.7	33.4	43.6	42.1	27.7	65.3	13.8	76.8	23.5	55.2	74.9	13.2	48.4
EnGAN [TIP 2021]	89.8	59.3	73.4	32.9	31.2	44.7	49.2	26.1	67.1	14.6	77.5	24.4	58.2	71.1	8.2	48.5
Zero-DCE [CVPR 2020]	90.9	60.8	72.4	32.1	31.4	44.0	46.2	26.2	67.1	13.8	79.6	23.7	59.9	67.2	14.6	48.7
SSIENet [arXiv 2020]	89.4	59.5	71.8	29.9	31.8	45.5	43.8	24.2	66.9	10.7	78.3	22.5	51.7	71.4	5.8	46.9
Xue et al. [ACM MM 2022]	93.1	73.6	79.1	44.5	46.7	48.0	51.2	37.6	68.2	15.1	79.6	22.0	61.4	85.9	36.1	56.1
FeatEnHancer [ICCV 2023]	93.4	70.4	75.7	41.2	33.6	50.7	55.2	35.6	68.7	14.3	81.2	27.5	61.2	79.9	51.0	56.0
YOLA [NeurIPS 2024]	93.3	74.1	79.2	45.3	44.5	52.7	63.8	49.0	71.3	9.1	83.5	39.5	68.4	87.5	50.8	60.8
FRBNet [NeurIPS 2025]	**94.6**	74.8	79.1	45.5	45.1	51.9	64.6	49.6	**72.5**	9.6	**84.5**	40.3	69.8	88.2	49.4	61.3
HDRNet (Ours)	93.5	**75.8**	**79.9**	**46.5**	**46.9**	**53.2**	**65.0**	**51.4**	70.2	**15.3**	82.0	**41.4**	**70.6**	**88.7**	**51.2**	**62.1**

Table 3: Instance segmentation results on LIS (Chen et al., 2023). Values are AP (%).

Method	AP_{seg}	AP^{50}_{seg}	AP^{75}_{seg}	AP_{box}	AP^{50}_{box}	AP^{75}_{box}
Mask R-CNN [ICCV 2017]	34.8	52.1	34.2	35.7	57.3	42.4
MBLLEN [BMVC 2018]	33.6	51.8	33.3	36.9	58.6	44.1
DarkIR [CVPR 2025]	37.6	56.2	37.1	43.2	62.3	51.6
FeatEnHancer [ICCV 2023]	32.1	51.4	31.2	35.9	58.7	42.5
YOLA [NeurIPS 2024]	39.9	62.5	41.5	47.8	68.6	52.8
FRBNet [NeurIPS 2025]	40.5	63.4	41.7	48.4	69.3	54.2
HDRNet (Ours)	**42.0**	**66.0**	**42.7**	**49.9**	**72.7**	**54.5**

Table 4: Main component ablation with YOLOv3. Values are mAP (%).

SLRE	LHA	SSR	HPR†	ExDark	DarkFace
✓	×	×	×	72.7	58.6
✓	✓	×	×	73.4	59.5
✓	✓	✓	×	74.0	60.9
✓	✓	×	✓	73.8	60.1
✓	✓	✓	✓	**74.6**	**61.2**

front-end is evaluated with matched data splits, backbones, input configurations, and optimization schedules. YOLA (Hong et al., 2024) and FRBNet (Sun et al., 2025) are the most relevant competitors. Table 1 reports the detection results. Specifically, on ExDark, HDRNet reaches 74.6%, 75.4%, and 72.9% mAP with YOLOv3, TOOD, and VFNet. It achieves the highest mAP across all three detectors. Notably, on VFNet, our approach improves recall by **2.6%** and mAP by **2.2%** compared to FRBNet. On DarkFace, it obtains 61.2%, 61.9%, and 73.4% mAP, and achieves optimal performance, most notably yielding a **3.1%** increase in recall and a **2.1%** improvement in mAP on YOLOv3. These results demonstrate the generality and robustness of HDRNet across different detectors. **Appendix B.1** reports class-wise ExDark results.

Low-light Segmentation. We evaluate HDRNet on nighttime semantic segmentation and low-light instance segmentation. For ACDC-Night (Sakaridis et al., 2021), we prepend HDRNet to DeepLabV3+ (Chen et al., 2018) with an ImageNet-pretrained ResNet-50 backbone. Following FRBNet (Sun et al., 2025), we evaluate 15 ACDC-Night classes, excluding truck, bus, rider, and motorcycle. For LIS (Chen et al., 2023), we prepend HDRNet to Mask R-CNN (He et al., 2017) with a ResNet-50 backbone. All methods use matched task configurations. Detailed segmentation settings are provided in **Appendix A.6**. On ACDC-Night, HDRNet reaches 62.1% mIoU in Table 2. It exceeds FRBNet by 0.8 points and achieves the best performance in **12** of the 15 classes, including *traffic sign* (51.4%) and *sidewalk* (75.8%). Table 3 shows the same trend on LIS. HDRNet ranks first in all six metrics, with 42.0% mask AP, 66.0% mask AP_{50}, 49.9% box AP, and 72.7% box AP_{50}. **Appendix B.5** extends the evaluation to image classification and video recognition.

4.2 Ablation Studies and Analysis

Effectiveness of main components. Table 4 evaluates the main components with YOLOv3. Adding LHA to the SLRE baseline improves mAP from 72.7% to 73.4% on ExDark and from 58.6% to 59.5% on DarkFace, supporting the benefit of highlight correction before ratio construction. SSR raises the scores to 74.0% and 60.9%, HPR in its place reaches 73.8% and 60.1%, and the full model attains 74.6% and 61.2%. Adding HPR to the model with SSR yields further gains on both datasets, suggesting that its training constraints complement spectral calibration. HPR acts only during training and adds no inference cost. **Appendix B.2** analyzes its individual constraints.

Efficiency and performance tradeoff. Figure 3 plots detection accuracy against inference latency. HDRNet attains 61.9% mAP with 10.5K parameters at 12.4 ms. It surpasses the second-best FRBNet by 2.6 points for 1.1K additional parameters and 0.9 ms of latency. The margin widens to 3.1 points over YOLA and 4.2 points over FeatEnHancer. FeatEnHancer also carries 13× more parameters and runs 3.3 ms slower. Non-architectural methods share a fixed 7.7 ms budget and remain far less accurate. HDRNet therefore occupies the most favorable region of the accuracy and cost plane. **Appendix B.2** reports the remaining measurements.

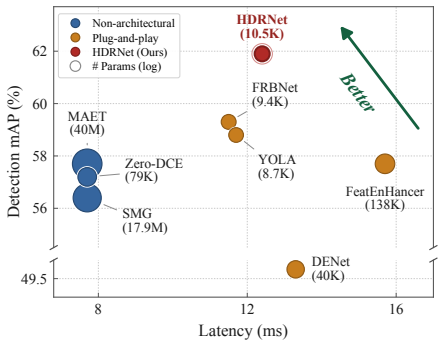


Figure 3: Efficiency-performance tradeoff on low-light detection.

4.3 Qualitative Comparison

Figure 4 connects the quantitative gains to visible recognition behavior. HDRNet recovers dim cars and pedestrians, keeps road boundaries cleaner, and reduces unstable responses near lamps and reflections. YOLA (Hong et al., 2024) and FRBNet (Sun et al., 2025) retain stronger nuisance responses, while HDRNet keeps target cues visible.

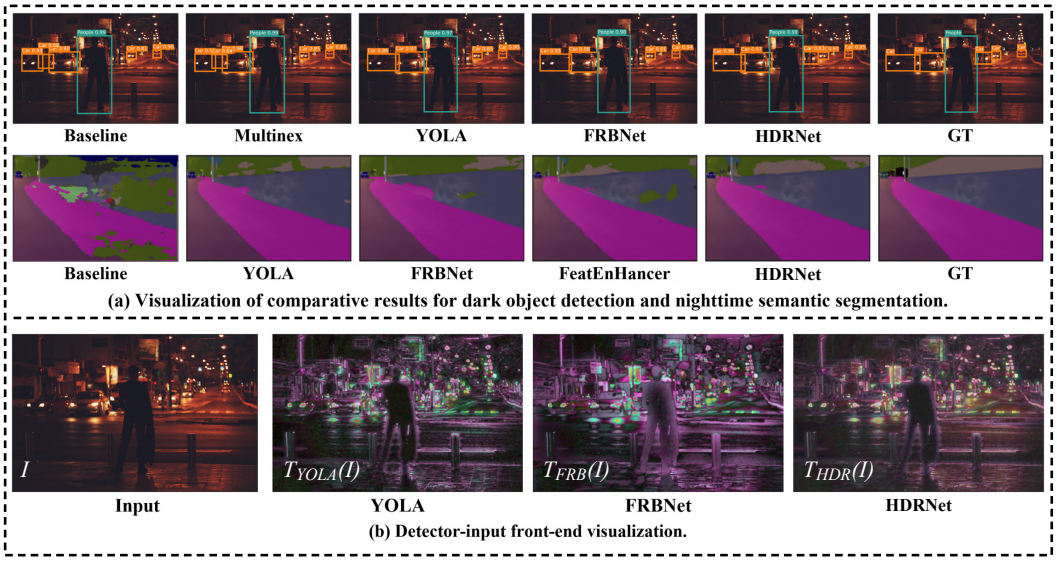


Figure 4: Qualitative comparison. HDRNet recovers targets under strong lights and keeps segmentation boundaries cleaner. Its front-end response is more selective than YOLA (Hong et al., 2024) and FRBNet (Sun et al., 2025).

5 Conclusion

This paper identifies an overlooked failure mode of channel-ratio features, where highlight-induced log-domain bias survives cross-channel subtraction and corrupts the ratio representation. We present HDRNet, a compact plug-and-play front-end that mitigates this failure mode. Its core module, LHA, predicts a bounded and spatially localized attenuation map and subtracts it from the log image before ratio construction, reducing highlight-induced contamination in the resulting features. On top of the corrected signal, a dual-domain design combines local invariant structure from SLRE with non-local spectral calibration from SSR, whose real, non-negative response preserves the ratio-spectrum phase by construction. Controlled recovery, highlight-stratified, and phase-intervention studies provide evidence for selective attenuation and improved structural alignment under the evaluated conditions. Extensive experiments on six low-light tasks show consistent gains over strong baselines without modifying the downstream model.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, grammatical correction, and improving the clarity and organization of the manuscript. These tools were not used to generate the core research ideas or methods, implement experimental code, execute experiments, or produce quantitative results. All AI-assisted revisions and suggestions were manually reviewed and verified against the authors' original analyses and experimental evidence to ensure that they did not alter the scientific claims. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All datasets used in this work are publicly available. Section 3 and **Appendices A.1 and A.2** provide the complete model formulation. Implementation details and dataset statistics are provided in **Appendices A.3 and A.4**, respectively, while task-specific training and evaluation settings are described in **Appendices A.5–A.7**. An anonymous implementation and the corresponding configurations will be available at `https://HDRnet-anon.github.io/HDRNet-Website/`.

REFERENCES

Shubham Agarwal, Raz Birman, and Ofer Hadar. Warlearn: Weather-adaptive representation learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4978–4987. IEEE Computer Society, 2025.

Alexandru Brateanu, Tingting Mu, Codruta O Ancuti, and Cosmin Ancuti. Multinex: Lightweight low-light image enhancement via multi-prior retinex. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29887–29896. Computer Vision Foundation, 2026.

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

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European conference on computer vision*, pp. 833–851. Springer, 2018.

Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023.

Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2533–2542. IEEE Computer Society, 2021.

Ziteng Cui, Kunchang Li, Lin Gu, Shenghan Su, Peng Gao, ZhengKai Jiang, Yu Qiao, and Tatsuya Harada. You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, pp. 238. BMVA Press, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009.

Zhipeng Du, Miaojing Shi, and Jiankang Deng. Boosting object detection with zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12666–12676. IEEE, 2024.

Daniel Feijoo, Juan C Benito, Alvaro Garcia, and Marcos V Conde. Darkir: Robust low-light image restoration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10879–10889. Computer Vision Foundation / IEEE, 2025.

Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3490–3499. Ieee, 2021.

Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1777–1786. Computer Vision Foundation / IEEE, 2020.

Khurram Azeem Hashmi, Goutham Kallempudi, Didier Stricker, and Muhammamd Zeshan Afzal. Featenhancer: Enhancing hierarchical features for object detection and beyond under low-light vision. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6702–6712. IEEE Computer Society, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 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.

Mingbo Hong, Shuiwang Li, Yuchao Yang, Feiyu Zhu, Qijun Zhao, and Li Lu. Sspnet: Scale selection pyramid network for tiny person detection from uav images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

Mingbo Hong, Shen Cheng, Haibin Huang, Haoqiang Fan, and Shuaicheng Liu. You only look around: Learning illumination invariant feature for low-light object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Hai Jiang, Ao Luo, Xiaohong Liu, Songchen Han, and Shuaicheng Liu. Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. In *European Conference on Computer Vision (ECCV)*, volume 15106, pp. 161–179. Springer, 2024.

Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Te-Te Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Attila Lengyel, Sourav Garg, Michael Milford, and Jan C. van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4379–4389. IEEE, 2021.

Chongyi Li, Chun-Le Guo, Man Zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2023a.

Shuiwang Li, Yangxiang Yang, Dan Zeng, and Xucheng Wang. Adaptive and background-aware vision transformer for real-time uav tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13943–13954, 2023b.

Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 10561–10570. Computer Vision Foundation / IEEE, 2021.

Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019.

Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, pp. 220. BMVA Press, 2018.

Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 5627–5636. IEEE, 2022.

MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. `https://github.com/open-mmlab/mmaction2`, 2020.

MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020.

Qingpao Qin, Kan Chang, Mengyuan Huang, and Guiqing Li. Denet: Detection-driven enhancement network for object detection under adverse weather conditions. In *Asian Conference on Computer Vision (ACCV)*, volume 13843, pp. 491–507. Springer, 2022.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10745–10755, 2021.

Steven A Shafer. Using color to separate reflection components. *Color Research & Application*, 10 (4):210–218, 1985.

Fangtong Sun, Congyu Li, Ke Yang, Yuchen Pan, Hanwen Yu, Xichuan Zhang, and Yiying Li. Frbnet: Revisiting low-light vision through frequency-domain radial basis network. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Chenxi Wang, Hongjun Wu, and Zhi Jin. Fourllie: Boosting low-light image enhancement by fourier frequency information. In *ACM International Conference on Multimedia (ACM MM)*, pp. 7459–7469. ACM, 2023.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pp. 20–36. Springer, 2016.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, pp. 155. BMVA Press, 2018.

Ke Xu, Xin Yang, Baocai Yin, and Rynson W. H. Lau. Learning to restore low-light images via decomposition-and-enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 2278–2287. Computer Vision Foundation / IEEE, 2020.

Xiaogang Xu, Ruixing Wang, and Jiangbo Lu. Low-light image enhancement via structure modeling and guidance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 9893–9903. IEEE, 2023.

Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. In *International workshop on deep learning for human activity recognition*, pp. 70–84. Springer, 2021.

Xinwei Xue, Jia He, Long Ma, Yi Wang, Xin Fan, and Risheng Liu. Best of both worlds: See and understand clearly in the dark. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2154–2162, 2022.

Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pp. 12872–12881. IEEE, 2023.

Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 3060–3069. Computer Vision Foundation / IEEE, 2020a.

Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J. Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, Yuqiang Zheng, Yanyun Qu, Yuhong Xie, Liang Chen, Zhonghao Li, Chen Hong, Hao Jiang, Siyuan Yang, Yan Liu, Xiaochao Qu, Pengfei Wan, Shuai Zheng, Minhui Zhong, Taiyi Su, Lingzhi He, Yandong Guo, Yao Zhao, Zhenfeng Zhu, Jinxiu Liang, Jingwen Wang, Tianyi Chen, Yuhui Quan, Yong Xu, Bo Liu, Xin Liu, Qi Sun, Tingyu Lin, Xiaochuan Li, Feng Lu, Lin Gu, Shengdi Zhou, Cong Cao, Shifeng Zhang, Cheng Chi, Chubing Zhuang, Zhen Lei, Stan Z. Li, Shizheng Wang, Ruizhe Liu, Dong Yi, Zheming Zuo, Jianning Chi, Huan Wang, Kai Wang, Yixiu Liu, Xingyu Gao, Zhenyu Chen, Chang Guo, Yongzhou Li, Huicai Zhong, Jing Huang, Heng Guo, Jianfei Yang, Wenjuan Liao, Jiangang Yang, Liguo Zhou, Mingyue Feng, and Likun Qin. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020b.

Xiangchen Yin, Zhenda Yu, Zetao Fei, Wenjun Lv, and Xin Gao. Pe-yolo: Pyramid enhancement network for dark object detection. In *International Conference on Artificial Neural Networks (ICANN)*, pp. 163–174, 2023.

Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8514–8523, 2021.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pp. 1632–1640. ACM, 2019.

Yu Zhang, XiaoGuang Di, Bin Zhang, and ChunHui Wang. Self-supervised image enhancement network: Training with low light images only. *ArXiv*, abs/2002.11300, 2020.

Appendix

CONTENTS

A Method Details **15**

A.1 Derivation and Discussion of the Ratio Model 15

A.2 Highlight Prior Regularization Details 15

A.3 Implementation Details 16

A.4 Dataset Statistics 16

A.5 Detection Protocols 17

A.6 Segmentation Protocols 17

A.7 Classification and Video Recognition Protocols 17

B Additional Experiments and Analyses **17**

B.1 Class-wise ExDark Detection Results 17

B.2 Additional Ablations 18

B.3 Mechanism and Phase-Preservation Analysis 20

B.4 Extended Highlight-Stratified Analysis 22

B.5 Additional Recognition Tasks 23

B.6 Task-Specific Qualitative Results 24

C Limitations and Discussion **25**

C.1 Limitations 25

C.2 Future Work 26

This appendix provides the derivation of the ratio model, the full formulation of Highlight Prior Regularization (HPR), implementation details, and supporting experiments. Within each task, all methods share the same downstream architecture, data split, input resolution, training schedule, optimizer, batch size, and backbone initialization, and only the front-end differs.

A Method Details

A.1 Derivation and Discussion of the Ratio Model

Image formation. Under the Lambertian assumption (Shafer, 1985), the surface exhibits only diffuse reflection and contains no specular component. Its observation can be written as

$$I_c(\mathbf{x}) = h(\mathbf{n}(\mathbf{x}), \boldsymbol{\omega}_l(\mathbf{x}))\; E_c(\mathbf{x})\, \rho_c(\mathbf{x}) + n_c(\mathbf{x}), \tag{24}$$

where $h(\mathbf{n}, \boldsymbol{\omega}_l)$ describes the geometric interaction between the surface normal $\mathbf{n}$ and the illumination direction $\boldsymbol{\omega}_l$, and E_c is the incident illumination in channel c. The illumination E_c is in general channel dependent, since common nighttime sources such as street lamps and vehicle headlights are colored. We therefore factor it into a channel-shared strength and a chromaticity term, $E_c(\mathbf{x}) = \bar{E}(\mathbf{x})\, e_c(\mathbf{x})$, and absorb the geometric shading into a single achromatic factor $\ell(\mathbf{x}) = h(\mathbf{n}, \boldsymbol{\omega}_l)\bar{E}(\mathbf{x})$, which yields Eq. 1.

Local illumination assumption. The condition $e_c(\mathbf{x} + \mathbf{u}) \approx e_c(\mathbf{x})$ for $\mathbf{u} \in \Omega$ states that all pixels inside the local window share approximately the same illuminant color as its center, so e_c is locally constant while it may still vary across the image. The assumption constrains only the spatial variation of the illuminant color rather than its spectrum, so colored light sources remain fully admissible. The factor ℓ varies with illumination strength and surface shading, but describes neither the intrinsic color of the observed material nor the color of the illuminant. Removing both nuisance terms therefore suppresses variation caused by local darkness and illuminant color while retaining crucial reflectance-dependent color cues.

Log-domain approximation. The approximation in Eq. 2 holds when the signal dominates noise, so that the logarithm converts the multiplicative relation into an additive one. The filtered log-difference ratio then preserves material-dependent color differences while removing both the spatially varying illumination strength and the smooth illuminant color.

A.2 Highlight Prior Regularization Details

This subsection defines the task and tone consistency losses and details the construction behind the Highlight Prior Regularization terms of Eq. 22.

Task and tone consistency losses. Let D_Φ denote the downstream model and let y denote the task annotation. The task loss is the native objective of the downstream model and remains unchanged,

$$\mathcal{L}_{\text{task}} = \ell_{\text{task}}(D_\Phi(I_{\text{out}}), y)\,. \tag{25}$$

For tone consistency we sample $\gamma \sim \mathcal{U}(0.6, 0.8)$ and form the tone jittered view I^γ. The loss is

$$\mathcal{L}_{\text{cons}} = \text{SmoothL1}(f_{\text{HDR}}(I), f_{\text{HDR}}(I^\gamma))\,. \tag{26}$$

We apply it to the RGF output before RGB cue fusion. The constraint therefore acts on the ratio representation itself rather than on the final three channel image, which leaves the absolute appearance branch free to respond to tone.

Synthetic highlight generation. We generate a chromatic Gaussian highlight for each training image. The synthetic highlight in channel c is

$$S_c^{\text{syn}}(\mathbf{x}) = \alpha t_c \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}\|_2^2}{2\sigma_h^2}\right). \tag{27}$$

The strength parameter follows $\alpha \sim \mathcal{U}(0.10, 0.28)$. The tint parameter follows $t_c \sim \mathcal{U}(0.85, 1.15)$. Each color channel receives an independent tint parameter. The center $\boldsymbol{\mu}$ is sampled within the relative spatial range $[0.15, 0.85]$. The spatial scale is $\sigma_h = 0.08 \min(H, W)$. The synthetic view is

$$I^{\text{syn}} = \text{clip}(I + S^{\text{syn}}, 0, 1)\,. \tag{28}$$

We define the numerically stable inputs as

$$I_\epsilon = \mathrm{clip}(I, \epsilon_{\log}, 1)\,, \qquad I_\epsilon^{\mathrm{syn}} = \mathrm{clip}(I^{\mathrm{syn}}, \epsilon_{\log}, 1)\,. \tag{29}$$

The injected perturbation has a known log domain target

$$\Delta^{\mathrm{syn}} = \log I_\epsilon^{\mathrm{syn}} - \log I_\epsilon. \tag{30}$$

Clipping removes information at saturated pixels. We exclude these pixels through the validity mask

$$V_c^{\mathrm{syn}}(\mathbf{x}) = \mathbf{1}[I_{\epsilon,c}(\mathbf{x}) + S_c^{\mathrm{syn}}(\mathbf{x}) < 1] \cdot \mathbf{1}[G(\mathbf{x}) > \tau_g]\,, \tag{31}$$

where the second factor with $\tau_g = 0.18$ restricts alignment to the highlight support, where the target perturbation is non-negligible.

Synthetic Specular Feature Consistency. This term supervises the complete corrected ratio representation rather than the attenuation map alone. It preserves the channel differences required for highlight compensation. A direct channel equality constraint on A would suppress these differences.

Synthetic Log Specular Alignment. The difference $A^{\mathrm{syn}} - \mathrm{sg}[A]$ isolates the incremental response to the synthetically injected highlight. The validity mask prevents unreliable targets in saturated regions from contributing to the alignment objective.

Highlight Sparsity Prior. Highlight attenuation should remain inactive outside necessary regions. Penalizing the average clean image attenuation discourages spatially diffuse correction and global intensity suppression.

Highlight Smoothness Regularization. Real highlight responses usually exhibit local spatial coherence. We regularize both attenuation maps and down weight the synthetic map by one half. The total variation operator is

$$\mathrm{TV}(X) = \mathrm{mean}(|\nabla_x X|) + \mathrm{mean}(|\nabla_y X|)\,. \tag{32}$$

This objective suppresses isolated responses and local prediction noise.

A.3 Implementation Details

Our implementation is based on PyTorch. We use MMDetection (Chen et al., 2019) for object detection, face detection, and instance segmentation, MMSegmentation (MMSegmentation Contributors, 2020) for semantic segmentation, and MMAction2 (MMAction2 Contributors, 2020) for video action recognition. Image classification uses a separate PyTorch implementation. Experiments are conducted on NVIDIA A40 GPUs.

A.4 Dataset Statistics

Table 5 summarizes the datasets and evaluation splits used in our experiments. Our evaluation spans six low-light visual recognition tasks. The four primary benchmarks cover object detection, face detection, semantic segmentation, and instance segmentation. CoDaN (Lengyel et al., 2021) extends the evaluation to image classification with separate daytime and nighttime test sets. For ARID (Xu et al., 2021), we use the first predefined training/test split (split 1) included in the dataset package, comprising 2,288 training clips and 823 test clips across 11 action classes.

Table 5: Overview of the evaluated benchmarks. Counts refer to images except for ARID, which uses video clips. Eval denotes the validation or test split used for reporting results. CoDaN evaluation counts are listed as day/night; an additional 500 daytime images are used for validation.

Dataset	Task	# Classes	# Train	# Eval
ExDark (Loh & Chan, 2019)	Dark object detection	12	3,000	1,800
DarkFace (Yang et al., 2020b)	Dark face detection	1	4,800	600
ACDC-Night (Sakaridis et al., 2021)	Nighttime semantic segmentation	15	400	106
LIS (Chen et al., 2023)	Low-light instance segmentation	8	1,561	669
CoDaN (Lengyel et al., 2021)	Image classification	10	10,000	2,500 / 2,322
ARID (Xu et al., 2021)	Video action recognition	11	2,288	823

A.5 Detection Protocols

Here we provide additional implementation details for the low-light detection experiments described in Section 4.1. Across both benchmarks, YOLOv3 uses a Darknet-53 backbone, while TOOD and VFNet use ResNet-50 (He et al., 2016) with FPN. On ExDark (Loh & Chan, 2019), YOLOv3 and TOOD are trained for 24 epochs with an initial learning rate of 1×10^{-3} and a weight decay of 5×10^{-4}. VFNet uses an initial learning rate of 1×10^{-2} and a weight decay of 1×10^{-4}. YOLOv3 and TOOD use a 608×608 evaluation scale, whereas VFNet uses 1333×800. YOLOv3 and TOOD follow the matched FRBNet and YOLA protocols, with a linear warm-up of 1000 iterations. They are trained on 2 GPUs with 8 images per GPU. VFNet uses a warm-up of 500 iterations and it is trained on 4 GPUs with 4 images per GPU. On DarkFace (Yang et al., 2020b), all detectors are configured for single-class face detection and use an identical data split within each front-end comparison. YOLOv3 is trained for 20 epochs with an initial learning rate of 1×10^{-3}, whereas TOOD and VFNet are trained for 12 epochs with initial learning rates of 1×10^{-3} and 1×10^{-2}, respectively. All three detectors use a weight decay of 1×10^{-4} and a linear warm-up over the first 500 iterations. YOLOv3 adopts multiscale training from 320×320 to 608×608 and is evaluated at 608×608. Both TOOD and VFNet randomly resize training images within the range from 750×500 to 1500×1000, while evaluation is performed at 1500×1000 with the aspect ratio preserved. All three detectors are trained on 4 GPUs with 4 images per GPU, resulting in a total batch size of 16.

A.6 Segmentation Protocols

For ACDC-Night (Sakaridis et al., 2021), we adopt DeepLabV3+ (Chen et al., 2018) with an ImageNet-pretrained (Deng et al., 2009) ResNet-50 backbone (He et al., 2016). During training, input images are randomly resized around the reference scale of 2048×1024 using a scale ratio sampled from $[0.5, 2.0]$, followed by random crops of size 1024×512. Evaluation is conducted at a reference resolution of 2048×1024 with the original aspect ratio preserved. All methods are trained for 20K iterations under the same MMSegmentation training protocol. For LIS (Chen et al., 2023), we use Mask R-CNN (He et al., 2017) with a ResNet-50 backbone in the MMDetection framework. During training, the input scale is randomly selected from $\{(1333, 640), (1333, 704), (1333, 768), (1333, 800)\}$ while preserving the aspect ratio. Evaluation is performed at a reference resolution of 1333×800. All methods are trained for 24 epochs using SGD under an identical training protocol.

A.7 Classification and Video Recognition Protocols

For CoDaN image classification (Lengyel et al., 2021), we prepend HDRNet to an ImageNet-pretrained ResNet-101 backbone. The front-end operates on RGB inputs in $[0, 1]$ before ImageNet normalization. Training uses random 224×224 crops, horizontal flips, color jitter, and random rotation up to 20 degrees. The model is trained for 175 epochs using SGD with an initial learning rate of 0.05, momentum 0.9, and weight decay 10^{-5}. We report day and night accuracy on the two test splits. For ARID video recognition (Xu et al., 2021), we use TSN (Wang et al., 2016) with a ResNet-50 backbone in an MMAction2-style pipeline. HDRNet is applied independently to each decoded frame. The pipeline inverts the ImageNet normalization, applies the front-end on $[0, 1]$ RGB frames, and then restores the normalization before the backbone. The temporal architecture and recognition losses remain unchanged. The metrics are top-1 and top-5 accuracy.

B Additional Experiments and Analyses

B.1 Class-wise ExDark Detection Results

To complement the aggregate ExDark results reported in Section 4.1, Tables 6 and 7 provide a detailed class-wise comparison of AP across all 12 object categories using YOLOv3 and TOOD under identical experimental protocols. HDRNet achieves the highest overall mAP with both detectors, reaching 74.6% with YOLOv3 and 75.4% with TOOD. With YOLOv3, it yields particularly notable improvements in the challenging *table* and *motorbike* categories, outperforming the best competing results by 2.8 and 2.7 AP points, respectively. These class-wise gains, together with the consistent

Table 6: Class-wise ExDark detection results with YOLOv3.

Method	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motorbike	People	Table	mAP
Baseline [arXiv 2018]	79.2	72.5	70.5	82.4	79.4	65.3	67.0	70.2	79.5	63.2	77.3	44.0	70.9
MBLLEN [BMVC 2018]	77.3	72.8	70.1	80.1	81.2	65.7	65.4	70.9	77.3	64.2	77.8	43.8	70.6
KinD [ACM MM 2019]	81.2	74.7	71.2	81.5	80.0	61.2	61.7	67.3	75.4	62.8	75.5	42.4	69.6
Zero-DCE [CVPR 2020]	81.5	75.6	72.2	81.8	80.4	62.0	65.9	72.6	79.5	62.4	77.6	43.0	71.2
EnGAN [TIP 2021]	81.3	75.2	71.8	82.9	79.0	62.6	65.9	68.7	75.8	63.2	77.7	42.1	70.5
RUAS [CVPR 2021]	76.3	68.3	63.5	78.1	73.9	59.5	64.8	65.2	75.8	54.9	72.1	39.6	66.0
SCI [CVPR 2022]	81.3	73.6	74.0	81.9	77.6	65.4	65.2	72.1	78.6	62.9	78.5	42.4	71.1
NeRCo [ICCV 2023]	81.0	73.5	65.9	81.3	75.8	62.3	62.5	67.3	75.9	61.5	75.5	38.8	68.4
SMG [CVPR 2023]	77.5	72.1	65.4	84.2	77.5	65.3	65.7	68.5	75.3	58.3	72.2	42.2	68.7
LightenDiffusion [ECCV 2024]	81.5	74.6	73.7	85.0	80.1	62.7	67.8	71.1	74.9	63.4	75.8	46.5	71.4
DarkIR [CVPR 2025]	78.5	73.0	65.8	85.3	76.6	59.8	62.5	65.1	74.4	61.6	73.9	41.7	68.2
Multinex [CVPR 2026]	81.5	74.1	70.5	82.3	78.8	63.3	65.9	69.0	75.6	62.0	76.3	41.5	70.1
DENet [ACCV 2022]	81.3	74.5	74.0	87.2	79.4	63.7	66.0	69.4	76.3	61.5	76.8	44.5	71.2
PE-YOLO [ICANN 2023]	76.3	72.3	66.8	84.5	78.1	60.4	64.8	66.2	74.1	63.2	74.8	44.2	68.8
WARLearn [WACV 2025]	81.2	73.8	73.1	88.3	80.8	68.1	66.5	71.8	79.4	60.1	77.8	47.0	72.3
MAET [ICCV 2021]	81.4	73.6	74.1	88.4	81.5	**68.9**	67.2	71.7	79.2	60.5	78.7	46.4	72.6
IAT [BMVC 2022]	81.2	73.4	74.0	88.0	80.5	68.8	67.7	71.7	79.8	60.1	78.4	46.3	72.5
FeatEnHancer [ICCV 2023]	79.4	**75.8**	73.6	87.4	81.3	62.2	65.1	68.3	75.2	64.0	76.4	45.4	71.2
YOLA [NeurIPS 2024]	82.5	75.5	71.6	86.4	80.7	68.2	66.8	72.4	**81.7**	65.4	78.6	45.4	72.9
FRBNet [NeurIPS 2025]	82.8	73.4	73.7	88.5	80.4	66.1	**69.6**	73.1	80.5	63.9	79.1	44.7	73.0
HDRNet (Ours)	**83.8**	75.6	**74.2**	**89.3**	**82.2**	68.6	68.2	**73.8**	81.0	**68.1**	**80.3**	**49.8**	**74.6**

Table 7: Class-wise ExDark detection results with TOOD.

Method	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motorbike	People	Table	mAP
Baseline [ICCV 2021]	80.8	75.5	71.4	88.5	76.4	70.3	66.9	69.7	85.7	61.2	76.5	48.3	72.6
MBLLEN [BMVC 2018]	81.5	78.3	74.1	88.2	79.6	71.8	67.5	70.2	85.5	62.3	77.1	47.6	73.6
KinD [ACM MM 2019]	82.1	77.0	70.8	88.7	77.4	70.2	67.4	66.7	84.7	61.4	76.7	47.5	72.6
Zero-DCE [CVPR 2020]	82.3	79.0	72.1	89.7	78.3	71.2	68.6	69.4	84.7	64.5	78.0	48.2	73.8
EnGAN [TIP 2021]	81.3	76.9	70.7	89.5	77.1	70.2	68.3	68.2	84.0	62.3	77.9	49.5	73.0
RUAS [CVPR 2021]	78.6	74.1	67.0	85.2	72.6	67.1	67.9	67.2	77.2	57.6	71.4	47.5	69.5
SCI [CVPR 2022]	81.2	78.5	71.3	89.8	76.1	73.6	67.0	70.3	85.7	63.2	77.5	49.1	73.6
NeRCo [ICCV 2023]	78.9	75.4	70.9	87.2	75.4	69.6	66.8	69.6	82.8	59.4	75.7	49.1	71.7
SMG [CVPR 2023]	78.5	76.2	70.3	87.8	73.6	72.6	65.9	67.2	84.6	60.7	73.2	47.3	71.5
LightenDiffusion [ECCV 2024]	81.2	76.4	73.2	88.2	78.4	71.0	66.5	70.8	85.9	61.9	76.0	48.3	73.2
DarkIR [CVPR 2025]	78.9	78.7	70.2	88.5	76.3	70.6	68.1	66.4	83.6	59.2	75.8	49.0	72.1
Multinex [CVPR 2026]	78.5	78.3	71.2	87.6	75.2	71.5	68.2	67.1	83.0	59.2	75.1	48.7	72.0
DENet [ACCV 2022]	81.1	78.5	71.2	88.0	77.1	71.2	67.5	70.3	**87.3**	62.0	77.2	**50.0**	73.5
PE-YOLO [ICANN 2023]	73.8	72.9	66.7	84.3	73.2	65.7	62.9	65.9	79.5	53.2	71.6	44.5	67.9
WARLearn [WACV 2025]	81.5	78.2	71.6	88.4	77.1	71.3	67.5	71.2	87.2	62.5	76.5	49.8	73.6
MAET [ICCV 2021]	80.2	77.7	73.8	**90.2**	78.6	73.6	69.7	70.2	86.4	64.9	77.8	48.9	74.3
IAT [BMVC 2022]	81.5	78.8	71.2	89.2	77.3	70.3	68.5	68.5	82.9	62.5	77.8	49.7	73.2
FeatEnHancer [ICCV 2023]	83.1	77.2	73.5	89.9	78.5	72.2	67.3	**72.1**	85.6	63.9	77.2	48.9	74.1
YOLA [NeurIPS 2024]	83.3	78.6	74.5	87.8	78.3	72.3	69.0	71.5	85.7	65.5	77.4	48.2	74.3
FRBNet [NeurIPS 2025]	82.8	78.8	**75.0**	90.1	**79.9**	73.0	68.4	71.7	84.6	64.1	77.9	49.1	74.6
HDRNet (Ours)	**83.4**	**79.1**	74.5	89.6	**79.9**	**74.3**	**70.2**	**72.1**	87.0	**66.3**	**78.6**	49.4	**75.4**

improvements across two detector architectures, demonstrate that the benefits of HDRNet are not confined to a specific object category or detector.

B.2 Additional Ablations

The main experiments establish the effectiveness of the proposed components. We further investigate whether HPR imposes the intended physical constraints and whether the performance of HDRNet depends on narrowly selected hyperparameters.

Effect of highlight prior regularization. HPR is designed to prevent LHA from learning arbitrary attenuation by restricting the correction to a selective and physically plausible space. To examine

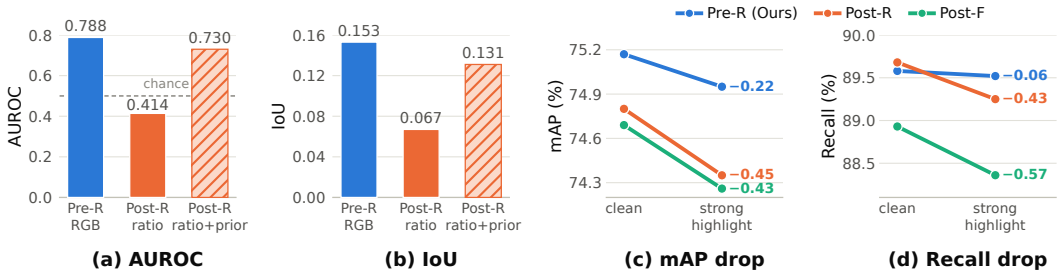


Figure 5: Correction-location analysis on ExDark. Panels (a) and (b) show highlight localization with equal-capacity correction heads. Panels (c) and (d) report detection accuracy on clean images and under strong synthetic highlights, with performance drops annotated at line endpoints.

Table 8: Leave-one-out ablation of HPR losses. Detection mAP (%).

Setting	ExDark	DarkFace
w/o $\mathcal{L}_{\mathrm{SFC}}$	74.4	61.1
w/o $\mathcal{L}_{\mathrm{SLA}}$	74.5	61.0
w/o $\mathcal{L}_{\mathrm{HSP}}$	73.9	60.7
w/o $\mathcal{L}_{\mathrm{HSR}}$	74.5	61.1
Full HPR	**74.6**	**61.2**

Table 9: Parameter sensitivity. Detection mAP (%) on ExDark and DarkFace.

Setting	Changed values	ExDark	DarkFace
LHA sensitive prior	$\tau_v = 0.40,\ \tau_s = 0.50$	74.0	60.9
LHA strict prior	$\tau_v = 0.50,\ \tau_s = 0.60$	74.2	60.6
LHA high attenuation	$a_{\max} = 0.75$	74.0	60.7
RGF weak residual	$g_0 = -4.5$	74.3	60.7
RGF strong residual	$g_0 = -3.0$	**74.6**	61.0
SSR narrow band	$r_l = 0.025$	74.4	60.4
SSR broad band	$r_l = 0.050$	74.0	**61.2**
Compact capacity	$C_{\mathrm{mid}} = 8,\ K = 4,\ K_f = 4$	74.5	60.8
Enlarged capacity	$C_{\mathrm{mid}} = 24,\ K = 12,\ K_f = 12$	74.3	60.8

the role of each constraint, Table 8 performs a leave-one-out ablation of the four HPR terms, with the tone-consistency loss kept unchanged. Removing the Highlight Sparsity Prior $\mathcal{L}_{\mathrm{HSP}}$ produces the largest degradation, reducing mAP from 74.6% to 73.9% on ExDark and from 61.2% to 60.7% on DarkFace. This result highlights the importance of keeping the attenuation spatially localized and sparse. Removing $\mathcal{L}_{\mathrm{SFC}}$, $\mathcal{L}_{\mathrm{SLA}}$, or $\mathcal{L}_{\mathrm{HSR}}$ causes smaller and less uniform changes. These terms respectively preserve the fused ratio representation under synthetic highlights, align the predicted attenuation change with the known log-domain perturbation, and promote spatially smooth attenuation. Their behavior supports the intended role of HPR as a set of complementary physical constraints rather than independent accuracy objectives. The full HPR objective therefore provides the most balanced correction across both datasets.

Parameter sensitivity. To determine whether the observed gains rely on a narrow configuration, Table 9 varies the highlight prior thresholds, maximum attenuation, fusion residual strength, SSR bandwidth, and model capacity while keeping the remaining settings fixed. C_{mid} denotes the intermediate channel width of the LHA predictor. Across all variants, mAP remains within 74.0% to 74.6% on ExDark and 60.4% to 61.2% on DarkFace, demonstrating that HDRNet is stable over a broad range of settings. The preferred configuration varies moderately across datasets. A stronger fusion residual performs best on ExDark, whereas a broader SSR band performs best on DarkFace. Increasing the model capacity provides no additional benefit on either benchmark. These results indicate that the improvements arise primarily from the constrained formulation rather than parameter scale or a narrowly tuned operating point.

Ablation on LHA Placement. A natural question is whether the correction must precede ratio construction or can be applied afterwards. Our analysis in Section 3.2 suggests that channel subtraction removes absolute intensity information useful for highlight localization, making localization more difficult for a corrector that observes only ratio maps. To test this, we compare three placements under the same ExDark protocol with YOLOv3. Pre-R is our LHA and observes the RGB image. Post-R observes the three log-ratio maps and predicts a signed residual per pair. Post-F suppresses highlights at the feature level after SLRE and SSR. We first isolate the information effect with equal-capacity correction heads on synthetic highlights in Figure 5 (a,b). The head trained on

Table 10: Efficiency and segmentation performance of different low-light scene understanding methods. Detection accuracy and inference latency are reported in Figure 3.

Category	Metric	Non-architectural Methods			End-to-End Trained Plug-and-Play Module				
		Zero-DCE	SMG	MAET	DENet	FeatEnHancer	YOLA	FRBNet	HDRNet
Efficiency	# Params ↓	79K	17.9M	40M	40K	138K	**8.7K**	9.4K	10.5K
	FLOPs (G) ↓	**50.6**			61.7	79.5	51.9	52.0	52.4
Performance	Seg (mIoU, %) ↑	48.7	49.7	-	52.2	56.0	60.8	61.3	**62.1**

ratio maps reaches an AUROC of 0.414, below chance, whereas the head trained on RGB reaches 0.788. Supplying the RGB prior to the ratio-domain head recovers most of the gap, which supports the importance of RGB-derived cues for highlight localization. We further evaluate whether these localization differences are accompanied by differences in detection robustness in Figure 5 (c,d). Pre-R achieves the highest clean mAP of 75.17% and the smallest performance drops under strong synthetic highlights. Its mAP decreases by only 0.22, compared with 0.45 for Post-R and 0.43 for Post-F, while its recall decreases by just 0.06, versus up to 0.57 for the alternatives (all drops in percentage points). These results support our RGB-conditioned LHA design, which uses absolute intensity cues to localize highlights and attenuate their influence before ratio construction.

Additional Efficiency-performance Trade-off Results. Figure 3 encodes detection accuracy, latency and parameter count. Segmentation accuracy and computational cost are reported here and in Table 10. HDRNet, FRBNet, and YOLA retain the same relative ranking across detection and segmentation. HDRNet reaches 62.1% mIoU and leads FRBNet (Sun et al., 2025) by 0.8 points and YOLA (Hong et al., 2024) by 1.3 points. FeatEnHancer (Hashmi et al., 2023) trails by 6.1 points and DENet (Qin et al., 2022) by 9.9 points. Zero-DCE (Guo et al., 2020) and SMG (Xu et al., 2023) fall to 48.7% and 49.7%. Computational cost supports the same conclusion. HDRNet raises the 50.6 GFLOPs of the detector to 52.4 GFLOPs for a relative overhead of 3.6%, and FRBNet and YOLA sit at 52.0 and 51.9 GFLOPs. DENet and FeatEnHancer instead consume 61.7 and 79.5 GFLOPs. FeatEnHancer spends 27.1 GFLOPs more than HDRNet while losing 4.2 points of detection accuracy and 6.1 points of segmentation accuracy. Across both tasks HDRNet therefore attains the strongest accuracy while adding only 1.8 GFLOPs and 10.5K parameters to the detector.

B.3 Mechanism and Phase-Preservation Analysis

To determine whether the observed gains arise from the intended mechanisms, we analyze the behavior of LHA and SSR. Specifically, we examine whether LHA selectively attenuates highlight-induced log-domain bias and whether SSR preserves spatial structure during spectral calibration.

Controlled recovery of ratio invariance. We evaluate the response of LHA under a controlled highlight perturbation. Each ExDark image receives the synthetic highlight defined in Eq. 27. The highlight is scaled by β and added to the input image. The spatial support covers 5%, 15%, or 30% of the image area. The clean ratio is denoted by Q_0. The corrupted ratio is denoted by Q_β. The LHA corrected ratio is denoted by $\widehat{Q}_\beta$. We construct a fixed validity mask from pixels unsaturated at $\beta = 1$. The same mask is used for every perturbation strength. This protocol prevents changes in the evaluation region from affecting the comparison. We report two regional errors. The highlight error E_h is the mean absolute deviation from Q_0 inside the injected support. The leakage error $E_{\bar{h}}$ is the corresponding deviation outside the injected support. We also measure the fraction of the corrupted ratio residual removed by LHA

$$R_{\mathrm{rec}}(\beta) = 1 - \frac{\left\|\mathbf{1}_{\Omega_h^\beta} \odot (\widehat{Q}_\beta - Q_0)\right\|_1}{\left\|\mathbf{1}_{\Omega_h^\beta} \odot (Q_\beta - Q_0)\right\|_1 + \epsilon}. \tag{33}$$

Figure 7 summarizes the results. The uncorrected curve provides a reference for the highlight-induced error in raw ratio maps before subsequent feature processing. At $\beta = 1.0$, LHA reduces the highlight region error E_h from 0.93 to 0.49. The non-highlight deviation $E_{\bar{h}}$ stays below 0.016. The small non-highlight deviation indicates limited leakage outside the injected support. The recovery

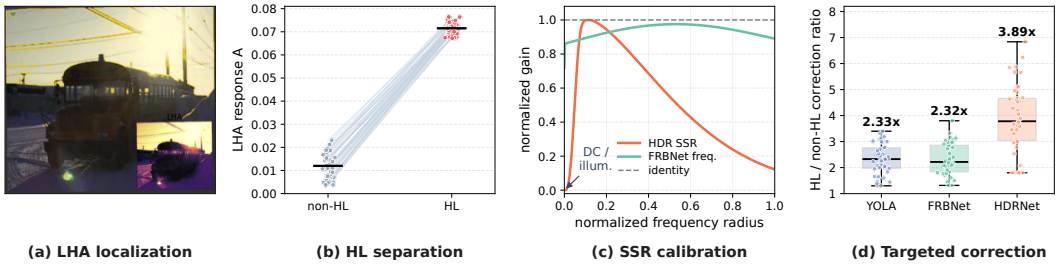


Figure 6: Internal mechanism diagnostics of HDRNet. The figure examines LHA localization, the SSR spectral response, and correction selectivity across different front-ends.

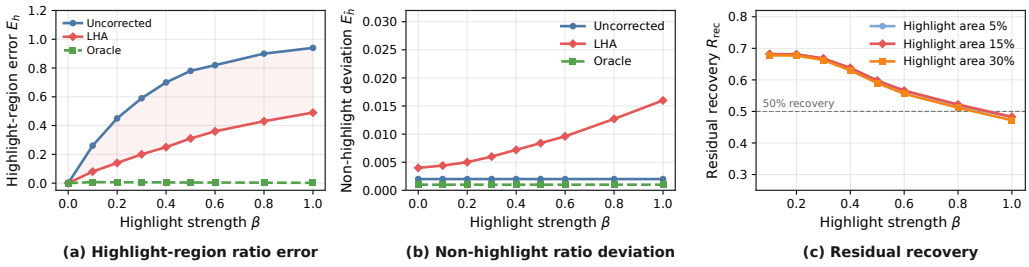


Figure 7: Controlled validation on ExDark. LHA reduces ratio error under known specular corruption, keeps leakage low in clean regions, and recovers a stable fraction of the corrupted residual.

rate R_{rec} exceeds 50% for $\beta \leq 0.8$ under all three coverages. LHA therefore attenuates the injected highlight bias selectively and preserves clean ratio responses outside the corrupted region.

Physical mechanism validation. To verify that HDRNet performs selective highlight correction rather than global intensity adjustment, Figure 6 visualizes the responses of LHA and SSR. Panels (a) and (b) show that LHA concentrates attenuation within highlight regions while producing consistently weaker responses elsewhere. Panel (c) further shows that SSR suppresses the DC component and low-frequency illumination energy while retaining informative mid-frequency structure. Panel (d) quantifies this selectivity. The ratio between correction strength inside and outside highlight regions reaches $3.89\times$ for HDRNet, compared with $2.33\times$ for YOLA and $2.32\times$ for FRBNet. These results confirm that HDRNet directs its correction toward plausible highlight interference instead of applying a global adjustment.

Ground truth evaluation of phase preserving spectral calibration. FRBNet (Sun et al., 2025) learns a complex spectral response. The response can modify spectral magnitude and phase. SSR instead applies a real nonnegative magnitude response. It preserves the input phase by construction. We evaluate structural alignment on 106 ACDC Night images with semantic boundary annotations. The strongest 12% of spectral gradient responses form the predicted edge set. We use a boundary tolerance of $\tau = 5$ pixels. The Boundary F score measures the harmonic mean of boundary precision and boundary recall. The distance metric D measures the mean distance from annotated boundaries to spectral ratio edges. Figure 8 reports paired differences for each image. The reported intervals are 95% paired bootstrap confidence intervals. Panels (a) and (b) examine the effect of spectral phase. The trained checkpoint and magnitude response remain fixed. Hermitian compatible phase shifts produce spatial displacements of 1, 2, and 4 pixels. The Boundary F score decreases by 0.03, 0.25, and 1.64 percentage points. The boundary distance increases by 0.50, 0.51, and 1.12 pixels. Larger phase displacements produce larger alignment errors. This controlled intervention confirms the geometric importance of phase fidelity. Panel (c) compares independently trained front-ends. HDRNet achieves a Boundary F score of 15.60%. FRBNet achieves 12.12%. The paired improvement is 3.49 percentage points. Its 95% confidence interval is $[3.06, 3.92]$. HDRNet performs better on 98.1% of the images. This comparison evaluates the complete front-end. Panel (d) isolates the spectral response within a fixed HDRNet pipeline. The native SSR response is replaced by an RMS gain matched FRBNet response. Native SSR achieves a Boundary F score of 15.60%. The transplanted

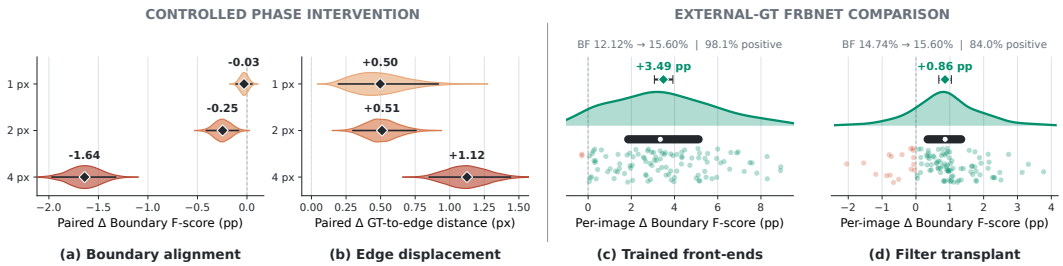


Figure 8: Semantic boundary evaluation of spectral calibration on ACDC-Night. Panels (a) and (b) show paired changes under controlled phase shifts. Panel (c) compares independently trained HDRNet and FRBNet front-ends. Panel (d) replaces SSR in the fixed HDRNet pipeline with the RMS gain-matched FRBNet response. Diamonds and intervals indicate paired means and 95% bootstrap confidence intervals. Positive values in (c) and (d) favor HDRNet and native SSR, respectively.

response achieves 14.74%. The paired improvement is 0.86 percentage points. Its 95% confidence interval is $[0.68, 1.05]$. Native SSR performs better on 84.0% of the images. The matched gain excludes global response scale as the sole explanation. The controlled intervention and transplant experiment jointly support phase preserving spectral calibration as a conservative design for semantic boundary alignment.

B.4 Extended Highlight-Stratified Analysis

Highlight-stratified robustness. Table 11 examines robustness across highlight severity and location. We stratify ExDark with the soft prior P_{hl} in Eq. 7 and construct two diagnostic partitions, each containing a light and a heavy subset of 300 images. The background-only partition accumulates highlight responses outside annotated objects, and the target-inclusive partition retains the full image. All methods use the same VFNet (Zhang et al., 2021) detector. HDRNet achieves the best mAP on all four subsets. All methods obtain higher absolute mAP on the heavy subsets. This trend reflects differences in subset composition and intrinsic detection difficulty. It does not indicate a beneficial effect of highlights. We therefore compare methods only within each fixed subset. In the background only partition, the HDRNet margin over the strongest baseline increases from 1.02 points on the light subset to 3.29 points on the heavy subset. In the target inclusive partition, HDRNet maintains a lead of more than 2 points on both subsets. The corresponding margins are 2.92 and 2.16 points. The HDRNet hit rate remains close to 96% in every subset. It exceeds both baselines by at least 3 points. These results demonstrate consistent robustness across all subsets. The largest relative advantage occurs under strong background nuisance highlights. This pattern agrees with the goal of correcting highlight induced bias before ratio construction.

Nuisance false positives. We verify that the higher hit rate reflects target recovery rather than additional detections on highlights. We split the heavy subset of the background-only partition in Table 11 into Medium-HL and High-HL groups and measure the normalized nuisance H-FP rate, defined as highlight-overlapping false positives per 100 retained detections. Figure 9(a) shows that HDRNet reduces this rate by 18% on Medium-HL and 31% on High-HL relative to the stronger of YOLA (Hong et al., 2024) and FRBNet (Sun et al., 2025) in each group. Figure 9(b) jointly evaluates near-highlight Recall@0.5 and the H-FP rate. On High-HL, HDRNet raises recall from approximately 94% to 98% while maintaining a lower H-FP rate than both baselines. These results support genuine target recovery rather than increased highlight-induced false positives.

Robustness to synthetic nuisance highlights. We evaluate robustness independently of the prior based stratification. We introduce four levels of synthetic background highlight interference on ExDark. Figure 10 shows that HDRNet consistently achieves the highest mAP across all perturbation levels. All methods exhibit modest positive mAP changes under stronger perturbations. The synthetic highlights are confined to background regions. They do not occlude annotated targets. The added illumination can improve local exposure and contextual contrast. We therefore interpret the clean to strong change as a stability measure. The positive values do not imply a general benefit of highlights for detection. From the clean input to the strongest perturbation, HDRNet changes

Table 11: Highlight-stratified robustness on ExDark using the same VFNet (Zhang et al., 2021) detector. Each split reports mAP, hit rate, and the heavy-subset gain.

Method	Background-only					Target-inclusive				
	mAP (%)			Hit rate (%)		mAP (%)			Hit rate (%)	
	Light	Heavy	Δ_H	Light	Heavy	Light	Heavy	Δ_H	Light	Heavy
YOLA [NeurIPS 2024]	67.06	76.43	–	90.99	92.61	61.21	73.39	–	91.17	92.68
FRBNet [NeurIPS 2025]	65.62	75.16	–	92.13	92.70	60.48	71.44	–	92.75	92.54
HDRNet (Ours)	**68.08**	**79.72**	**+3.29**	**95.80**	**96.29**	**64.13**	**75.55**	**+2.16**	**95.78**	**95.97**

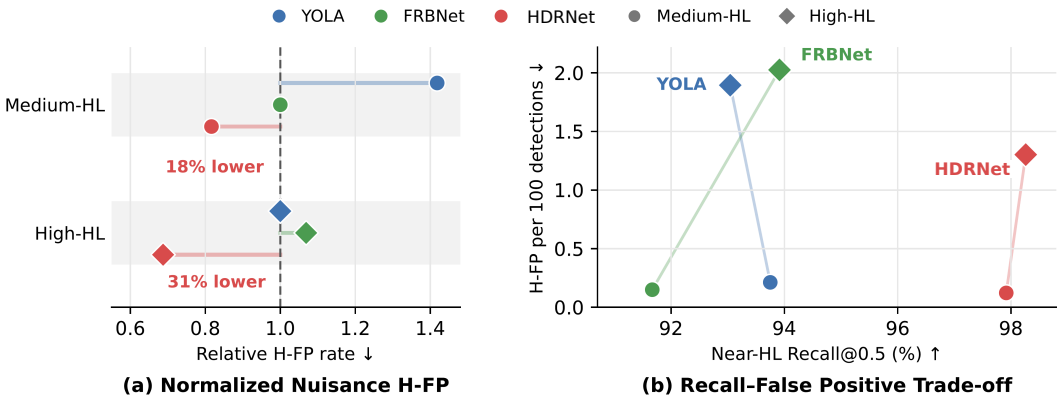


Figure 9: False-positive and recall analysis near highlights. The panels report the normalized nuisance H-FP rate and its trade-off with near-highlight Recall@0.5.

by $+0.85$ points. YOLA and FRBNet change by $+0.25$ and $+0.43$ points. For targets detected in the clean images, HDRNet retains a mean confidence ratio of 1.011. The corresponding ratios for YOLA and FRBNet are 0.999 and 1.006. These results demonstrate stable detection under nonoccluding background highlights. They also confirm the preservation of valid target responses.

B.5 Additional Recognition Tasks

To examine whether the learned correction generalizes beyond detection and segmentation, we further evaluate HDRNet on low-light image classification and video action recognition. Tables 12 and 13 report the results on CoDaN and ARID, respectively.

Table 12: Image classification on CoDaN.

Method	Day Acc (%)	Night Acc (%)
Baseline [CVPR 2016]	74.9	42.9
YOLA [NeurIPS 2024]	75.1	43.9
FRBNet [NeurIPS 2025]	78.4	42.1
HDRNet (Ours)	**80.5**	**46.6**

Table 13: Video recognition on ARID.

Method	Top-1	Top-5
Baseline [ECCV 2016]	26.9	68.5
YOLA [NeurIPS 2024]	30.9	78.1
FRBNet [NeurIPS 2025]	31.7	84.7
HDRNet (Ours)	**35.1**	**87.6**

Image classification on CoDaN. To evaluate transfer to global image-level recognition, we apply HDRNet to the CoDaN classification model without modifying its recognition architecture. As shown in Table 12, HDRNet achieves the highest accuracy on both the day and night splits, reaching 80.5% and 46.6%, respectively. Compared with FRBNet (Sun et al., 2025), HDRNet improves day accuracy by 2.1 points and night accuracy by 4.5 points. Notably, FRBNet falls 0.8 points below the baseline on the night split, whereas HDRNet improves the baseline by 3.7 points. The consistent gains under both lighting conditions suggest that HDRNet mitigates low-light corruption while preserving the appearance cues required for image-level recognition.

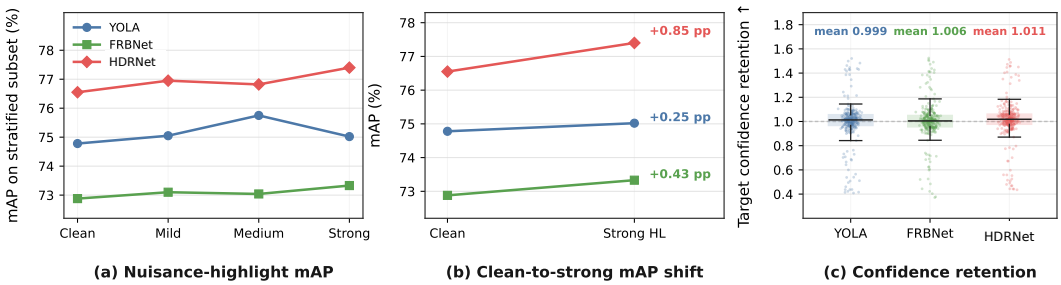


Figure 10: Robustness to synthetic nuisance highlights. Panels show mAP versus perturbation strength, clean-to-strong mAP changes, and confidence retention for clean-image detections.



Figure 11: Dark face detection on DarkFace with VFNet. HDRNet recovers the occluded foreground face and two distant faces under strong storefront illumination.

Video recognition on ARID. To further test transfer to spatiotemporal recognition, we apply HDRNet independently to each video frame without introducing an additional temporal module. Table 13 shows that HDRNet achieves 35.1% top-1 and 87.6% top-5 accuracy. It surpasses FRBNet (Sun et al., 2025) by 3.4 and 2.9 points in top-1 and top-5 accuracy, respectively. It also exceeds YOLA (Hong et al., 2024) by 4.2 and 9.5 points. The gains over the TSN (Wang et al., 2016) baseline reach 8.2 points in top-1 accuracy and 19.1 points in top-5 accuracy. Since the temporal architecture remains unchanged, these improvements indicate that the corrected frame-level representation remains informative after temporal aggregation. Together, the CoDaN and ARID results support the transferability of HDRNet beyond static detection and segmentation.

B.6 Task-Specific Qualitative Results

We complement the quantitative evaluation with representative examples involving severe darkness, localized light sources, narrow semantic structures, and crowded instances. All methods use the same downstream architecture and evaluation protocol and differ only in their low-light front-ends. GT denotes ground truth in all figures. The results are best viewed in color and zoomed in.

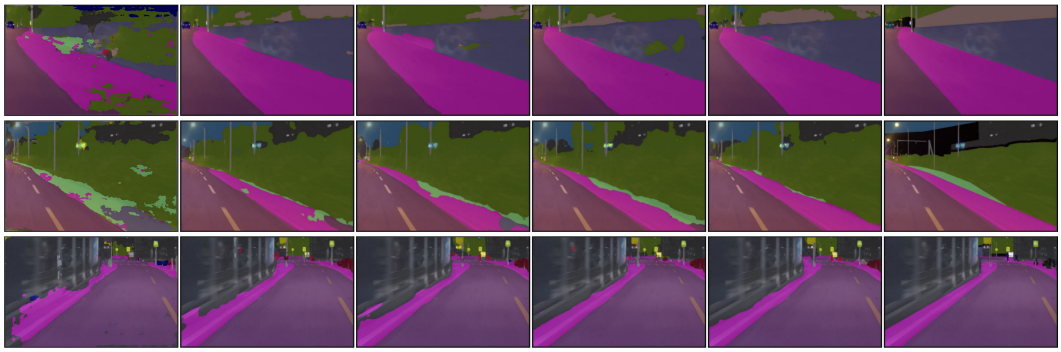


Figure 12: Nighttime semantic segmentation on ACDC-Night. HDRNet produces more continuous road and sidewalk regions with boundaries closer to the ground truth.

Dark face detection with VFNet. Figure 11 shows a challenging DarkFace scene in which bright storefront lights coexist with severely underexposed faces. The baseline and several competing front-ends detect the two distant faces but miss the heavily occluded face on the left. FRBNet (Sun et al., 2025) recovers this face but produces overlapping detections around the same region. In contrast, HDRNet detects all three annotated faces with a single box for each target and increases the confidence of the two distant faces from 0.35 and 0.39 in the baseline to 0.57 and 0.61. This result shows that HDRNet recovers weak facial cues without confusing bright signs and local lights with valid targets.

Nighttime semantic segmentation. Figure 12 compares semantic predictions in three nighttime road scenes. The baseline produces fragmented sidewalk and vegetation regions, particularly along curved road boundaries and weakly illuminated roadside areas. YOLA (Hong et al., 2024), FRBNet (Sun et al., 2025), and FeatEnHancer (Hashmi et al., 2023) reduce some of these fragments but still introduce holes, irregular boundaries, or local class confusion. HDRNet produces more continuous road and sidewalk regions, closely following the curved boundary in the first scene, the narrow roadside strip in the second, and the building-side boundary in the third. These examples indicate that the corrected representation preserves both the global scene layout and thin boundary structures under uneven nighttime illumination.

Low-light instance segmentation. Figure 13 covers three representative challenges on LIS, including small tabletop objects, adjacent parked vehicles, and a large bus surrounded by nearby instances. The baseline and competing front-ends frequently miss small objects, merge neighboring vehicles, or produce masks that extend across object boundaries. HDRNet more clearly separates the bottle from the supporting surface in the first scene, distinguishes adjacent cars and small roadside instances in the second, and aligns the mask of the large bus while retaining neighboring instances in the third. The improvements across objects of different scales show that HDRNet supports both instance localization and boundary delineation in complex low-light scenes.

C Limitations and Discussion

C.1 Limitations

Despite its compact parameter count, the FFT branch of HDRNet introduces additional latency compared with purely spatial front-ends. The current highlight prior relies on fixed intensity and saturation thresholds, which may not adapt optimally to different cameras and scenes and are less reliable for strongly colored light sources. In addition, the controlled mechanism analysis uses synthetic highlights because real images do not provide ground-truth clean ratios or highlight-induced log-domain bias. The highlight-stratified evaluation is limited to ExDark and is constructed using a deterministic prior. Its results should therefore be interpreted as controlled evidence for the proposed mechanism rather than dataset-independent validation.

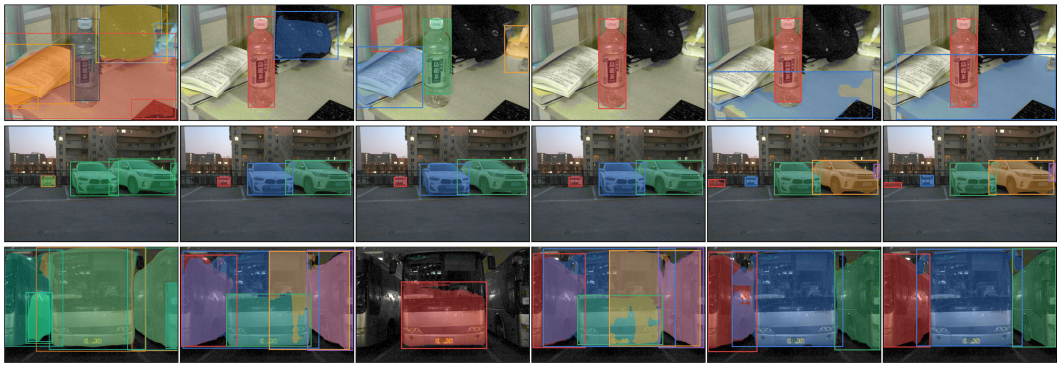


Figure 13: Low-light instance segmentation on LIS. HDRNet improves small-instance recovery, separation of adjacent vehicles, and mask alignment for large objects.

C.2 Future Work

Future work will explore a scene-adaptive and chroma-aware highlight prior learned from weak highlight annotations or paired observations while retaining the bounded log-domain correction. We will also investigate lower-resolution or spatial approximations of the phase-preserving spectral response to reduce the computational cost of the FFT branch. Finally, extending the highlight-stratified protocol to face detection and segmentation will provide broader validation across different target scales and prediction structures.