INTRINSIC FEATURE DECOMPOSITION FOR DOMAIN ADAPTIVE LOW-LIGHT OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Object detection under low-light conditions remains underexplored due to the challenge of learning discriminative representations across extreme illumination variations. We propose a domain-adaptive pretraining framework that learns noise-robust, illumination-invariant feature representations directly within the detector via feature-level decomposition. Inspired by Retinex theory, our approach treats the first-layer backbone feature as intrinsic structure and employs an auxiliary encoder-decoder during pretraining to factor features into intrinsic, illumination, and noise components. The auxiliary branch is discarded after pretraining, incurring no inference overhead. Using paired normal-dark images synthesized from COCO, we jointly optimize detection and decomposition losses to enforce intrinsic consistency. When fine-tuned on real-world low-light datasets, our pretraining consistently improves the COCO-pretrained baseline, with gains of 7.9 mAP on LOD, 4.2 on NOD, 3.7 on NuImages, 2.7 on BDD100K, 11.7 on ExDark, and 5.9 on DARK Face. These gains persist across detector architectures, including YOLO26-m and RT-DETR. Overall, these results show that embedding Retinex principles at the feature level effectively bridges the illumination gap for low-light detection, scales across model sizes and detector architectures, and preserves runtime efficiency. Our code will be publicly available.

1 INTRODUCTION

With the rapid advancement of computer vision, object detection has become one of the most representative tasks (Girshick et al., 2014; Girshick, 2015; Zhang et al., 2022; Redmon, 2018; Jocher et al., 2023; Zhu et al., 2020; Zhao et al., 2024b). Despite remarkable progress under normal illumination conditions, detection performance significantly degrades in low-light environments (Li et al., 2024; Satzoda & Trivedi, 2016; Neumann et al., 2018; Fan et al., 2023; Wang et al., 2024). Low-light scenes typically exhibit diminished signal-to-noise ratio, contrast attenuation, and motion-induced blur, which distort visual representations and weaken discriminative features (Li et al., 2024; Satzoda & Trivedi, 2016; Neumann et al., 2018; Fan et al., 2023; Wang et al., 2024; Chen et al., 2018; Zhang et al., 2019; Goyal et al., 2022; Xu et al., 2020). Such photometric degradation introduces a pronounced domain discrepancy between bright and dark conditions, resulting in significant performance drops for detection models in poorly illuminated scenarios.

To address the challenges of low-light perception, early studies primarily focused on image-level enhancement techniques (Wei et al., 2018; Wu et al., 2022; Cai et al., 2023; Liu et al., 2021a) that aimed to produce brighter images with reduced noise and blurriness. Such approaches often produce visually pleasing results for human observers. However, these improvements in perceptual quality do not always translate into better performance in downstream tasks such as object detection, since the enhancement process could introduce artifacts or distort semantic information which is crucial for recognition (Morawski et al., 2022; Chen et al., 2018; Hashmi et al., 2023; Cui et al., 2024).

Alternatively, pretraining on large-scale datasets followed by finetuning on low-light data has been widely adopted to obtain generalized feature representations and transfer knowledge across domains. Nevertheless, this strategy does not explicitly account for the intrinsic characteristics of low-light environments, thereby yielding only limited improvements in detection performance. Other approaches (Hashmi et al., 2023; Morawski et al., 2022; Peng et al., 2024; Cui et al., 2024) have attempted to extract enhanced features directly from low-light inputs to improve detection perfor-

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.

mance. However, these methods often require additional network layers, which typically cause higher computational cost and longer inference time. Since object detection task has been widely deployed for real-time applications, it is desirable to develop a method that can effectively handle low-light domains without requiring any additional modules during inference.

In this paper, we propose a novel pretraining strategy for object detection models to achieve effective domain adaptation to low-light environments. Inspired by Retinex theory, we introduce illumination-aware feature representations, along with an intrinsic feature decomposition that encourages the network to disentangle these components during pretraining. We further validate our approach by finetuning the pretrained model on several low-light datasets, achieving substantial improvement in detection performance. The main contributions are summarized as follows:

- We introduce a novel pretraining framework for robust low-light object detection. The proposed intrinsic feature decomposition explicitly formulate features into *intrinsic*, *illumination*, and *noise* components, allowing the network to learn illumination-invariant representations under low-light conditions.
- An auxiliary module for extracting illumination and noise features is employed solely in the pretraining phase and removed at finetuning, resulting in no additional computational cost or inference latency.
- The proposed strategy significantly outperforms existing methods across multiple low-light datasets, achieving state-of-the-art performance while maintaining strong generalization and scalability across diverse domains.

2 Related work

2.1 Low-Light Image Enhancement (LLIE)

A large body of work enhances low-light images before downstream processing, typically by learning a mapping to normal-light appearance from paired data (Chen et al., 2018; Zhang et al., 2019; Wu et al., 2022; Cai et al., 2023; Zhou et al., 2024; Yan et al., 2025; Feijoo et al., 2025). Building on Retinex theory, Wei et al. (2018) decomposed an image into reflectance and illumination and enhanced the illumination component, a strategy later extended with transformer-based long-range modeling (Cai et al., 2023) and frequency-aware restoration (Feijoo et al., 2025). To avoid the cost of paired supervision, self-supervised and unpaired alternatives have also been explored (Jiang et al., 2021; Liu et al., 2021b; Zhang et al., 2020; Liu et al., 2021a; Guo et al., 2020; Li et al., 2021; Ma et al., 2022). Regardless of supervision type, these methods operate purely at the *image level*: the resulting perceptual gains do not necessarily translate into better performance on downstream tasks such as object detection, and can even introduce artifacts that hurt recognition.

2.2 Domain Adaptation for Low-Light Detection

An alternative direction adapts the detector itself to low-light conditions rather than enhancing the input image. Early approaches fine-tune architecturally modified detectors on low-light data (Morawski et al., 2022; Yin et al., 2023; Zhao et al., 2024a) or learn representations that stay consistent across lighting conditions (Wang et al., 2021; Lengyel et al., 2021; Luo et al., 2023; Sasagawa & Nagahara, 2020), but remain limited by the scarcity of labeled low-light data. More closely related, three recent methods pursue illumination robustness through distinct mechanisms: MAET (Cui et al., 2021) learns a self-supervised auto-encoding transformation task and disentangles object- from degradation-related information via an orthogonality constraint in feature space; DAI-Net (Du et al., 2024) lets the detector's shallow backbone feature additionally decode a Retinex reflectance target supplied by a separately pretrained decomposition network; and YOLA (Hong et al., 2024) derives illumination-invariant feature maps from raw pixel neighborhoods via a Lambertian-inspired module and fuses them with the input image ahead of an otherwise unmodified backbone. In each case, invariance is introduced through a generic pretext task, an externally supplied decomposition target, or a hand-crafted input-level transform, rather than a decomposition that uses the Retinex structure as an inductive bias and is learned end to end within the detector's own feature hierarchy.

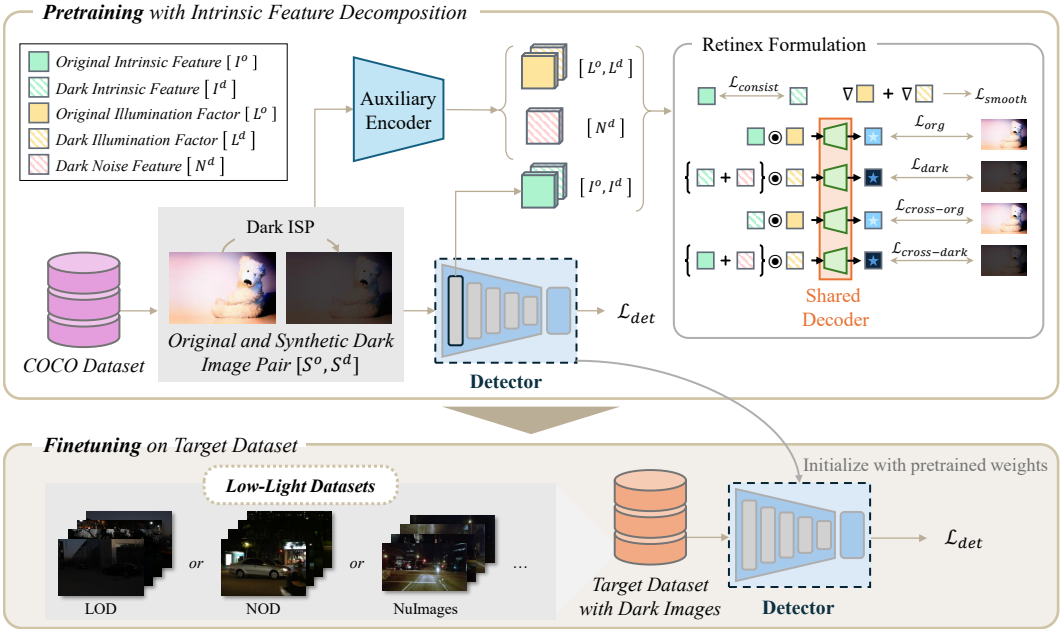

Figure 1: **Overview of the proposed pretraining framework based on feature-level Retinex decomposition.** Synthetic original–dark image pairs are generated from COCO using a Dark ISP pipeline. The original and low-light images are processed by the YOLO backbone, where the first-layer features serve as intrinsic features I. A lightweight auxiliary encoder extracts illumination factor L and noise feature N. A shared decoder reconstructs images via the feature-level Retinex formulation $(I+N)\odot L$, enabling the model to learn illumination-invariant intrinsic representations. During pretraining, both detection and decomposition losses are jointly optimized. The pretrained weights are then fine-tuned on target low-light datasets without any auxiliary modules at inference.

We instead reformulate the detector's own backbone feature directly as the intrinsic representation, jointly learning it together with illumination and noise components through a shared reconstruction objective. This requires no pretrained decomposition network, input transform, or auxiliary modules at inference, and it directly supervises the representation the detector actually uses to be illumination- and noise-invariant.

3 Method

3.1 Revisiting Retinex Theory

The Retinex theory (Brainard & Wandell, 1986), inspired by the human visual system, models the formation of an image as the element-wise product of two components: the intrinsic reflectance R (also referred to as intrinsic feature) and the illumination L. Formally, a captured image S can be expressed as

$$S = R \odot L, \tag{1}$$

where R corresponds to the inherent surface property of the scene and L represents the spatially varying illumination. Retinex-based methods have been widely used to enhance low-light images by estimating and removing the effects of L, thereby obtaining an illumination-invariant image.

Recent deep learning approaches (Wei et al., 2018; Wu et al., 2022; Liu et al., 2021b; Ma et al., 2022; Cai et al., 2023) have operationalized this idea using neural networks to decompose a low-light image into illumination and reflectance maps, which are then used to generate enhanced inputs for downstream perception tasks such as object detection. However, since the decomposition occurs outside the detector, it does not affect the detector's internal feature representations, and image-level enhancement often introduces artifacts or alters local textures, leading to unstable detection results under challenging illumination conditions.

3.2 Intrinsic Feature Decomposition

In the classical Retinex formulation, an image S is expressed as the element-wise product of its reflectance R and illumination L. We extend this concept to the *feature domain* by representing the effect of illumination as a multiplicative factor. The intrinsic feature $I \in \mathbb{R}^{H\times W\times C}$ is intended to capture semantic information that is invariant to illumination and noise. Under low-light conditions, reduced signal levels and ISP amplification make noise degradation non-negligible. We therefore introduce a noise feature N, together with an illumination factor L that modulates feature responses. The resulting intrinsic feature decomposition formulation is

$$F = (I + N) \odot L, \tag{2}$$

where F denotes the composed feature and $\odot$ denotes element-wise multiplication.

For normally exposed inputs, the relatively high signal-to-noise ratio limits the effect of sensor and ISP noise on the feature representation; thus, we assume $N^o \approx 0$. In contrast, noise becomes prominent under low-light conditions and introduces feature discrepancies that cannot be explained by illumination modulation alone. Explicitly modeling N^d prevents such degradation from being absorbed into the intrinsic feature, encouraging I to remain stable across illumination conditions.

Given paired normal–dark inputs (S^o, S^d), the corresponding features are defined as

$$F^o = I^o \odot L^o, \tag{3}$$

$$F^d = (I^d + N^d) \odot L^d. \tag{4}$$

Here, F, I, L, and N denote feature-domain quantities rather than image-level Retinex components. They are latent factors shaped by this structure and the objectives in Sec. 3.4, not physically measured signals.

3.3 Network Architecture

Building upon the proposed Intrinsic Feature Decomposition, we integrate it into the detector backbone during pretraining while preserving the original inference architecture. Rather than explicitly decomposing a backbone feature F into (I, L, N) using additional modules, we use the output of the first convolutional layer as the intrinsic feature I, since it preserves fine-grained information from the input. The training objectives then encourage I to become invariant to illumination and noise.

To estimate the remaining components, we employ a lightweight auxiliary encoder consisting of three convolutional layers only during pretraining. The encoder predicts an illumination factor L and a noise feature N, both having the same spatial and channel dimensions as I. These components are combined according to Eq. 2, allowing illumination-dependent modulation and noise degradation to be explicitly modeled.

A shared decoder $\mathcal{D}$ maps the composed features back to the image space. Specifically, the original and dark inputs are reconstructed as $\mathcal{D}(I^o \odot L^o)$ and $\mathcal{D}((I^d + N^d) \odot L^d)$, respectively. These reconstruction objectives anchor the feature decomposition to the input images. Cross reconstruction further swaps I^o and I^d while retaining the illumination and noise components of each reconstruction branch, encouraging the intrinsic features to remain consistent across illumination conditions.

During pretraining, the detection and decomposition objectives are jointly optimized using original and synthetic low-light pairs (S^o, S^d) generated by the Dark ISP pipeline following MAET (Cui et al., 2021). After pretraining, the auxiliary encoder and decoder are discarded, and the pretrained detector is fine-tuned on real low-light datasets without additional inference cost.

3.4 Decomposition Loss

Based on the pretraining framework described in Section 3.3, we design a decomposition objective that separates intrinsic structure from illumination-dependent modulation and noise degradation while preserving the information required to reconstruct the input images. Given (I^o, L^o) for the original input and (I^d, L^d, N^d) for the dark input, the total decomposition loss is defined using the shared decoder $\mathcal{D}$ as

$$\begin{aligned}\mathcal{L}_{\text{decom}} = \mathcal{L}_{\text{org}} + \mathcal{L}_{\text{dark}} + \lambda_{\text{consist}}\mathcal{L}_{\text{consist}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} \\ + \lambda_{\text{cross}}\left(\mathcal{L}_{\text{cross-org}} + \mathcal{L}_{\text{cross-dark}}\right).\end{aligned} \tag{5}$$

Each loss term is defined as follows:

- **Original reconstruction loss** reconstructs the normally exposed input:

$$\mathcal{L}_{\text{org}} = \|S^o - \mathcal{D}(I^o \odot L^o)\|_1 . \tag{6}$$

- **Dark reconstruction loss** reconstructs the synthetic low-light input while accounting for noise degradation:

$$\mathcal{L}_{\text{dark}} = \left\|S^d - \mathcal{D}\big((I^d + N^d) \odot L^d\big)\right\|_1 . \tag{7}$$

- **Intrinsic consistency loss** aligns the intrinsic features of the paired inputs:

$$\mathcal{L}_{\text{consist}} = \left\|I^o - I^d\right\|_1 . \tag{8}$$

- **Illumination-factor smoothness loss** regularizes the spatial variation of the illumination factors:

$$\mathcal{L}_{\text{smooth}} = \sum_{k\in\{o,d\}} \left(\left\|\nabla_x L^k\right\|_1 + \left\|\nabla_y L^k\right\|_1\right) . \tag{9}$$

  This regularization discourages abrupt changes in the modulation factors and encourages structural details to be encoded by I.

- **Cross-reconstruction loss** swaps the intrinsic features while retaining the illumination and noise components of each reconstruction branch:

$$\begin{aligned} \mathcal{L}_{\text{cross-org}} &= \left\|S^o - \mathcal{D}(I^d \odot L^o)\right\|_1 , \\ \mathcal{L}_{\text{cross-dark}} &= \left\|S^d - \mathcal{D}\big((I^o + N^d) \odot L^d\big)\right\|_1 . \end{aligned} \tag{10}$$

  This objective encourages I to remain consistent across illumination conditions while preserving condition-specific modulation and noise.

We set $\lambda_{\text{consist}} = 0.05$, $\lambda_{\text{smooth}} = 0.1$, and $\lambda_{\text{cross}} = 0.05$. The decomposition loss is jointly optimized with the detection loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{decom}}. \tag{11}$$

This joint objective promotes intrinsic features that are robust to illumination changes and noise degradation without modifying the inference-time architecture.

4 Experiments

This section evaluates the proposed pretraining framework on low-light object detection benchmarks (LOD, NOD), real-world driving datasets (NuImages, BDD100K), and additional detectors and benchmarks, followed by feature-level analysis and ablation studies. Additional implementation details are provided in the appendix.

4.1 Evaluations on the LOD Dataset

The LOD dataset (Hong et al., 2021) consists of 2,230 short / long-exposure image pairs captured under various low-light and normal lighting conditions. It provides paired ground-truth images and bounding box annotations for multiple object categories, making it well suited for fair comparisons with both supervised enhancement methods and Retinex decomposition-based approaches.

Settings For all experiments, we adopt the YOLOv8-m detector (Jocher et al., 2023) as a unified detection backbone. We randomly split the dataset into 80% for training and 20% for validation. All models are fine-tuned on the LOD for 50 epochs.

Results Table 1 summarizes the detection performance on the LOD dataset. Compared with the baseline detector, existing low-light image enhancement approaches, including both self-supervised (Guo et al., 2020; Liu et al., 2021b; Ma et al., 2022) and supervised methods (Cai et al., 2023; Zhou et al., 2024; Feijoo et al., 2025; Yan et al., 2025), yield only marginal improvements or even slight performance degradation. This observation indicates that improving the visual appearance of dark images through image enhancement does not necessarily translate into better detection performance. Domain adaptation-based approaches (Cui et al., 2021; Du et al., 2024; Hong et al.,

Table 1: Detection performance (mAP) on the LOD and NOD datasets.

		LOD										NOD				
Approach	**Method**	**Bicy.**	**Bus**	**Bottle**	**Car**	**Chair**	**Table**	**M.bike**	**TV**	**mAP**		**Pers.**	**Bicy.**	**Car**	**mAP**	
Baseline	YOLOv8-m	58.7	68.0	52.2	71.6	55.0	50.8	45.7	44.8	55.8	–	52.8	44.1	48.0	48.3	–
Self-sup. LLIE	Zero-DCE	59.0	70.1	51.6	71.4	53.9	43.0	43.3	50.0	55.3	(-0.5)	52.4	43.9	47.5	47.9	(-0.4)
	RUAS	57.7	69.1	49.0	72.8	54.1	48.7	47.9	53.2	56.6	(+0.8)	50.5	43.3	45.7	46.5	(-1.8)
	SCI	58.5	68.4	54.7	72.7	53.4	42.7	46.5	47.3	55.5	(-0.3)	51.7	44.3	46.2	47.4	(-0.9)
Sup. LLIE	Retinexformer	59.0	67.2	50.8	71.8	52.0	43.8	47.7	52.6	55.6	(-0.2)	52.4	44.5	49.4	48.8	(+0.5)
	GLARE	56.8	70.9	54.1	73.2	53.9	44.4	47.4	54.0	56.8	(+1.0)	52.9	44.6	47.2	48.2	(-0.1)
	DarkIR	58.4	69.1	52.2	72.5	54.4	44.9	47.0	53.5	56.5	(+0.7)	53.1	43.7	47.0	48.0	(-0.3)
	HVI-CIDNet	59.1	69.3	53.6	72.5	54.0	46.1	46.3	53.0	56.7	(+0.9)	52.4	44.9	45.7	47.7	(-0.6)
Domain Adaptation	MAET	61.0	**72.6**	60.3	73.2	57.7	53.8	48.3	62.3	61.1	(+5.3)	54.2	45.1	51.7	50.4	(+2.1)
	DAI-Net	61.6	69.6	**63.1**	72.9	58.0	53.4	51.7	60.8	61.4	(+5.6)	53.8	45.3	52.3	50.5	(+2.2)
	YOLA	61.5	70.5	60.8	72.5	60.3	51.9	50.6	61.4	61.2	(+5.4)	54.0	46.2	52.6	50.9	(+2.6)
	Ours	**64.3**	71.5	62.5	**74.5**	**61.0**	**54.9**	**54.0**	**66.8**	**63.7**	**(+7.9)**	**55.4**	**46.7**	**55.4**	**52.5**	**(+4.2)**

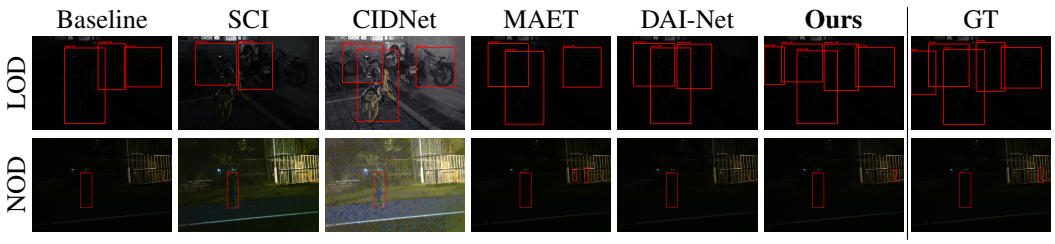

Figure 2: Qualitative results after finetuning on the NOD and LOD datasets.

2024) provide noticeable gains, achieving around 61 mAP. In contrast, our domain adaptive pretraining strategy brings a substantial improvement of +7.9 mAP over the baseline, reaching 63.7 mAP. This represents the highest performance among all compared methods across most object categories, as shown in Table 1.

Fig. 3 further presents the detection performance across different YOLOv8 model variants (n, s, m, l, x), illustrating the mAP envelope (shaded region) between the baseline and our pretraining approach, with every variant fine-tuned on LOD for the same number of epochs from its respective pretrained weights. Notably, the baseline detectors pretrained on COCO exhibit a clear *saturation trend*: performance improvements taper off from the s model onward, indicating that larger model capacities are not effectively utilized under low-light conditions. In contrast, our domain-adaptive pretraining enables steady performance gains as the model size increases, fully leveraging the representational capacity of larger models. These results highlight a key advantage of the proposed approach—rather than being limited by the domain gap between pretraining and target low-light data, the detector with our proposed pretraining strategy effectively scales with model complexity, achieving significant improvements.

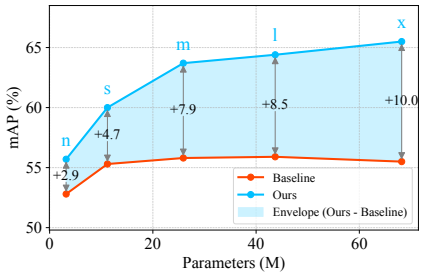

Figure 3: mAP across YOLOv8 model scales.

4.2 Evaluations on the NOD Dataset

The NOD dataset (Morawski et al., 2021) consists of images captured under various illumination levels using the Sony RX100 VII camera. It contains bounding box annotations (person, bicycle, car), covering a wide range of low-light conditions, from moderately dark scenes to extremely low-light environments. Unlike LOD (Hong et al., 2021), NOD does not provide paired normal-light ground-truth images, making it suitable for evaluating model generalization under realistic low-light conditions.

Settings All models were trained or fine-tuned using the YOLOv8-m detector structure. We fine-tuned all models on the NOD dataset for 25 epochs.

Results Table 1 presents the comparison results on the NOD dataset. The baseline YOLOv8-m model achieved 48.3 mAP. Enhancement methods (Guo et al., 2020; Liu et al., 2021b; Ma et al., 2022; Cai et al., 2023; Zhou et al., 2024; Feijoo et al., 2025; Yan et al., 2025) showed marginal improvements or slight degradation, indicating that image-level enhancement alone is insufficient for improving detection robustness in real low-light conditions. Domain adaptation methods (Cui et al., 2021; Du et al., 2024; Hong et al., 2024) achieved moderate gains of +2.1 mAP, +2.2 mAP and +2.6 mAP, respectively. In contrast, our pretraining method delivered the best performance, reaching 52.5 mAP and outperforming the baseline by +4.2 mAP. Note that improvements are consistent across all classes: person (+2.6), bicycle (+2.6), and car (+7.4) compared to the baseline.

4.3 Adaptation to Night Driving Scenes

Nighttime driving scenarios pose unique challenges for object detectors due to complex illumination variations, strong headlights, specular reflections, and motion blur. To further validate the generalization capability of our pretrained detector in realistic, large-scale autonomous driving environments, we evaluate the proposed method on the NuImages (Caesar et al., 2020) and BDD100K (Yu et al., 2020) datasets, which contain diverse urban scenes captured under various low-light conditions.

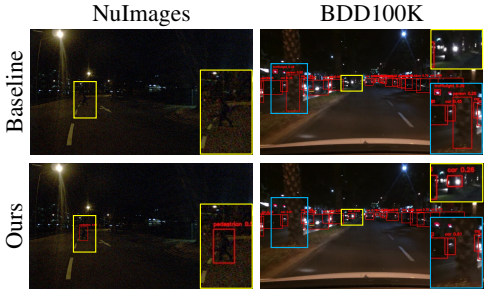

Figure 4: Qualitative results on night driving.

Settings For NuImages, we first parsed the dataset to extract nighttime scenes based on timestamp metadata, resulting in 4,485 night-time images. We then re-split[1] this subset into training and validation sets following the original ratio, and fine-tuned the detectors for 25 epochs. For BDD100K, we selected night images based on metadata and performed fine-tuning exclusively on this subset. The training split contains approximately 28,000 images, and we fine-tuned the models for 10 epochs to match the larger dataset size. In both datasets, we adopted the YOLOv8-m detector as the backbone, initialized with either the standard COCO-pretrained weights or our domain-adaptive pretrained weights.

Results Fig. 4 summarizes the results. On NuImages, our method improves the mAP from 29.0 to 32.7, achieving a +3.7 mAP gain over the baseline. On BDD100K nighttime scenes, it improves the mAP from 24.3 to 27.0, outperforming the baseline by +2.7 mAP and demonstrating effective adaptation to unseen driving domains. Qualitatively, our method exhibits greater robustness to blurry objects, suppressing false positives (blue boxes) and correctly detecting pedestrians and vehicles (yellow boxes).

These results highlight the practicality of the proposed framework for real-world autonomous driving applications. By learning illumination-invariant intrinsic features during pretraining, the model effectively bridges the domain gap between normal and low-light driving scenes, leading to improved detection robustness in nighttime environments.

4.4 Generalization Across Detectors and Benchmarks

The evaluations above are all based on a single detector, YOLOv8-m. To test whether the benefits of our pretraining hold beyond one architecture and beyond LOD/NOD, we extend the evaluation to two further detectors and two further low-light benchmarks: ExDark (Loh & Chan, 2019) and DARK Face (Yang et al., 2020), both widely used to assess low-light detection robustness.

Settings In addition to YOLOv8-m (Jocher et al., 2023), we evaluate a more recent real-time CNN detector, YOLO26-m (Jocher et al., 2026), and a transformer-based detector, RT-DETR (Zhao et al., 2024b). For each detector, we apply the same COCO pretraining protocol described above and then fine-tune the COCO-pretrained baseline and our pretrained weights on LOD, NOD, ExDark, and DARK Face, following the same per-dataset fine-tuning schedule.

[1]Since the original validation set does not include night scenes, we re-split the dataset.

Results Table 2 summarizes the results. Our pretraining improves every detector on every benchmark. For YOLO26-m, a substantially stronger COCO-pretrained baseline, still benefits clearly from our pretraining (up to +8.7 mAP on ExDark and +4.5 mAP on DARK Face), showing that the learned intrinsic representation is not tied to a specific detector design. Notably, the transformer-based RT-DETR also improves consistently (+3.3 to +7.4 mAP across the four benchmarks), indicating that the proposed pretraining generalizes beyond CNN-based detectors to attention-based architectures as well. Together with the additional gains observed on ExDark and DARK Face, these results confirm that the proposed intrinsic feature decomposition is a broadly applicable pretraining strategy, independent of the specific detector architecture or target low-light benchmark.

Table 2: Generalization across detectors and benchmarks.

Detector	Method	LOD	NOD	ExDark	DARK Face
YOLOv8-m	Baseline	55.8	48.3	67.4	65.6
	Ours	**63.7**	**52.5**	**79.1**	**71.5**
YOLO26-m	Baseline	57.1	50.1	72.9	69.1
	Ours	**63.0**	**52.3**	**81.6**	**73.6**
RT-DETR	Baseline	48.5	46.9	66.7	63.4
	Ours	**51.8**	**51.0**	**74.1**	**67.9**

4.5 Feature Robustness Analysis

To quantify the stability of intrinsic features under illumination degradation, we compute the Pearson correlation ($\rho^{(l)}$) and cosine similarity ($s^{(l)}$) between the feature maps of the original and degraded images. Let $F^{(0)}, F^{(l)} \in \mathbb{R}^{C\times H\times W}$ denote the feature maps from the original and degraded image at level l, respectively. After vectorization $\mathbf{f}^{(0)} = \mathrm{vec}(F^{(0)})$ and $\mathbf{f}^{(l)} = \mathrm{vec}(F^{(l)})$, the metrics are defined as

$$\rho^{(l)} = \frac{(\mathbf{f}^{(0)} - \bar{\mathbf{f}}^{(0)})^\top (\mathbf{f}^{(l)} - \bar{\mathbf{f}}^{(l)})}{\|\mathbf{f}^{(0)} - \bar{\mathbf{f}}^{(0)}\|_2 \|\mathbf{f}^{(l)} - \bar{\mathbf{f}}^{(l)}\|_2}, \quad s^{(l)} = \frac{\mathbf{f}^{(0)\top} \mathbf{f}^{(l)}}{\|\mathbf{f}^{(0)}\|_2 \|\mathbf{f}^{(l)}\|_2}.$$

Higher values of ρ and s indicate stronger feature consistency under illumination degradation. Fig. 5 shows the mean and variance of both metrics across degradation levels on the COCO validation split. As degradation increases, the baseline exhibits a sharp decline in both metrics, whereas our method shows only a gradual decrease, indicating more stable intrinsic feature representations even under strong degradation. Qualitative examples are shown in Fig. 6.

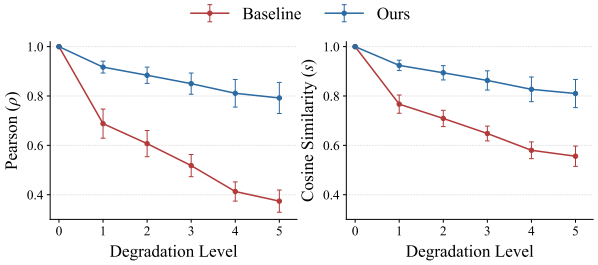

Figure 5: Feature Robustness Evaluation.

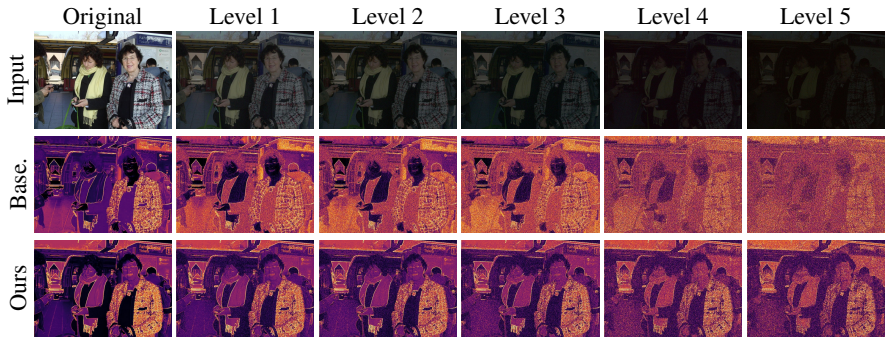

Figure 6: Visualization of feature representations across different degradation levels. *Zoom in for best view.*

Table 3: Ablation study of each component.

Method	Dark ISP	Feat. Consist.	Decomposition		Noise Comp.	mAP
			image-level	feature-level		
Baseline	–	–	–	–	–	55.8
	✓	–	–	–	–	58.1
	✓	✓	–	–	–	58.4
	✓	–	✓	–	–	57.2
	✓	✓	–	✓	–	62.5
Ours	✓	✓	–	✓	✓	**63.7**

Table 4: Effect of the backbone stage for the intrinsic feature.

Method	B5	B3	B1	mAP
Baseline	–	–	–	55.8
	✓	–	–	62.8
	–	✓	–	63.4
Ours	–	–	✓	**63.7**

4.6 Ablation Study

For ablation study, we vary the pretraining strategies and compare the detection performance of models finetuned on the LOD dataset. Finetuning is performed under controlled conditions, where only the pretraining stage is altered. The results are summarized in Table 3.

Compared with the baseline detector pretrained on normal-light data, applying Dark ISP to generate darkened image pairs leads to a clear improvement (+2.3 mAP), demonstrating that low-light augmentation during pretraining benefits subsequent fine-tuning. Adding only the consistency loss $\mathcal{L}_{\text{consist}}$ between the paired first-layer features yields a negligible further gain (+0.3 over Dark ISP alone), indicating that simply aligning paired features does not account for the improvement. Similarly, incorporating image-level Retinex decomposition yields only marginal gains, suggesting that pixel-level illumination correction alone is insufficient for learning robust representations.

In contrast, introducing intrinsic feature decomposition, which uses the same consistency loss, substantially improves detection performance (+6.7 mAP) by enabling the model to disentangle illumination and intrinsic information directly in the feature space. Furthermore, explicitly modeling the noise component provides an additional gain (+7.9 mAP), resulting in the best overall performance. These results confirm that combining intrinsic feature formulation with noise-aware representation learning significantly enhances robustness to illumination degradation.

In addition, we analyze the effect of extracting intrinsic features from different backbone stages. In our framework, intrinsic features are extracted from the first layer of the backbone (B1). To examine whether this choice is optimal, we also perform pretraining using intrinsic features from the third layer (B3) and the fifth layer (B5), and compare their performance after fine-tuning on the LOD dataset. As shown in Table 4, using B1 achieves the best performance with 63.7 mAP. This result suggests that earlier backbone features preserve richer structural information, which tends to degrade in deeper layers due to increasing semantic abstraction. Moreover, improving the quality of early backbone features can propagate beneficial effects to subsequent layers, leading to stronger feature representations throughout the network.

5 Conclusion

We present a pretraining strategy for low-light object detection that learns illumination-invariant intrinsic representations without inference-time overhead. Using COCO with Dark ISP pairs and an auxiliary encoder–decoder optimized alongside detection, the approach consistently outperforms enhancement and domain adaptation baselines across various benchmark datasets. On LOD, it reveals a clear mAP envelope across YOLOv8 variants, overcoming the saturation of COCO-pretrained baselines and scaling gains with model capacity. A limitation shared with related domain-adaptive methods (e.g., MAET, DAI-Net) is that pretraining relies on synthetic dark images from the Dark ISP pipeline, leaving a residual synthetic-to-real gap. Also, the auxiliary encoder–decoder adds computational cost during pretraining despite no inference overhead. Despite these limitations, our proposed method generalizes robustly to real-world low-light benchmarks (e.g., NOD, NuImages, BDD100K), leaving the exploration of narrowing synthetic-to-real gap through more realistic noise/degradation modeling as a future work.

AI USE STATEMENT

In this work, we used generative AI tools for adapting baseline implementations to our experimental setup. We have not used generative AI tools for generating synthetic datasets, developing the conceptual framework, proposing or refining hypotheses, designing the research methodology or experiments, implementing the proposed method, translation, cleaning or reformatting datasets, or interpreting results, and formulating mathematical claims, writing proofs, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for experimental scripting and for drafting and editing parts of the manuscript. We have reviewed all AI-assisted work: the AI-assisted code for baseline adaptations and experimental scripts was verified and tested for correctness by the authors, and all AI-drafted or edited text was revised by the authors, who checked every number, claim, and reference against our results and the original sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 specifies the decomposition, architecture and loss weights, and Section 4 describes the datasets, splits, baselines, evaluation metric and fine-tuning schedules. Appendix A details the Dark ISP synthesis, degradation levels and auxiliary encoder–decoder, and Appendix B lists every pre-training and fine-tuning hyperparameter together with the per-dataset epochs, resolution and batch settings. All compared methods share the same detector, fine-tuning schedule and settings on each dataset, and fixed random seeds make the Dark ISP synthesis deterministic in the robustness analysis. All experimental code and model checkpoints will be publicly released.

REFERENCES

David H Brainard and Brian A Wandell. Analysis of the retinex theory of color vision. *Journal of the Optical Society of America A*, 3(10):1651–1661, 1986.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3291–3300, 2018.

Xiaohan Cui, Long Ma, Tengyu Ma, Jinyuan Liu, Xin Fan, and Risheng Liu. Trash to treasure: Low-light object detection via decomposition-and-aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2553–2562, 2021.

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, 2024.

Zhentao Fan, Xianhao Wu, Xiang Chen, and Yufeng Li. Learning to see in nighttime driving scenes with inter-frequency priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4218–4225, 2023.

Daniel Feijoo, Juan C Benito, Alvaro Garcia, and Marcos V Conde. Darkir: Robust low-light image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10879–10889, 2025.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

Bhavya Goyal, Jean-François Lalonde, Yin Li, and Mohit Gupta. Robust scene inference under noise-blur dual corruptions. In *2022 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–12. IEEE, 2022.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1780–1789, 2020.

Khurram Azeem Hashmi, Goutham Kallempudi, Didier Stricker, and Muhammad Zeshan Afzal. Featenhancer: Enhancing hierarchical features for object detection and beyond under low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6725–6735, 2023.

Mingbo Hong, Shen Cheng, Haibin Huang, Haoqiang Fan, and Shuaicheng Liu. You only look around: learning illumination-invariant feature for low-light object detection. *Advances in neural information processing systems*, 37:87136–87158, 2024.

Yang Hong, Kaixuan Wei, Linwei Chen, and Ying Fu. Crafting object detection in very low light. In *British Machine Vision Conference*, 2021.

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.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. URL `https://github.com/ultralytics/ultralytics`.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics yolo26: unified real-time end-to-end vision models. *arXiv preprint arXiv:2606.03748*, 2026.

Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4399–4409, 2021.

Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(8):4225–4238, 2021.

Zhiqiang Li, Jian Xiang, and Jiawen Duan. A low illumination target detection method based on a dynamic gradient gain allocation strategy. *Scientific Reports*, 14(1):29058, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10561–10570, 2021a.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10561–10570, 2021b.

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.

Rundong Luo, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Similarity min-max: Zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8104–8114, 2023.

Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Igor Morawski, Yu-An Chen, Yu-Sheng Lin, and Winston H Hsu. Nod: Taking a closer look at detection under extreme low-light conditions with night object detection dataset. *arXiv preprint arXiv:2110.10364*, 2021.

Igor Morawski, Yu-An Chen, Yu-Sheng Lin, Shusil Dangi, Kai He, and Winston H Hsu. Genisp: Neural isp for low-light machine cognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 630–639, 2022.

Lukáš Neumann, Michelle Karg, Shanshan Zhang, Christian Scharfenberger, Eric Piegert, Sarah Mistr, Olga Prokofyeva, Robert Thiel, Andrea Vedaldi, Andrew Zisserman, et al. Nightowls: A pedestrians at night dataset. In *Asian Conference on Computer Vision*, pp. 691–705. Springer, 2018.

Daxin Peng, Wei Ding, and Tong Zhen. A novel low light object detection method based on the yolov5 fusion feature enhancement. *Scientific reports*, 14(1):4486, 2024.

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

Yukihiro Sasagawa and Hajime Nagahara. Yolo in the dark-domain adaptation method for merging multiple models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pp. 345–359. Springer, 2020.

Ravi Kumar Satzoda and Mohan Manubhai Trivedi. Looking at vehicles in the night: Detection and dynamics of rear lights. *IEEE Transactions on Intelligent Transportation Systems*, 20(12): 4297–4307, 2016.

Wenjing Wang, Wenhan Yang, and Jiaying Liu. Hla-face: Joint high-low adaptation for low light face detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16195–16204, 2021.

Xiaofeng Wang, Xiao Liu, Hong Yang, Zhengyong Wang, Xiaoyue Wen, Xiaohai He, Linbo Qing, and Honggang Chen. Degradation modeling for restoration-enhanced object detection in adverse weather scenes. *IEEE Transactions on Intelligent Vehicles*, 2024.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018.

Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5901–5910, 2022.

Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2281–2290, 2020.

Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. Hvi: A new color space for low-light image enhancement. In *Proceedings of the computer vision and pattern recognition conference*, pp. 5678–5687, 2025.

Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5525–5533, 2016.

Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020.

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*, pp. 163–174. Springer, 2023.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

Wenqiang Zhang, Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Qian Zhang, and Wenyu Liu. Featurized query r-cnn. *arXiv preprint arXiv:2206.06258*, 2022.

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*, pp. 1632–1640, 2019.

Yu Zhang, Xiaoguang Di, Bin Zhang, and Chunhui Wang. Self-supervised image enhancement network: Training with low light images only. *arXiv preprint arXiv:2002.11300*, 2020.

Dewei Zhao, Faming Shao, Sheng Zhang, Li Yang, Heng Zhang, Shaodong Liu, and Qiang Liu. Advanced object detection in low-light conditions: Enhancements to yolov7 framework. *Remote Sensing*, 16(23), 2024a.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16965–16974, 2024b.

Han Zhou, Wei Dong, Xiaohong Liu, Shuaicheng Liu, Xiongkuo Min, Guangtao Zhai, and Jun Chen. Glare: Low light image enhancement via generative latent feature based codebook retrieval. In *European Conference on Computer Vision*, pp. 36–54. Springer, 2024.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

A Additional Implementation Details

A.1 DarkISP Implementation Details

To construct paired normal-/low-light training samples during pretraining, we generate a synthetic dark image S^d from each normal-light RGB image S^o using a Dark ISP pipeline. We follow the DarkISP synthesis pipeline adopted in MAET (Cui et al., 2021). The implementation follows a three-stage procedure: RGB-to-RAW unprocessing, low-light corruption, and RAW-to-RGB ISP rendering.

A.1.1 RGB-to-RAW Unprocessing

Given an input RGB image $S^o \in \mathbb{R}^{3\times H\times W}$, we first approximate the inverse camera ISP. Specifically, we apply inverse tone mapping, inverse gamma correction, color space conversion to a camera space, and inverse white balance. The gamma value is sampled from $\gamma \sim \mathcal{U}(2.0, 3.5)$. For color conversion, we randomly select one camera color correction matrix (CCM) from a predefined set of four matrices and combine it with the standard RGB-to-XYZ transform to obtain an RGB-to-camera mapping. Inverse white balance is applied using randomly sampled gains, where the global RGB gain follows $\mathcal{N}(0.8, 0.1^2)$, while the red and blue channel gains are sampled from $\mathcal{U}(1.9, 2.4)$ and $\mathcal{U}(1.5, 1.9)$, respectively.

A.1.2 Low-light Corruption

To simulate low-photon imaging conditions, the RAW signal is attenuated by a darkness factor α sampled from a truncated normal distribution within $[0.01, 0.1]$. The attenuation is applied multiplicatively to the RAW signal:

$$x_d = \alpha x,$$

where x denotes the RAW signal and x_d denotes the darkened signal. The noise scale β multiplicatively scales both shot and read noise to control the overall corruption level. The shot noise level is sampled from a log-uniform distribution $\log \sigma^2_{\text{shot}} \sim \mathcal{U}(\log 10^{-4}, \log 0.012)$, while the read noise follows a log-log linear relation

$$\log \sigma^2_{\text{read}} = 2.18 \log \sigma^2_{\text{shot}} + 1.20 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.26^2).$$

The total noise variance is modeled as

$$\text{Var}(x_d) = x_d \, \sigma^2_{\text{shot}} + \sigma^2_{\text{read}},$$

where x_d denotes the darkened RAW signal. Gaussian noise with this variance is added pixel-wise. We further inject quantization noise using a uniform distribution, where the quantization bit depth is randomly selected from $\{12, 14, 16\}$.

A.1.3 RAW-to-RGB Rendering

The corrupted RAW signal is then converted back to RGB space through the forward ISP. We apply white balance restoration using the sampled gains, transform the color back to RGB space using the inverse CCM, and finally apply gamma correction using $1/\gamma$.

A.2 Degradation Level Setup for Feature Robustness Analysis

In Sec. 4.5, we evaluate the robustness of intrinsic features under progressively severe illumination degradation using a set of predefined degradation levels within the Dark ISP pipeline. Starting from the original image (Level 0), we construct five synthetic low-light conditions. Each degradation level is defined by a pair of parameters controlling the darkness attenuation factor α and the overall sensor noise scale β:

The first value specifies the multiplicative darkness factor applied to the RAW-like signal after RGB-to-RAW unprocessing, while the second value scales both shot and read noise during sensor noise synthesis. Level 0 corresponds to the original RGB image and is used as the reference feature for comparison.

Table 5: Predefined degradation levels used for intrinsic feature robustness analysis.

Level	Darkness factor α	Noise scale β
0 (Original)	None	1.00
1	0.15	0.88
2	0.07	0.76
3	0.03	0.64
4	0.01	0.52
5	0.005	0.40

To ensure consistent degradation across levels, we reset the random seeds before synthesizing each degraded sample, making the stochastic Dark ISP process deterministic for a given image and level. For each degraded image, we extract the early backbone feature map (`P1`) from both the baseline model and our model, and compare it with the feature map of the original image. Feature robustness is quantified using the Pearson correlation and cosine similarity between the original and degraded features, averaged over the selected samples.

A.3 Hyperparameter Selection

The weights of the decomposition losses were chosen following the design principles commonly adopted in Retinex-based decomposition literature, rather than through an aggressive task-specific search. In particular, our objective was guided by prior works such as Deep Retinex Decomposition (Wei et al., 2018), which formulates decomposition using reconstruction, reflectance consistency, and illumination smoothness constraints, and DAI-Net (Du et al., 2024), which transfers a similar Retinex-style decomposition objective to low-light detection while introducing additional regularization terms for illumination-invariant representation learning.

Following this guideline, we treated the reconstruction terms $\mathcal{L}_{\text{org}}$ and $\mathcal{L}_{\text{dark}}$ as the primary supervision for stabilizing the feature-level Retinex formulation, and used the remaining terms only as auxiliary regularizers. We set $\lambda_{\text{consist}} = 0.05$ to encourage intrinsic agreement between original and dark pairs without over-constraining the feature space, $\lambda_{\text{smooth}} = 0.1$ to impose a sufficiently strong low-frequency prior on the illumination feature, and $\lambda_{\text{cross}} = 0.05$ so that cross-reconstruction improves illumination invariance while remaining secondary to direct reconstruction. In this way, the auxiliary decomposition objective regularizes the detector toward illumination-invariant intrinsic features while maintaining stable optimization of the detection objective during pretraining.

A.4 Auxiliary Encoder and Decoder Architecture

As briefly noted in the main paper, the auxiliary encoder–decoder pair is designed to be lightweight so that it adds only a small computational cost during pretraining, despite introducing no overhead at inference. Here we detail its architecture, which is not fully specified in the main paper due to space constraints.

The auxiliary encoder takes the input image and predicts the illumination factor L and the noise feature N, both produced with the same spatial resolution and channel dimensionality as the intrinsic feature I (i.e., the decomposition target extracted from the backbone's first convolutional layer, `B1`). Matching the channel dimensionality of L and N to I allows them to be combined directly through the intrinsic feature decomposition formulation in Eq. 2 without any additional projection layers. Concretely, the encoder is implemented with only three convolutional layers, and the decoder $\mathcal{D}$ that maps the composed feature back to image space is implemented with three convolutional layers. This is a deliberately lightweight architectural choice: since both modules are discarded after pretraining and never appear in the fine-tuned or deployed detector, we opted for the simplest design capable of producing well-formed L and N features rather than a larger, more expressive network.

To verify that this simple convolutional design is a reasonable choice and not a performance bottleneck, we compare it against two alternative architectures for the auxiliary encoder–decoder pair: (i) an **MLP-based** design, which replaces the convolutional layers with channel-wise MLP blocks operating at the same spatial resolution, and (ii) a **light Transformer-based** design, which uses



Figure 7: Feature visualization based on Retinex feature decomposition.

a shallow attention block followed by an MLP block. All variants are pretrained under the same protocol and evaluated on the LOD dataset after fine-tuning.

Table 6: Ablation study on the architecture of the auxiliary encoder–decoder pair, evaluated on the LOD dataset.

Auxiliary Encoder–Decoder Architecture	**mAP**
MLP-based	63.3
Convolutional	63.7
Light Transformer-based (Attention + MLP)	64.3

As shown in Table 6, mAP varies only within a narrow margin (63.3–64.3) across the three designs. The light Transformer-based variant achieves the highest mAP (64.3), a marginal +0.6 gain over convolutional design (63.7), while the MLP-based variant performs slightly worse (-0.4). Since the light Transformer design introduces additional computation during pretraining for only a marginal improvement, we adopt the three-layer convolutional design for its favorable balance between performance and simplicity. This choice reflects our design intent: the auxiliary modules are meant to provide training-time supervision rather than maximize representational capacity.

B Experimental Setup and Evaluation Protocol

B.1 Overview

We conduct extensive experiments to validate the effectiveness and generalization capability of the proposed intrinsic feature decomposition pretraining framework across various low-light datasets and scenarios. We conduct the evaluation on low-light detection benchmarks (LOD (Hong et al., 2021), NOD (Morawski et al., 2021)) with quantitative comparison and qualitative feature visualizations to provide deeper insights into the learned representations. Also, we implement our approach to large-scale real-world driving datasets (NuImages (Caesar et al., 2020), BDD100K (Yu et al., 2020)) to demonstrate its practical utility in autonomous driving.

B.1.1 Pretraining protocol

For the pretraining stage, we use the COCO dataset (Lin et al., 2014) as the large-scale source domain. Each method is pretrained for 100 epochs using the AdamW optimizer. The learning rate is set to 0.001 as the maximum learning rate, with a warmup ratio of 0.1, and is scheduled using a cosine annealing scheduler. For fair comparison, each domain adaptation method (Cui et al., 2021; Du et al., 2024; Hong et al., 2024), including ours, follows the original pretraining strategy described in its respective paper.

B.1.2 Finetuning protocol

After pretraining, we fine-tune the detectors on each target low-light dataset following a consistent protocol. The number of fine-tuning epochs is determined based on the size of each dataset to ensure stable loss convergence. The detailed configurations and fine-tuning schedules for each dataset are

provided in their respective experimental setup sections. All methods (Guo et al., 2020; Liu et al., 2021b; Ma et al., 2022; Cai et al., 2023; Zhou et al., 2024; Feijoo et al., 2025; Yan et al., 2025; Cui et al., 2021; Du et al., 2024; Hong et al., 2024) are fine-tuned under the same schedule and settings for each dataset to ensure fair comparisons.

B.1.3 Evaluation strategy

We evaluate the pretrained detectors on multiple downstream tasks: object detection on paired low-light datasets (LOD (Hong et al., 2021)), generalization to unpaired low-light scenes (NOD (Morawski et al., 2021)), adaptation to nighttime driving datasets (NuImages (Caesar et al., 2020) and BDD100K (Yu et al., 2020)), and feature visualization analysis. This comprehensive evaluation strategy allows us to assess not only the performance improvements over standard COCO-pretrained models but also the scalability of our approach across different domains, data scales, and model capacities.

B.2 Generalization Across Detectors and Benchmarks

This section details the protocol for the generalization experiments in Sec. 4.4, which evaluate three detectors—YOLOv8-m (Jocher et al., 2023), YOLO26-m (Jocher et al., 2026), and RT-DETR (Zhao et al., 2024b)—on four low-light benchmarks: LOD (Hong et al., 2021), NOD (Morawski et al., 2021), ExDark (Loh & Chan, 2019), and DARK Face (Yang et al., 2020).

Pretraining and fine-tuning protocol All three detectors are pretrained on COCO (Lin et al., 2014) using the same protocol described in Sec. B.1.1 above: 100 epochs with the AdamW optimizer, a maximum learning rate of 0.001, a warmup ratio of 0.1, and a cosine annealing scheduler. The Dark ISP synthesis parameters and the decomposition loss weights ($\lambda_{\text{consist}} = 0.05$, $\lambda_{\text{smooth}} = 0.1$, $\lambda_{\text{cross}} = 0.05$) are likewise kept identical across all three detectors, so that only the detector architecture varies across the comparison. For fine-tuning, each detector–dataset pair uses the same optimizer and learning rate schedule as pretraining. The number of fine-tuning epochs is fixed per dataset and shared across all three detectors: 50 epochs for LOD and 25 epochs for NOD, as described in Secs. 4.1 and 4.2, and 25 epochs for both ExDark and DARK Face.

ExDark ExDark (Loh & Chan, 2019) contains 7,363 low-light images spanning ten lighting conditions, from extremely dark scenes to twilight, annotated with bounding boxes over 12 object categories. We adopt the same train/test split used by MAET (Cui et al., 2021).

DARK Face DARK Face (Yang et al., 2020) targets face detection (a single object category) under real nighttime conditions. Since the officially held-out test set used for the UG^2+ challenge leaderboard is not publicly annotated, we follow the same protocol as HLA-Face (Wang et al., 2021) and use the publicly annotated low-light images, split into 5,500 training and 500 validation images. As faces occupy only a small fraction of the image, we follow the standard input resolution convention of face detection benchmarks (Yang et al., 2016) and fine-tune and evaluate at 1280×1280, rather than the 640×640 resolution used for LOD, NOD, and ExDark.

Common training settings Across all detector–dataset combinations, pretraining uses a batch size of 32 distributed over 8 RTX 5090 GPUs, while fine-tuning uses a batch size of 16 on a single RTX 5090 GPU.

C Additional Experimental Results

C.1 Quantitative Results

C.1.1 Scalability

Fig. 3 illustrates the scalability of our domain-adaptive pretraining method by showing the mAP scores of detectors across different model sizes. Table 7 additionally reports the corresponding quantitative mAP comparisons.

Table 7: Detection performance (mAP) across different model sizes.

Model	Params (M)	Baseline	Ours
YOLOv8-n	3.2	52.8	**55.7** (+2.9)
YOLOv8-s	11.2	55.3	**60.0** (+4.7)
YOLOv8-m	25.9	55.8	**63.7** (+7.9)
YOLOv8-l	43.7	55.9	**64.4** (+8.5)
YOLOv8-x	68.2	55.5	**65.5** (+10.0)

C.1.2 Feature Robustness

Fig. 5 illustrates how feature representations collapse as the degradation level increases, measured by Pearson correlation and cosine similarity between the feature maps of the original and degraded images. Table 8 further provides the corresponding quantitative metrics.

Table 8: Feature robustness under different degradation levels measured by Pearson correlation (ρ) and cosine similarity (s) between feature maps of the original and degraded images. Values are reported as mean $\pm$ standard deviation.

Degradation Level	Pearson Correlation (ρ)		Cosine Similarity (s)	
	Baseline	Ours	Baseline	Ours
0 (Original)	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
1	0.688 ± 0.059	$\mathbf{0.917 \pm 0.024}$	0.767 ± 0.037	$\mathbf{0.924 \pm 0.021}$
2	0.607 ± 0.053	$\mathbf{0.884 \pm 0.033}$	0.709 ± 0.033	$\mathbf{0.894 \pm 0.029}$
3	0.518 ± 0.045	$\mathbf{0.850 \pm 0.043}$	0.648 ± 0.030	$\mathbf{0.863 \pm 0.039}$
4	0.413 ± 0.039	$\mathbf{0.811 \pm 0.056}$	0.580 ± 0.034	$\mathbf{0.827 \pm 0.050}$
5	0.374 ± 0.045	$\mathbf{0.792 \pm 0.063}$	0.556 ± 0.041	$\mathbf{0.810 \pm 0.057}$

C.2 Qualitative Results

C.2.1 Qualitative Analysis of Intrinsic Feature Decomposition

To better understand the behavior of our intrinsic feature decomposition during pretraining, we visualize the intrinsic feature (I), illumination factor (L), and noise feature (N) for paired original and synthetic low-light images, as shown in Fig. 7. We observe that the intrinsic features remain largely consistent between the original and synthetic low-light inputs, preserving structural cues such as object boundaries and textures despite severe illumination attenuation. In contrast, the illumination factors exhibit a clear reduction in overall intensity under low-light conditions, reflecting the applied low-light degradation. Furthermore, the noise feature N^d highlights localized degradation patterns such as amplified sensor noise and detail loss that arise under low-light conditions. Note that no explicit noise component is modeled for original inputs ($N^o \approx 0$), aligning with our intrinsic feature formulation $S^o \approx \mathcal{D}(I^o \odot L^o)$ and $S^d \approx \mathcal{D}((I^d + N^d) \odot L^d)$.

C.2.2 Additional Qualitative Results on LOD

We provide additional qualitative results of object detection on LOD (Hong et al., 2021) and NOD (Morawski et al., 2021) datasets in the supplementary material.