Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Detection Utility, Not Visual Quality: Degradation-Aware Dual-Branch Fusion for Low-Light Object Detection

Hui Zong
Huai'an University
Pengfei Xu
2124096112340hau.edu.cn

Huai'an University
Chenbin Wang
Tongji University
Haoran Liu
Huai'an University
Chunpeng Wang
Qilu University of Technology
Xin Cheng
Huai'an University

Research Article

Keywords: Low-light object detection, degradation-aware representation learning, dual-branch feature fusion, enhancement utility estimation, boundary consistency

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11076264/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Detection Utility, Not Visual Quality: Degradation-Aware Dual-Branch Fusion for Low-Light Object Detection

Hui Zong[1], Pengfei Xu[1], Chenbin Wang[2], Haoran Liu[1], Chunpeng Wang[3*], Xin Cheng[1*]

[1]Faculty of Computer and Software Engineering, Huai'an University, Huai'an, 223001, Jiangsu, China.
[2]School of Computer Science and Technology, Tongji University, Shanghai, 201804, China.
[3]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China.

*Corresponding author(s). E-mail(s): mpeng1122@163.com; chengxin0314@126.com;

Abstract

Enhanced visibility does not necessarily translate into improved object detection under low-light conditions. On ExDark, eight enhanced variants evaluated with the same detector produce mAP@0.5 scores ranging from 76.4% to 77.6%, with five falling below the un-enhanced baseline of 77.0%, indicating that enhancement quality and detection utility can be misaligned. This observation motivates a key question: how can the detector assess the reliability of enhanced representations and adapt their contribution to the detection task? To address this problem, we propose DADF-Det, a degradation-aware dual-branch framework for low-light object detection. A Global Degradation Discriminator (GDD) estimates the detection utility of the enhanced image from the original–enhanced image pair and provides a global reliability prior for feature fusion. Guided by this prior, the Enhanced Texture Branch (ETB) extracts semantic texture representations from the enhanced image, while the Low-Light Edge Branch (LEB) preserves structural and boundary cues from the original image. A Dynamic Discrepancy Calibration Fusion (DDCF) module further calibrates local feature discrepancies according to the estimated global reliability, while a boundary consistency loss

constrains the fused representations to preserve localization-sensitive structures. Experiments on ExDark, DarkFace, and ACDC-Night show that DADF-Det improves mAP@0.5 on ExDark by 1.1 percentage points over the strongest publicly reported method while achieving competitive performance on the remaining benchmarks.

Keywords: Low-light object detection, degradation-aware representation learning, dual-branch feature fusion, enhancement utility estimation, boundary consistency

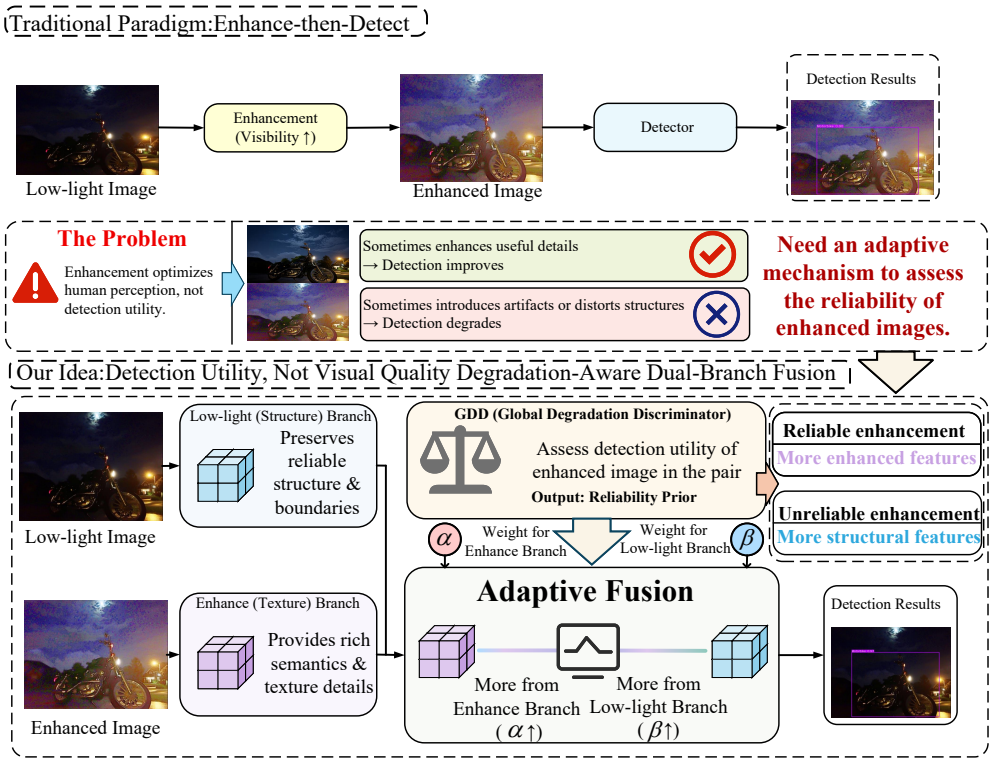


Fig. 1 Motivation and overview of DADF-Det. The proposed framework estimates the detection utility of enhanced information and adaptively balances texture and structural representations for low-light object detection.

1 Introduction

Applications such as intelligent transportation, public security surveillance, and industrial inspection require object detectors to operate reliably under low-light conditions. Low-light imaging suppresses object visibility, weakens texture and boundary cues, and increases the ambiguity between targets and their surrounding backgrounds, making reliable object representation difficult to obtain. In practice, image enhancement has become a common strategy for alleviating these degradations before detection. However, this paradigm implicitly assumes that making low-light images more visually informative will also make them more useful for object detection.

This assumption does not always hold. Representative methods, including Retinex-based models [1–3], curve-estimation approaches [4], and generative enhancement networks [5], improve image visibility and perceptual quality. However, perceptual enhancement methods are primarily optimized for human perception rather than object detection. Enhanced images may contain hallucinated textures, amplified noise, or distorted structures. Improved visibility may coexist with artifacts that

interfere with semantic discrimination and localization. This mismatch between perceptual enhancement and detection objectives limits the effectiveness of conventional enhance-then-detect pipelines.

To reduce dependence on a single enhanced image, existing methods [6] combine features from the original low-light image and its enhanced counterpart through cross-domain feature interaction. Enhanced representations may provide additional appearance cues, while the original input retains structural information unaltered by enhancement. Their relative usefulness, however, varies with degradation severity and scene content. We investigate whether a sample-level branch preference learned from detection-oriented supervision can provide explicit guidance for adaptive fusion.

In our controlled ablation on ExDark (Table 11), fixed equal-weight fusion of the two branches achieves 77.1% mAP_{50}, compared with 78.4% for the structural branch alone. This observation motivates sample-adaptive regulation of the enhanced representation rather than assuming that fixed equal-weight fusion is sufficient. We therefore propose DADF-Det, a degradation-aware dual-branch framework for low-light object detection. A Global Degradation Discriminator (GDD) predicts a sample-level branch-preference coefficient learned under detection-utility supervision and provides a global gating prior for collaborative representation learning. Guided by this prior, the Enhanced Texture Branch (ETB) models semantic representations from the enhanced image, while the Low-Light Edge Branch (LEB) preserves structural cues from the original image. A Dynamic Discrepancy Calibration Fusion (DDCF) module adaptively coordinates the two branches under different degradation conditions. A boundary consistency constraint is further introduced during training to preserve structural continuity and improve localization accuracy.

The main contributions of this work are summarized as follows.

1. We show that the complementarity between original and enhanced representations in dual-branch low-light detectors is not unconditional. Naive fusion of the two branches can underperform the structural branch alone, revealing that enhanced representations are not uniformly beneficial for object detection under low-light conditions.
2. We learn a sample-level gating coefficient under detection-utility supervision to regulate the contributions of the original and enhanced representations. GDD is initialized using degradation statistics and subsequently refined through detection feedback, providing explicit task-supervised guidance for adaptive fusion rather than predicting an absolute detection-utility value.
3. We demonstrate that the proposed utility-guided representation is not tied to a specific enhancement output. It achieves consistent gains across eight enhanced variants on ExDark, with improvements ranging from 2.4 to 3.4 percentage points in mAP_{50} over the corresponding conventional enhance-then-detect pipelines. With task-specific adaptation, DADF-Seg achieves 60.8% mIoU on ACDC-Night, outperforming YOLA by 2.1 percentage points.

2 Related Work

2.1 Generic Object Detection

Object detection has evolved from convolutional neural network (CNN)-based architectures to Transformer-based frameworks. Two-stage detectors, such as Faster R-CNN [7], achieve high detection accuracy through region proposal generation. Single-stage detectors, including YOLO [8] and SSD [9], improve the trade-off between detection accuracy and inference speed. DETR [10] further formulates object detection as a set prediction problem. Subsequent methods, including Deformable DETR [11], DINO [12], and RT-DETR [13], improve training convergence and inference efficiency, promoting the development of end-to-end object detection.

Most existing detectors, however, are designed under normal illumination conditions. Low-light environments introduce noise, texture degradation, and low contrast, which deteriorate low-level visual representations and weaken high-level semantic features. These degradations often result in missed detections, false positives, and inaccurate localization. Improving feature representation under low-light conditions has therefore become a major research topic in object detection.

2.2 Enhancement-Based Low-Light Object Detection

Enhancement-based low-light object detection methods can be broadly categorized into two groups.

The first group follows an enhance-then-detect paradigm. Representative enhancement methods, such as KinD [2], Zero-DCE [4], and Retinexformer [3], improve image visibility before object detection. However, these methods are mainly optimized for perceptual image quality rather than detection-oriented feature representation.

The second group adopts joint enhancement and detection optimization. Methods such as MAET [14] and GDIP [15] reduce the inconsistency between enhancement and detection through end-to-end training. Nevertheless, enhancement errors can still propagate into the detector, and joint optimization increases the coupling between the two tasks.

Although these methods have improved low-light object detection, most of them are still based on single-stream feature learning. The complementary information provided by the original low-light image and its enhanced counterpart is not fully utilized. This limitation has motivated recent studies on dual-branch representation learning.

2.3 Dual-Branch Low-Light Representation

The original low-light image preserves relatively complete structural information, whereas the enhanced image contains richer texture details and semantic cues. These two representations naturally complement each other. Motivated by this observation, recent studies have explored dual-branch architectures for collaborative representation learning.

DENet [16] adopts a detection-guided enhancement framework to enable information interaction between the enhancement and detection branches. PENet, which forms the enhancement component of PE-YOLO [17], models different frequency

components using a multi-branch pyramid architecture to improve detail recovery. FCMA-Det [6] constructs a dual-domain input from the original low-light image and its enhanced counterpart, strengthening feature representation through cross-domain feature interaction.

Dual-branch architectures improve representation learning by exploiting complementary information from different image domains. However, most existing methods employ fixed fusion strategies or static attention mechanisms. The contribution of each branch is therefore independent of the degradation level and the reliability of the enhanced image. This limitation reduces the adaptability of feature fusion under different degradation conditions.

2.4 Degradation-Aware Feature Modeling

Another line of research improves low-light object detection by explicitly modeling image degradation. YOLA [18] learns illumination-invariant representations through degradation-aware feature learning. FRBNet [19] combines image-formation analysis with frequency-domain modeling to suppress illumination interference and improve feature discrimination. DAI-Net [20] models reflectance to bridge day-to-night domain discrepancies, improving cross-domain generalization under complex illumination conditions. These studies demonstrate that degradation-aware representation learning can improve feature robustness under low-light conditions.

Despite these advances, degradation-aware modeling is mainly designed for single-stream feature learning or image enhancement. In most existing methods, degradation information serves only as an auxiliary cue during feature extraction and rarely participates in feature fusion. Consequently, degradation information is not fully exploited to regulate information interaction between different feature streams.

To address this limitation, we propose DADF-Det, which integrates degradation-aware modeling with dual-branch collaborative representation. A Global Degradation Discriminator (GDD) estimates the detection utility of the enhanced image and provides a degradation prior for adaptive branch modulation. Guided by this prior, a Dynamic Discrepancy Calibration Fusion (DDCF) module dynamically coordinates information exchange between ETB and LEB. This design enables degradation-aware collaborative representation learning for low-light object detection.

3 Method

3.1 Problem Definition and Overall Framework

Given a low-light image and its enhanced counterpart, denoted by (I_{low}, I_{enh}), the objective is to learn an object detector that exploits the complementary information of the two inputs while avoiding the adverse effects introduced by low-light enhancement. Unlike conventional enhancement-based detection pipelines that implicitly treat the enhanced representation as uniformly beneficial, we consider its contribution to be degradation-dependent. Specifically, the proposed framework learns a detection model

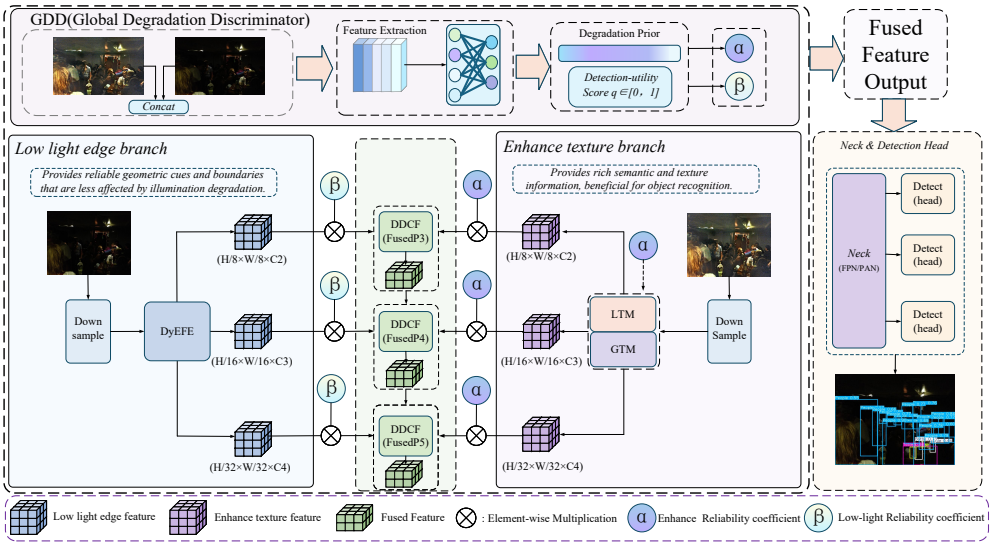


Fig. 2 Overall architecture of the proposed DADF-Det framework.

f together with a global enhancement-utility estimator

$$q = \mathcal{D}(I_{low}, I_{enh}), \qquad q \in [0, 1], \tag{1}$$

where q is a relative branch-preference coefficient learned under detection-utility supervision, rather than an absolute measure of detection performance. A larger q favors the enhanced representation, whereas a smaller q favors the original low-light representation. The term utility score refers to this utility-guided gating coefficient throughout the paper. It is converted into complementary branch modulation weights to regulate feature learning and fusion.

The problem formulation is based on three practical assumptions. **(A1) Enhanced-input availability.** The enhanced image I_{enh} is assumed to be available during both training and inference and is generated by an external low-light enhancement model. Therefore, the proposed detector is designed as a degradation-aware detection framework built upon an enhancement-assisted input pair rather than as a standalone enhancement model. **(A2) Global-to-local reliability decomposition.** The scalar q provides a sample-level prior expressing the relative preference between the two representations, while their reliability may still vary spatially within an image. Accordingly, GDD supplies global branch guidance, whereas DDCF uses local feature discrepancies to spatially modulate the fused responses. **(A3) Structural-reference reliability.** Although the original low-light image may contain noise and weak responses, its structural representation does not suffer from enhancement-induced distortions. We therefore assume that the LEB representation provides a relatively reliable structural reference for boundary preservation, rather than assuming that it is free of low-light degradation.

Based on this formulation, we propose DADF-Det, a degradation-aware dual-branch framework for low-light object detection, as illustrated in Fig. 2. The framework consists of four main components: the Low-Light Edge Branch (LEB), the Enhanced Texture Branch (ETB), the Dynamic Discrepancy Calibration Fusion (DDCF) module, and the Boundary Consistency Loss. Given (I_{low}, I_{enh}), LEB extracts multi-scale structural representations $\{F^3_{edge}, F^4_{edge}, F^5_{edge}\}$ directly from the original low-light image, while ETB produces the corresponding texture representations $\{F^3_{tex}, F^4_{tex}, F^5_{tex}\}$ from the enhanced image.

To estimate the global utility of enhancement, the Global Degradation Discriminator (GDD) jointly analyzes I_{low} and I_{enh} and predicts $q \in [0, 1]$. The resulting degradation prior directly determines the complementary global modulation weights $\alpha_g = q$ and $\beta_g = 1 - q$ of ETB and LEB, respectively, as defined in Section 3.3.1. This prior is shared by the subsequent texture modeling and feature fusion processes, allowing the contribution of the two branches to vary according to the overall degradation condition of each sample.

The two branches are then processed according to their respective roles. ETB uses the global degradation prior to regulate texture representation learning and reduce unreliable enhancement-induced responses, whereas LEB focuses on preserving structural and boundary cues from the original low-light input. DDCF subsequently performs cross-branch fusion by combining the global reliability prior with local response discrepancies, thereby extending the global utility estimate with position-aware feature calibration. Finally, the Boundary Consistency Loss constrains the gradient-domain consistency between the fused representation and the LEB-derived structural reference. Through this global-to-local formulation, DADF-Det jointly models enhancement reliability, cross-branch complementarity, and structural preservation for low-light object detection.

3.2 Low-Light Edge Branch (LEB)

The original low-light image preserves the original imaging cues without being altered by the enhancement process. It therefore retains potentially useful object boundaries and geometric information [21, 22]. However, low-light conditions reduce the signal-to-noise ratio, causing genuine edge responses to be easily corrupted by random noise. Accordingly, the low-light edge branch is designed not to recover complete semantic representations, but to extract relatively reliable structural cues from the degraded input for complementary representation learning.

Considering the scale variations of object boundaries under low-light conditions, DyEFE first performs multi-scale contextual modeling to establish spatial correlations. Structural responses extracted with different receptive fields provide complementary contextual information for subsequent boundary analysis [23, 24]. Given an input feature map F_{in}, projection operators with different receptive fields are employed to extract structural responses, which are aggregated along the channel dimension as

$$F_{ctx} = \mathrm{Concat}(P_1(F_{in}), P_2(F_{in}), \ldots, P_n(F_{in})), \tag{2}$$

where $P_i(\cdot)$ denotes the feature projection operator with the ith receptive field, and $\mathrm{Concat}(\cdot)$ represents channel-wise concatenation. The multi-scale responses provide

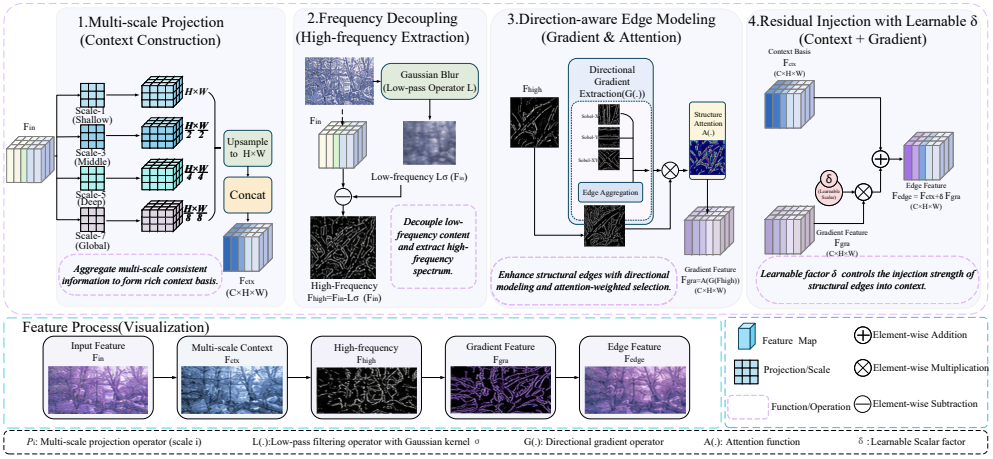


Fig. 3 Architecture of the proposed Dynamic Edge Feature Enhancement (DyEFE) module.

contextual information at different spatial ranges, which supports the representation of weak and scale-varying structural responses.

Spatial contextual modeling alone does not explicitly distinguish genuine boundaries from random noise, as both may contain high-frequency components [25]. DyEFE therefore further performs frequency decoupling to reduce low-frequency illumination variations while retaining high-frequency responses for subsequent structural analysis:

$$F_{high} = F_{in} - L_{\sigma}(F_{in}), \tag{3}$$

where $L_{\sigma}(\cdot)$ denotes the low-pass filtering operator. This operation separates the low-frequency component associated with illumination and smooth background variations from the remaining high-frequency responses, thereby providing a frequency-domain representation for subsequent boundary modeling.

Nevertheless, F_{high} may contain both object boundaries and noise responses. Frequency decomposition alone is therefore insufficient to characterize their structural differences. DyEFE subsequently applies the Sobel gradient operator $G(\cdot)$ [26] to introduce directional gradient information. Continuous object boundaries generally exhibit spatially coherent gradient patterns, whereas noise may produce irregular local responses. Based on these gradient features, a lightweight adaptive selection function $A(\cdot)$ is introduced to further refine the structural responses [27, 28]:

$$F_{gra} = A(G(F_{high}))\,, \tag{4}$$

where $G(\cdot)$ denotes the Sobel gradient operator and $A(\cdot)$ denotes the lightweight adaptive feature selection function. The gradient representation provides explicit directional structural cues, while the adaptive selection function adjusts the contribution of different gradient responses according to their local characteristics.

To integrate the structural information with the contextual representation, a learnable coefficient δ is introduced to regulate the contribution of the gradient features.

The refined gradient representation is incorporated into the multi-scale contextual features through a residual formulation inspired by residual learning [29]:

$$F_{edge} = F_{ctx} + \delta F_{gra}, \tag{5}$$

where δ is a learnable parameter that controls the contribution of the gradient-based structural information. This residual formulation allows the contextual representation to serve as the basic feature representation while introducing gradient cues as an additional structural prior.

The above procedure is performed independently at multiple feature scales, producing the multi-scale structural representations

$$\left\{F^3_{edge},\, F^4_{edge},\, F^5_{edge}\right\}. \tag{6}$$

Since this branch operates directly on the original low-light input, its representations retain structural cues that are not directly modified by the enhancement process. These multi-scale features are subsequently provided to the degradation-aware dynamic fusion module and the boundary consistency constraint as structural references.

3.3 Enhanced Texture Branch (ETB)

LEB extracts stable yet semantically limited structural priors directly from the original image, whereas the enhanced image provides richer texture details and semantic information. Nevertheless, the enhancement process inevitably introduces unreliable responses, including enhancement artifacts, hallucinated textures, and local structural distortions [5, 30]. The reliability of these responses varies with degradation severity and scene content. Texture representation should therefore adaptively balance semantic preservation and artifact suppression instead of relying on a fixed feature enhancement strategy.

To address this issue, we develop the Enhanced Texture Branch (ETB). The proposed branch utilizes the degradation prior predicted by the Global Degradation Discriminator (GDD) to regulate both local and global texture modeling. The degradation prior is shared throughout the texture modeling pipeline. It provides a unified modulation signal for the Local Texture Module (LTM), the Global Texture Module (GTM), and the Dynamic Discrepancy Calibration Fusion (DDCF) module.

The LTM selects locally discriminative texture patterns during feature downsampling. The GTM aggregates multi-scale contextual information under large receptive fields while suppressing anomalous texture responses. It further refines the local representations generated by the LTM into globally consistent texture representations. The complementary roles of the two modules enable degradation-aware texture representation learning under different degradation conditions.

ETB generates multi-scale texture representations $\{F^3_{tex}, F^4_{tex}, F^5_{tex}\}$ at the P3, P4, and P5 feature levels. These representations are fused with the structural representations extracted by LEB through the Dynamic Discrepancy Calibration Fusion (DDCF) module described in Section 3.4.

3.3.1 Global Degradation Discriminator (GDD)

Although low-light enhancement restores richer texture details, its contribution to object detection is not solely determined by visual quality. Under complex low-light conditions, enhancement-induced errors often have a limited impact on perceptual quality, but they can impair semantic discrimination and localization accuracy [30]. A fixed contribution assigned to enhanced features therefore cannot accurately characterize their utility for the downstream detection task.

To guide branch preference using detection utility, we introduce a Global Degradation Discriminator (GDD). The proposed module jointly analyzes the original low-light image and its enhanced counterpart to predict a relative gating coefficient. Unlike conventional image quality assessment methods that measure perceptual similarity, such as FSIM [31], GDD does not predict perceptual image quality. Instead, detection-utility supervision teaches it how to regulate the enhanced representation relative to the original representation for the current task. Its output serves as a global degradation prior, not a calibrated absolute utility value.

Given the original low-light image I_{low} and its enhanced counterpart I_{enh}, the two images are concatenated along the channel dimension to form a six-channel input. The concatenated input is then fed into the GDD to predict a degradation-aware coefficient:

$$q = D_\theta(I_{low}, I_{enh}), \qquad 0 \leq q \leq 1,$$

where $D_\theta(\cdot)$ denotes a lightweight convolutional network parameterized by θ.

Direct supervision for detection utility is not readily available. Training the GDD solely with the detection loss may also lead to unstable optimization, since the detector is continuously updated during training. To address this issue, we adopt a two-stage learning strategy. This strategy allows the GDD to progressively establish a mapping from degradation statistics to task-related detection utility.

Before training the GDD, the discrepancy modulation path is task-aware initialized by exposing DDCF to a narrow range of fusion gates around the neutral state $q = 0.5$. Specifically, q is randomly sampled from $[0.4, 0.6]$ during this initialization period, allowing the discrepancy encoder to receive task gradients while preserving a near-balanced fusion regime. After convergence, the detector and discrepancy encoder are fixed for the subsequent two-stage GDD training. This gate-perturbed initialization prepares the utility evaluator and is distinct from the fixed-$q = 0.5$ No-GDD ablation setting.

Stage I: Degradation Attribute Learning

This stage constructs weak supervisory signals from degradation statistics to provide a stable initialization for the GDD. Multiple no-reference degradation attributes, including brightness, blur, noise, color distortion, and compression artifacts, are computed for each enhanced image and normalized as

$$d_k \in [0, 1]. \tag{7}$$

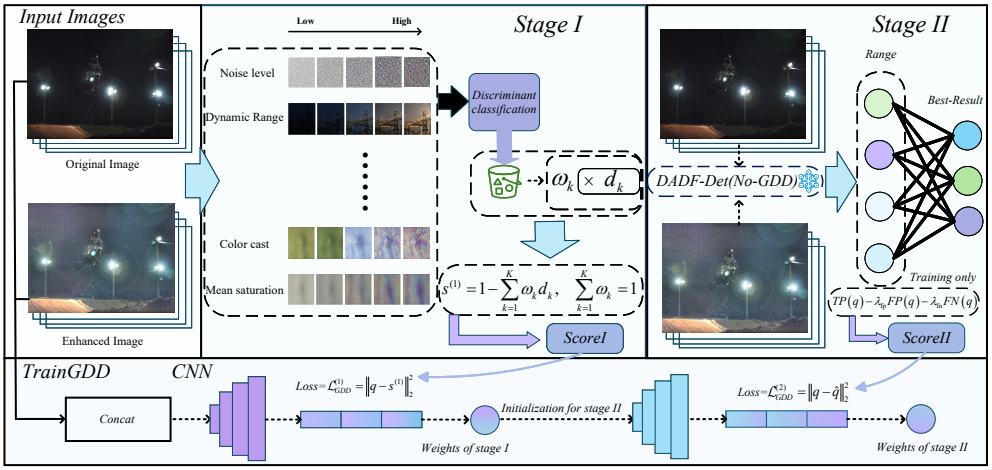


Fig. 4 Two-stage training strategy of the Global Degradation Discriminator (GDD).

Here, a larger value of d_k indicates a stronger degree of the corresponding degradation. To avoid introducing additional manually tuned priors, all degradation attributes are assigned uniform weights. The overall degradation score is therefore obtained by

$$s^{(1)} = 1 - \sum_{k=1}^{K} \omega_k d_k, \qquad \omega_k = \frac{1}{K}, \qquad \sum_{k=1}^{K} \omega_k = 1, \tag{8}$$

where K denotes the number of degradation attributes and ω_k is the uniform weight assigned to the k-th attribute. The degradation score serves as a weak supervisory target for initializing the GDD:

$$\mathcal{L}_{GDD}^{(1)} = \left\| q - s^{(1)} \right\|_2^2. \tag{9}$$

This stage does not aim to model task-specific detection performance. Instead, it provides a coarse degradation-aware initialization for the GDD, establishing an initial correspondence between global degradation characteristics and the gating coefficient. By using uniform weights without manually tuned attribute importance, Stage I introduces only a lightweight and task-independent prior, thereby avoiding an overinterpretation of perceptual degradation as detection utility. The task relevance of the gating coefficient is subsequently learned through the detection-oriented optimization in Stage II.

Stage II: Detection Utility Refinement

Degradation statistics describe perceptual image degradation, but they do not necessarily reflect whether an enhanced image improves object detection. Moreover, direct supervision from detection utility is available through ground-truth-based evaluation, but obtaining such supervision requires multiple detector forward passes for each training sample and is therefore computationally expensive. Therefore, rather than directly

optimizing the GDD from scratch with detection feedback, the second stage further refines the GDD initialized in Stage I using task-level detection utility.

For each training image pair, a predefined set of candidate gating coefficients is evaluated:

$$\mathcal{Q} = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}.$$

Each evaluated candidate coefficient $\tilde{q} \in \mathcal{Q}$ is applied to the gate-initialized DADF-Det model without the GDD module during forward passes on the training split. The detector and discrepancy encoder parameters are kept fixed throughout this process, and only the candidate gating coefficient is varied for utility evaluation. The resulting predictions are compared with the corresponding ground-truth annotations using the detection utility function

$$S_{\text{det}}(\tilde{q}) = TP(\tilde{q}) - \lambda_{fp} FP(\tilde{q}) - \lambda_{fn} FN(\tilde{q}), \tag{10}$$

where $TP(\tilde{q})$, $FP(\tilde{q})$, and $FN(\tilde{q})$ denote the numbers of true positives, false positives, and false negatives, respectively. We set $\lambda_{fp} = 0.5$ and $\lambda_{fn} = 1.0$, thereby assigning a higher penalty to missed detections. A prediction is considered a true positive when its IoU with a ground-truth object of the same class is no less than 0.5, with a confidence threshold of 0.25 and an NMS IoU threshold of 0.45.

Let $\mathcal{Q}_{eval} \subseteq \mathcal{Q}$ denote the candidates actually evaluated for the current training sample. The coefficient with the highest utility among these candidates is used as the utility-guided gating target:

$$\hat{q} = \arg\max_{\tilde{q} \in \mathcal{Q}_{eval}} S_{\text{det}}(\tilde{q}), \tag{11}$$

which is then used to fine-tune the GDD through

$$\mathcal{L}_{GDD}^{(2)} = \|q - \hat{q}\|_2^2 . \tag{12}$$

The gating target is computed only on the training split and is never used at test time. It is the selected control coefficient, not the utility value $S_{\text{det}}(\hat{q})$ itself. This stage transfers task-level detection feedback to GDD through a fixed, task-initialized evaluator, avoiding the use of an untrained GDD to generate its own supervision. In practice, a binary interval-halving heuristic progressively narrows the candidate gate range and returns the best evaluated coefficient as an approximate target, requiring approximately 3–5 forward passes per training sample on average and at most $|\mathcal{Q}| = 11$ passes.

After the two-stage training, GDD is frozen and used to provide sample-dependent gating coefficients during subsequent detector training and inference.

Deployment: Global Degradation Prior

At inference, the frozen GDD predicts the utility-guided coefficient q, which is directly converted into complementary global gating weights for the two branches:

$$\alpha_g = q, \tag{13}$$

$$\beta_g = 1 - q. \tag{14}$$

Here, α_g and β_g denote the global modulation weights for ETB and LEB, respectively. Since $q \in [0, 1]$, the two coefficients satisfy

$$\alpha_g + \beta_g = 1, \qquad \alpha_g, \beta_g \in [0, 1].$$

A larger q assigns greater relative weight to the enhanced texture branch, whereas a smaller q shifts the weighting toward the low-light edge branch. Thus, the GDD output expresses a task-supervised branch preference for texture modeling and cross-branch fusion; it does not by itself indicate the absolute detection performance of an image.

Rather than directly participating in feature fusion, the GDD serves as a shared degradation prior that guides the LTM, GTM, and DDCF. Unlike conventional enhancement-then-detection paradigms that rely primarily on the enhanced image for downstream detection, the proposed GDD jointly considers the original low-light image and its enhanced counterpart to characterize degradation and branch reliability. This design further introduces detection utility into the representation learning process, enabling the framework to adjust the contributions of the two branches under varying degradation conditions, following the general principle of degradation-aware representation learning [18–20].

3.3.2 Local Texture Module (LTM)

Low-light enhancement can reveal texture details, but the resulting representations may also contain enhancement-induced errors [30]. These unreliable responses coexist with informative textures and may propagate into deeper feature representations through conventional downsampling, degrading subsequent semantic representation learning. Accordingly, LTM aims to retain locally discriminative textures while limiting the propagation of unreliable responses during feature encoding.

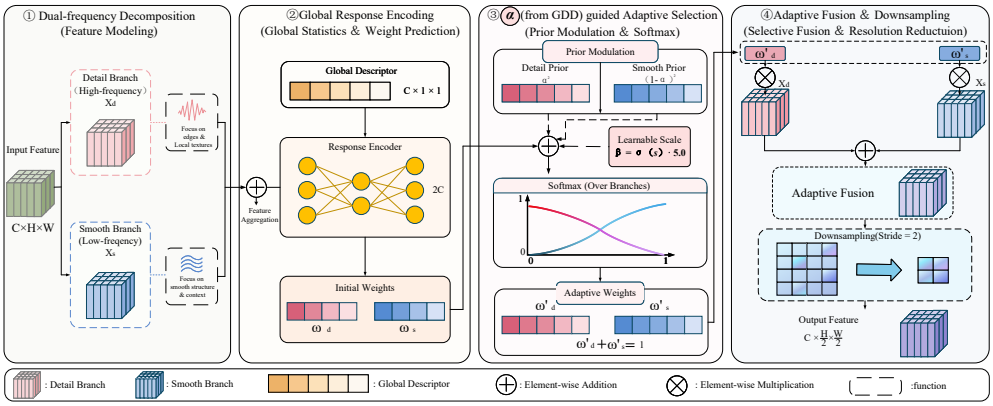


Fig. 5 Architecture of the proposed Local Texture Module (LTM).

To achieve this objective, the Local Texture Module (LTM) uses two complementary branches to model local detail and broader spatial context, respectively [23, 24]:

$$X_d = \phi_d(X), \quad X_s = \phi_s(X), \tag{15}$$

where the detail branch $\phi_d(\cdot)$ models local textures and edge details, while the smooth branch $\phi_s(\cdot)$ aggregates contextual information with a larger receptive field. Both outputs have shape $C \times H \times W$ for subsequent aggregation and weighted fusion.

Global average pooling summarizes $X_d + X_s$ into channel-wise statistics [27], which $\psi(\cdot)$ maps to initial logits $[o_d, o_s]$, arranged as a $2 \times C$ array. The GDD prior $\alpha_g = q$ then regulates branch selection through additive logit offsets:

$$[\omega_d, \omega_s] = \mathrm{Softmax}_{\mathrm{branch}}(\psi(\mathrm{GAP}(X_d + X_s)) + \Delta(\alpha_g)) \,. \tag{16}$$

The detail and smooth priors are defined as

$$b_d = \alpha_g^2, \qquad b_s = (1 - \alpha_g)^2.$$

With the bounded learnable scale $\gamma = 5\sigma(\beta)$, the offsets are stacked along the branch dimension:

$$\Delta(\alpha_g) = \gamma \begin{bmatrix} (b_d - \frac{1}{2})\mathbf{1}_C \\ (b_s - \frac{1}{2})\mathbf{1}_C \end{bmatrix}.$$

Here, β is an LTM-specific learnable scalar, distinct from the global weight β_g; σ is the sigmoid function, and $\mathbf{1}_C$ is an all-ones row vector. Softmax is applied across the two branches independently for each channel, yielding $\omega_{d,c} + \omega_{s,c} = 1$. For fixed initial logits, the prior shifts the detail-minus-smooth logit difference by $\gamma(2\alpha_g - 1)$. Thus, relative to the unmodulated weights, $\alpha_g > 0.5$ favors the detail branch and $\alpha_g < 0.5$ favors the smooth branch; at $\alpha_g = 0.5$, the common offset cancels and the original weights are preserved.

The channel-wise weights are broadcast over the spatial dimensions to fuse the two representations before downsampling:

$$F_{LTM} = \mathrm{Down}(\omega_d \odot X_d + \omega_s \odot X_s) \,, \tag{17}$$

where $\mathrm{Down}(\cdot)$ denotes the downsampling operation.

In this way, LTM combines image-dependent channel statistics with the task-supervised global prior to regulate the relative contributions of detail and contextual responses before downsampling.

3.3.3 Global Texture Module (GTM)

The Local Texture Module (LTM) performs texture selection at the local scale. However, enhancement-induced errors may still affect deep feature distributions through large-scale illumination shifts, texture discontinuities, and accumulated artifacts. Local receptive fields alone are therefore insufficient to establish globally consistent texture representations, motivating long-range contextual modeling [32, 33]. To address this limitation, we introduce a Global Texture Module (GTM). Guided by the degradation

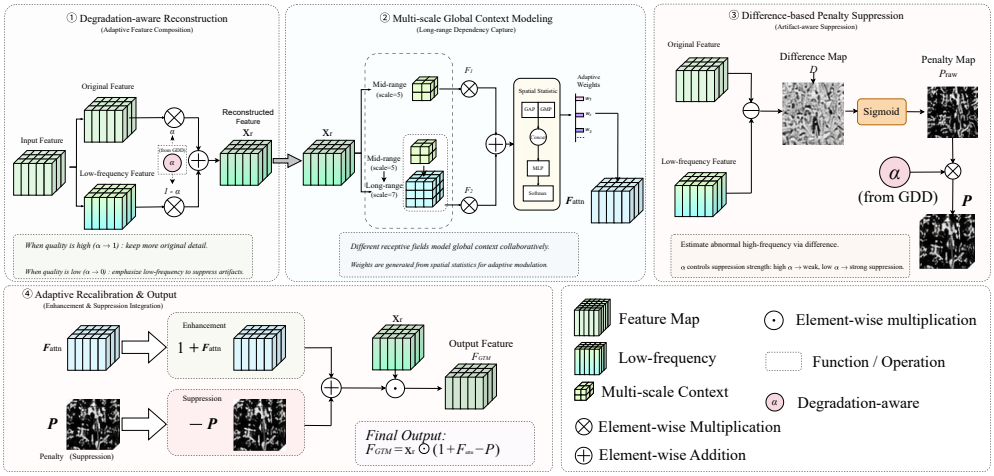


Fig. 6 Architecture of the proposed Global Texture Module (GTM).

prior provided by the Global Degradation Discriminator (GDD), GTM jointly performs degradation-aware feature reconstruction, multi-scale contextual aggregation, and artifact suppression within a unified feature modulation framework.

Given an input feature map X, degradation-aware feature reconstruction is performed using the degradation prior α_g defined in Section 3.3.1. A low-frequency structural representation X_{low} is obtained through local average pooling, and the reconstructed feature is formulated as

$$X_r = \alpha_g X + (1 - \alpha_g) X_{low}. \tag{18}$$

This adaptive interpolation reduces the propagation of enhancement-induced errors into deep semantic representations and provides a more reliable input for subsequent global modeling.

Based on the reconstructed feature, contextual branches with different receptive fields are employed to capture complementary long-range spatial dependencies. Channel-wise global average pooling and max pooling are used to summarize spatial statistics and generate adaptive weights for aggregating multi-scale contextual information. The resulting global texture attention is formulated as

$$F_{attn} = \rho\Big(\hat{F}_1 \odot W_1 + \hat{F}_2 \odot W_2\Big), \tag{19}$$

where $\hat{F}_1$ and $\hat{F}_2$ denote channel-compressed contextual features extracted with different receptive fields, and W_1 and W_2 represent the corresponding adaptive weights derived from spatial statistics. Here, $\rho(\cdot)$ denotes a sigmoid activation function, such that $F_{attn} \in [0, 1]$. The complementary receptive fields promote semantic aggregation under global spatial consistency, thereby improving feature representation in complex low-light scenes.

In addition to contextual aggregation, GTM explicitly suppresses enhancement-induced artifacts. Genuine structures are generally supported by coherent low-frequency components, whereas enhancement artifacts are more likely to appear as abnormal high-frequency responses [25]. The discrepancy between the original feature and its low-frequency component is therefore used to estimate potentially unreliable responses. The degradation prior further regulates these responses through an adaptive penalty term:

$$P = (1 - \alpha_g)\sigma(\xi(|X - X_{low}|)) . \tag{20}$$

Here, $\xi(\cdot)$ denotes a lightweight mapping function and $\sigma(\cdot)$ is the sigmoid function. Since $\alpha_g \in [0, 1]$ and $\sigma(\cdot) \in (0, 1)$, the penalty remains bounded as $P \in [0, 1)$.

The global texture attention and adaptive penalty are jointly applied to the reconstructed feature through residual recalibration [29]:

$$F_{GTM} = X_r \odot (1 + F_{attn} - P) . \tag{21}$$

Since $F_{attn} \in [0, 1]$ and $P \in [0, 1)$, the modulation factor $1 + F_{attn} - P$ remains within $(0, 2]$. Thus, the recalibration operation can adaptively enhance or suppress feature responses without reversing their signs. Specifically, the attention term increases the responses associated with globally informative texture patterns, while the penalty term attenuates degradation-sensitive responses. This design enables GTM to improve texture discrimination while maintaining stable feature modulation under different degradation conditions.

GTM produces multi-scale texture representations $\{F^3_{tex}, F^4_{tex}, F^5_{tex}\}$ at the P3, P4, and P5 feature levels, respectively, which are subsequently provided to DDCF for fusion with the corresponding LEB representations.

3.4 Dynamic Discrepancy Calibration Fusion (DDCF)

ETB and LEB provide complementary semantic and structural information. However, the reliability of the two representations varies across degradation conditions and spatial locations. A fixed fusion strategy cannot simultaneously account for sample-level reliability and local response discrepancies, and adaptive attention-based fusion is generally more suitable for features with inconsistent semantics or scales [34]. To address this issue, we introduce a Dynamic Discrepancy Calibration Fusion (DDCF) module, which performs degradation-aware collaborative fusion under the guidance of the global degradation prior.

At each pyramid level $i \in \{3, 4, 5\}$, let $F^i_{enh} = F^i_{tex}$ and $F^i_{low} = F^i_{edge}$ denote the ETB and LEB features, respectively. The structural feature is first projected into the feature space of ETB, producing the aligned representation

$$F^{a,i}_{low} = P^f_i\left(F^i_{low}\right), \tag{22}$$

where $P^f_i(\cdot)$ denotes a learnable projection function. The global degradation prior predicted by GDD is subsequently used to determine the fusion weights of the two

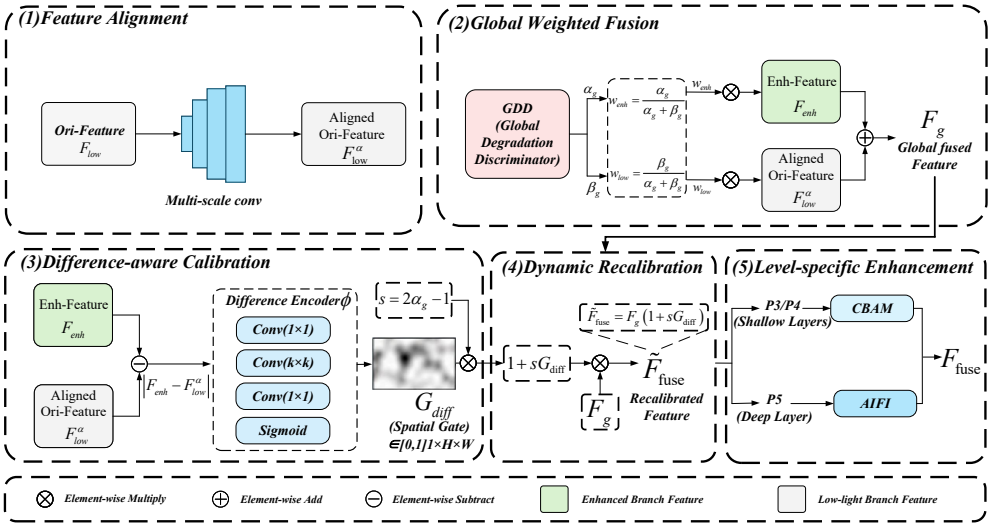


Fig. 7 Architecture of the proposed Dynamic Discrepancy Calibration Fusion (DDCF) module.

branches. The global fused representation is formulated as

$$F_g^i = w_{enh} F_{enh}^i + w_{low} F_{low}^{a,i}, \tag{23}$$

where

$$w_{enh} = \frac{\alpha_g}{\alpha_g + \beta_g}, \qquad w_{low} = \frac{\beta_g}{\alpha_g + \beta_g}. \tag{24}$$

Since $\alpha_g + \beta_g = 1$ in the proposed formulation, the fusion weights satisfy

$$w_{enh} = \alpha_g = q, \qquad w_{low} = \beta_g = 1 - q.$$

Here, α_g and β_g are complementary global branch modulation weights for ETB and LEB, respectively, derived from the same detection-utility score q as defined in (13) and (14). In general, the ratio form normalizes the relative branch contributions before fusion, making them invariant to a common positive rescaling of the modulation coefficients. In the present formulation, the complementary weights already sum to one, so this normalization leaves them unchanged. These weights adaptively adjust the overall contribution of each branch according to the detection utility of the enhanced image. As a result, the fused representation remains consistent with the degradation condition of the current sample.

Global weighting alone cannot characterize local variations in cross-branch responses. DDCF therefore uses the absolute feature difference as input to a learnable spatial gate for modulating the globally fused representation:

$$G_{diff}^i = \sigma\left(\phi_i\left(\left|F_{enh}^i - F_{low}^{a,i}\right|\right)\right), \tag{25}$$

where $\phi_i(\cdot)$ denotes a lightweight discrepancy encoder. The resulting G^i_{diff} is a learnable spatial gate derived from feature differences, rather than a direct measure constrained to increase monotonically with discrepancy magnitude.

Feature differences alone do not determine whether a fused response should be amplified or suppressed. The spatial gate is therefore combined with the global branch-preference coefficient α_g to set the direction and magnitude of response modulation. Specifically, we define a signed modulation factor as

$$s = 2\alpha_g - 1 = 2q - 1, \qquad s \in [-1, 1], \tag{26}$$

and apply it to the spatial discrepancy guidance through residual modulation [29]:

$$\widetilde{F}^i_{fuse} = F^i_g \left(1 + sG^i_{diff}\right). \tag{27}$$

Since $G^i_{diff} \in [0, 1]$ and $s \in [-1, 1]$, the modulation factor $1+sG^i_{diff}$ lies in $[0, 2]$. Thus, the spatial gate can amplify or suppress the globally fused responses without reversing their signs. When $\alpha_g > 0.5$, $s > 0$ and a larger gate value produces stronger amplification; when $\alpha_g < 0.5$, $s < 0$ and a larger gate value produces stronger attenuation. This operation modulates the fused response as a whole rather than selectively suppressing only the enhanced-branch contribution, and it does not recompute a pixel-wise ratio between the two branch weights.

The difference-derived gate acts after global weighted fusion as a spatial modulation of the fused responses. Applying this operation at P3–P5 produces the multi-scale calibrated features $\{\widetilde{F}^3_{fuse}, \widetilde{F}^4_{fuse}, \widetilde{F}^5_{fuse}\}$. To further accommodate different semantic levels, CBAM [28] refines the shallow P3 and P4 features, while the AIFI module from RT-DETR [13] models global dependencies in the deep P5 feature. The refined outputs are denoted by $\{F^3_{fuse}, F^4_{fuse}, F^5_{fuse}\}$ in the subsequent boundary constraint.

Overall, DDCF combines sample-level branch weighting, difference-derived spatial modulation, and level-specific attention. GDD determines the global branch preference, while the learnable spatial gate adjusts the strength of the already fused responses according to the feature differences.

3.5 Boundary Consistency Loss

To mitigate the attenuation of boundary responses during dual-branch feature fusion, we introduce a Boundary Consistency Loss in addition to the standard detection objective. Object boundaries are associated with structural discontinuities in the gradient domain [21], while enhancement-induced errors may introduce abnormal local gradients [22]. Therefore, constraining the fused representation to retain the gradient characteristics of the original low-light representation provides a way to preserve structural cues that may be weakened during enhancement and feature fusion.

Since boundary responses are spatially sparse, applying the gradient constraint uniformly over the entire feature map may dilute the supervision in non-boundary regions. To emphasize structurally informative locations, we derive a spatial edge prior from the LEB representation at each pyramid level. Specifically, the channel-averaged

gradient magnitude is first computed as

$$g_i(h,w) = \frac{1}{C_i^e} \sum_{c=1}^{C_i^e} \left| \mathrm{Sobel}\left(F_{\mathrm{edge}}^i\right) \right|_{c,h,w}, \quad i \in \{3,4,5\},$$

where C_i^e denotes the number of channels in F_{edge}^i, and $|\,\mathrm{Sobel}(\cdot)|$ denotes the per-channel gradient magnitude obtained using depth-wise Sobel filtering. The spatial gradient response is then max-normalized and detached from the computational graph to obtain the soft edge prior:

$$M_i = \mathrm{sg}\left[\frac{g_i}{\max_{h,w} g_i + \epsilon} \right], \tag{28}$$

where ϵ is a small constant for numerical stability and $\mathrm{sg}[\cdot]$ denotes the stop-gradient operation. The maximum is computed over the spatial dimensions separately for each image and pyramid level. Thus, M_i is a single-channel spatial prior with values in $[0,1]$, assigning relatively larger weights to locations with stronger gradient responses. The detached formulation treats the spatial weighting as fixed during each optimization step, blocking direct loss gradients through the weighting path.

To establish the structural reference, the corresponding LEB feature is projected into the representation space of the fused feature:

$$F_i^{\mathrm{ref}} = \mathrm{sg}\left[P_i^f\left(F_{\mathrm{edge}}^i\right) \right], \tag{29}$$

where $P_i^f(\cdot)$ is the same learnable projection used in DDCF, with parameters shared with the detection path. The stop-gradient operation is applied only to the output on the reference path when computing the boundary constraint; it does not freeze the shared projection or LEB. The reference therefore serves as a fixed target within each optimization step while remaining dependent on the current structural representation learned by the network.

The gradient discrepancy between the fused representation and the structural reference is defined as

$$D_i = G\left(F_i^{\mathrm{fuse}}\right) - G\left(F_i^{\mathrm{ref}}\right), \tag{30}$$

Here, $G(\cdot)$ denotes the per-channel Sobel gradient magnitude, consistent with the magnitude used to construct the spatial prior. To avoid imposing a channel-wise correspondence between the original LEB representation and the projected feature representation, the discrepancy is first reduced along the feature-channel dimension

to form two spatial residual maps:

$$R_{1,i}(h,w) = \frac{1}{C_i}\sum_{c=1}^{C_i} |D_i(c,h,w)|,$$
$$R_{2,i}(h,w) = \frac{1}{C_i}\sum_{c=1}^{C_i} D_i(c,h,w)^2,$$

where C_i denotes the number of channels in D_i. The spatial edge prior is then applied to these residual maps:

$$\begin{aligned} \mathcal{L}_{feat} &= \frac{1}{3}\sum_{i=3}^{5} \mathrm{mean}_{h,w}[M_i \odot R_{1,i}]\,, \\ \mathcal{L}_{match} &= \frac{1}{3}\sum_{i=3}^{5} \mathrm{mean}_{h,w}[M_i \odot R_{2,i}]\,. \end{aligned} \tag{31}$$

At each pyramid level, the spatial prior and residual maps share the same spatial resolution. The losses are computed per image and then averaged over the mini-batch. Here, the channel dimension is first averaged to obtain spatial residual maps, after which M_i performs spatial weighting. Consequently, the edge prior does not impose a correspondence between the channels of the original LEB feature and those of the projected reference. Moreover, the prior enters both penalties linearly, while the squared term is applied only to the gradient discrepancy. The L_1 penalty maintains sensitivity to local gradient discrepancies, whereas the squared residual penalty assigns greater cost to pronounced structural deviations.

The Boundary Consistency Loss is consequently defined as

$$\mathcal{L}_{bound} = \mathcal{L}_{feat} + \mathcal{L}_{match}. \tag{32}$$

The gradient-derived spatial prior is therefore an integral part of the boundary constraint rather than an additional optimization objective. Detaching the prior blocks direct loss gradients through the weighting path within each optimization step. The prior is nevertheless recomputed from the current LEB representation as training proceeds.

After the two-stage GDD training, the detector undergoes further task training with GDD frozen. The feature extraction, fusion, neck, and prediction head are optimized, while GDD supplies sample-dependent gating coefficients without parameter updates. The GDD regression losses are not used in this stage. The final detector training objective is formulated as

$$\mathcal{L} = \mathcal{L}_{det} + \lambda_{grad}\mathcal{L}_{bound}, \tag{33}$$

where λ_{grad} controls the contribution of the boundary constraint to the overall optimization. Since the fused representation is optimized against a detached structural

reference, an excessively large λ_{grad} may impose an overly strong structural constraint and interfere with task-oriented feature learning. The effect of λ_{grad} is therefore investigated in Section 4 to determine an appropriate regularization strength.

Overall, the proposed Boundary Consistency Loss constrains the fused representation in the feature-gradient domain while assigning greater spatial emphasis to gradient-salient regions derived from LEB. The detached spatial prior and projected structural reference block direct back-propagation through the weighting and reference paths during each optimization step, while the detector continues to learn through the detection objective and the fused-feature path of the boundary constraint.

4 Experimental Analysis

4.1 Dataset and Experimental Settings

Datasets: We evaluate DADF-Det on three challenging low-light benchmarks: ExDark [35], DarkFace [36], and ACDC-Night [37]. ExDark contains 7,363 images covering 12 object categories under diverse low-light conditions. Following the data split adopted in our detection experiments, 3,000 images are used for training, 1,800 for validation, and 2,563 for testing. The validation set is used for hyperparameter tuning and model selection, whereas the test set is kept untouched for final evaluation.

DarkFace contains 6,000 real-world nighttime images with more than 30,000 face annotations and is used to evaluate face detection under extremely low illumination. The dataset is divided into training, validation, and test sets at a ratio of 3:1:1, corresponding to 3,600, 1,200, and 1,200 images, respectively. Compared with ExDark, DarkFace contains smaller targets and stronger noise interference, providing a more challenging evaluation setting.

ACDC-Night is used to evaluate the cross-task applicability of the proposed degradation-aware representation to nighttime semantic segmentation. Its nighttime subset is divided into 400 training images and 106 validation images. Following the experimental configuration of FeatEnHancer [38], all segmentation methods are trained for 20K iterations at an input resolution of 2048×1024. DADF-Seg retains the detection network's feature extraction structure and replaces the detection head with a DeepLabV3+ [33] decoder. Both category-wise IoU and mean IoU (mIoU) are computed over the 15 categories listed in Table 3 and reported on the validation set.

Training Settings: Unless otherwise specified, all detection experiments are conducted on an NVIDIA RTX 4090 GPU using PyTorch 1.12 and CUDA 11.6. The input resolution is set to 640×640. The SGD optimizer is employed with an initial learning rate of 0.01, momentum of 0.937, and weight decay of 5×10^{-4}. All detection models are trained for 300 epochs with a batch size of 8. Data augmentation includes Mosaic, MixUp, random flipping, and HSV perturbation. Unless otherwise stated, EnlightenGAN [5] is adopted as the default enhancement method to generate enhanced inputs.

For semantic segmentation, images are resized to 2048×1024, and DeepLabV3+ is used as the segmentation baseline. All methods are trained for 20K iterations using the corresponding MMSegmentation schedule, following FeatEnHancer [38]. The mIoU over the 15 evaluated categories is used as the primary evaluation metric.



Fig. 8 Example images from different datasets. (a) ExDark, (b) DarkFace, and (c) ACDC-Night.

For result provenance, Tables 1 and 2 include comparison results reported in the original papers using the same dataset splits as our experiments, although detector architectures and other experimental configurations may differ. These values indicate the overall performance level rather than a strictly controlled ranking. The exception is the low-light enhancement block in Table 1, which is trained and evaluated by us under the same YOLOv13 [39] protocol as the baseline. All rows in Table 3 are trained and evaluated by us under the unified semantic segmentation protocol described above, while Tables 4–12 use the same YOLOv13 baseline and dataset split. ExDark is evaluated using class-wise AP_{50} and mAP_{50}, and DarkFace reports Recall and mAP_{50}.

4.2 Comparison With Existing Methods

Tables 1–3 compare DADF-Det with representative low-light vision methods on ExDark, DarkFace, and ACDC-Night. The comparison covers enhancement methods [1, 2, 4, 5, 40–46] and detection-oriented approaches [6, 14, 16–19, 38, 47–49]. In Table 1, the low-light enhancement methods are evaluated by us under the fixed YOLOv13 protocol, whereas the low-light detection methods use the best results reported in their original papers. Table 2 likewise uses publicly reported results for

the comparison methods. These cross-paper values may involve different detectors and configurations and therefore indicate overall performance rather than a strict ranking. By contrast, all methods in Table 3 are trained and evaluated by us under the same DeepLabV3+ protocol, providing a controlled comparison on ACDC-Night.

4.2.1 Detection Performance on the ExDark Dataset

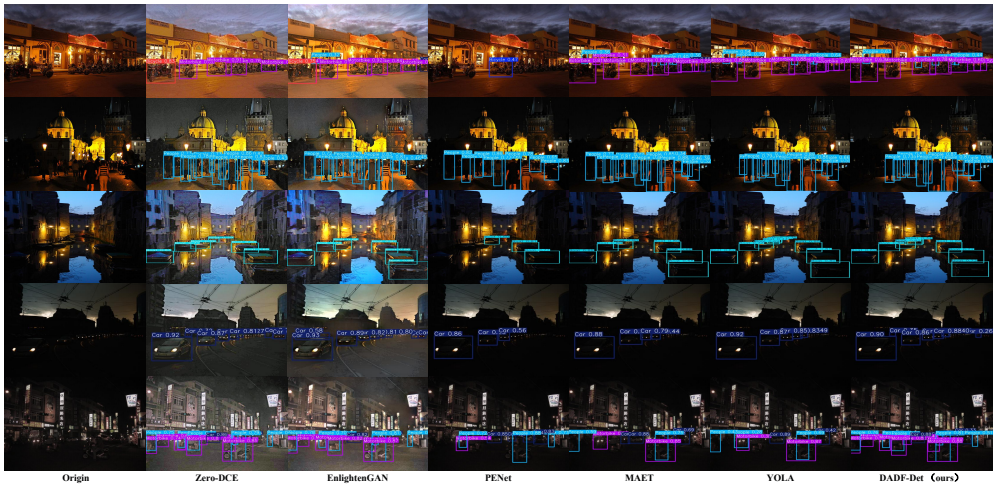


Fig. 9 Qualitative detection comparison of representative low-light enhancement and detection methods on the ExDark dataset.

Table 1 presents both controlled enhancement comparisons and publicly reported low-light detection results on ExDark. Under the controlled YOLOv13 protocol, the eight conventional enhance-then-detect pipelines achieve mAP_{50} values ranging from 76.4% to 77.6%, and five of them perform worse than the 77.0% baseline without enhancement. Because the detector, data split, and training settings are identical in this block, these results provide direct evidence that improved visual appearance does not necessarily translate into improved detection performance. DADF-Det instead reaches an mAP_{50} of 80.8%. In the separate publicly reported block, it exceeds FCMA-Det and EMV-Y by 1.2 and 1.1 percentage points, respectively; these cross-paper comparisons indicate the overall performance level rather than a strictly controlled ranking. Together, the results show that the proposed framework can exploit useful enhancement information while suppressing unreliable textures introduced during enhancement.

As illustrated in Fig. 9, DADF-Det produces more complete detections for small and densely distributed objects under uneven illumination. In the displayed examples, its predicted bounding boxes align more closely with the objects than those of the compared enhancement-based and low-light detection methods.

The response visualizations in Fig. 10 further show that DADF-Det concentrates activation more consistently on true object regions while reducing diffuse responses

Table 1 Class-wise detection performance on ExDark. The upper block reports low-light enhancement methods re-trained by us under the same YOLOv13 baseline, data split, and training protocol. The lower block reports the best results of low-light detection methods from their original papers using the same data split, although detector architectures and other configurations may differ.

Method	Bic.	Boa.	Bot.	Bus	Car	Cat	Cha.	Cup	Dog	Mot.	Ppl.	Tab.	mAP_{50}
Baseline	83.6	70.2	70.9	88.9	80.2	75.2	66.1	70.8	81.9	89.5	80.1	66.8	77.0
Low-Light Enhancement Methods													
KinD [2]	78.9	77.2	76.5	92.5	83.1	70.3	66.8	75.8	78.7	75.9	80.1	62.0	76.4
Zero-DCE [4]	85.4	75.8	70.9	90.2	81.2	76.3	65.6	71.3	81.9	88.2	80.6	64.0	77.6
PairLIE [40]	83.9	73.6	70.4	91.1	79.7	74.0	66.5	65.7	80.7	88.9	81.2	65.4	76.7
MBLLEN [41]	83.7	69.7	71.1	89.4	80.1	75.8	66.2	70.4	81.3	**90.4**	80.5	67.1	77.1
DarkIR [42]	82.6	70.4	72.8	90.0	77.4	78.1	67.7	71.0	82.1	86.7	80.1	63.0	76.8
LightDiff [43]	84.7	72.2	68.6	88.8	80.8	74.7	66.6	66.0	81.9	88.9	80.0	65.4	76.5
EnlightenGAN [5]	84.9	76.7	70.8	92.2	82.8	71.8	66.2	69.9	80.2	88.0	80.1	64.7	77.4
Multinex [44]	82.1	71.0	69.8	91.7	79.6	76.5	62.6	69.8	84.4	88.0	80.4	66.0	76.8
Low-Light Detection Methods													
MAET [14]	83.1	78.5	75.6	92.9	83.1	73.4	71.3	79.0	79.8	77.2	81.1	57.0	77.6
DENet [16]	80.9	79.2	**80.1**	90.7	84.5	70.7	72.0	79.3	80.1	76.7	82.4	58.0	77.9
PE-Y [17]	84.7	79.2	79.3	92.5	83.9	71.5	71.7	79.7	79.7	77.3	81.8	55.3	78.0
YOLA [18]	83.9	78.7	75.3	88.8	79.0	73.4	69.9	71.9	**86.8**	66.3	78.3	49.8	75.2
FCMA-Det [6]	83.9	80.2	79.4	92.8	84.5	74.8	73.6	81.2	81.1	80.7	83.4	59.2	79.6
FRBNet [19]	83.2	78.9	76.5	91.2	80.7	74.1	69.9	72.6	84.6	64.6	78.6	49.8	75.4
EMV-Y [47]	82.8	79.7	79.8	94.1	84.7	74.3	**74.1**	**83.1**	82.7	78.1	83.6	59.3	79.7
DADF-Det (Ours)	**86.1**	**80.7**	76.9	**94.3**	**85.2**	**79.8**	72.5	71.8	80.7	89.1	**84.6**	**68.0**	**80.8**

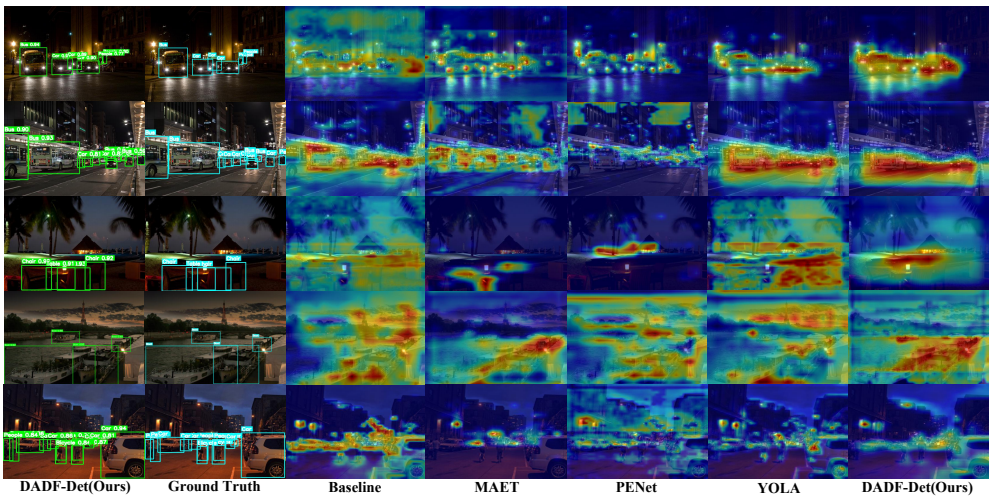


Fig. 10 Qualitative detection and response-heatmap comparison of representative low-light detection methods on the ExDark dataset. The first two columns show the DADF-Det predictions and ground-truth annotations, while the remaining columns visualize the responses of the baseline, MAET, PENet, YOLA, and DADF-Det.

from illumination changes and complex backgrounds. Compared with the baseline

and representative low-light detectors, the proposed method preserves clearer target-focused responses for vehicles, pedestrians, and densely distributed objects, which is consistent with its improved class-wise detection performance on ExDark.

The results suggest that improving low-light detection performance depends not only on recovering image visibility but also on effectively utilizing enhancement information for the detection task. DADF-Det establishes dynamic collaboration between enhanced semantic representations and original structural information through degradation-aware dual-branch fusion.

4.2.2 Detection Performance on the DarkFace Dataset

Table 2 presents the detection performance of different methods on the DarkFace dataset. DarkFace primarily targets face detection under extremely low illumination. It therefore places greater emphasis on detection stability for small-scale targets with weak boundary responses.

Table 2 Publicly Reported Original-System Results of Representative Methods on the DarkFace Dataset (%). Results for the comparison methods are taken from the original papers.

Method	Recall	mAP_{50}
Baseline	78.1	54.8
Low-Light Enhancement Methods		
EnlightenGAN [5]	74.8	52.4
LightDiff [43]	81.0	58.7
DarkIR [42]	81.4	60.4
Multinex [44]	81.5	59.6
Low-Light Detection Methods		
IAT [45]	79.7	58.3
MAET [14]	80.7	59.6
EMV-Y [47]	81.5	57.6
IADNet [49]	76.8	54.6
FeatEnHancer [38]	81.7	55.2
YOLA [18]	83.1	56.3
FRBNet [19]	82.7	63.6
DADF-Det (Ours)	**83.6**	**65.3**

DADF-Det achieves the highest Recall and mAP_{50}, reaching 83.6% and 65.3%, respectively. It improves upon FRBNet by 0.9 and 1.7 percentage points in the two metrics. Compared with DarkIR, the corresponding improvements are 2.2 and 4.9 percentage points. These results show that the proposed method improves face recall and overall detection performance.

As illustrated in Fig. 11, DADF-Det produces more complete face detections in extremely dark regions and reduces missed detections for small and densely distributed targets compared with the baseline and the conventional enhanced-image input.

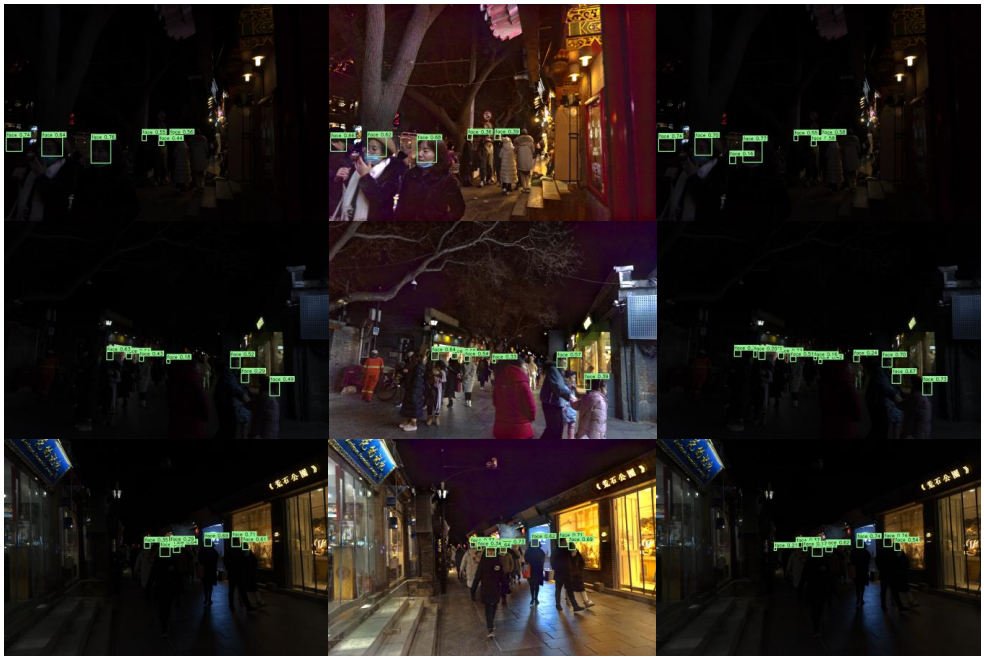


Fig. 11 Qualitative detection comparison on the DarkFace dataset. From left to right, the columns show the baseline results, the results obtained using enhanced images, and the results of DADF-Det, respectively.

Compared with conventional enhance-then-detect methods, DADF-Det does not rely on the enhanced image as the sole information source. Instead, it integrates structural features from the original image with semantic features from the enhanced image, providing additional cues for low-light detection.

4.2.3 Semantic Segmentation Performance on the ACDC-Night Dataset

Table 3 Semantic Segmentation Results of Representative Methods on the ACDC-Night Dataset (%). All rows are trained and evaluated by us under the same training and evaluation protocol. Both class-wise IoU and mIoU use the 15 listed categories.

Method	RO	SI	BU	WA	FE	PO	TL	TS	VE	TE	SK	PE	CA	TR	BI	mIoU
Baseline	90.0	61.4	74.2	32.8	34.4	45.7	49.8	31.2	68.8	14.6	80.4	27.1	62.1	76.3	14.4	50.9
RetinexNet [1]	89.4	61.0	70.6	30.1	28.1	42.4	47.6	25.7	65.8	8.6	77.3	21.5	54.8	67.4	8.2	46.6
FIDE [50]	90.0	60.7	72.8	32.4	34.1	43.3	47.9	26.1	67.0	13.7	78.0	26.5	57.1	71.0	12.4	48.8
KinD [2]	90.0	61.0	73.2	31.9	32.8	43.5	42.7	27.7	65.5	13.3	77.4	22.8	55.1	74.5	11.5	48.1
EnGAN [5]	89.7	58.9	73.7	32.8	31.8	44.7	49.2	26.2	67.3	14.2	77.8	25.0	59.0	71.2	7.8	48.6
ZeroDCE [4]	90.6	59.9	73.9	32.6	31.7	44.3	46.2	25.8	67.2	14.6	79.1	24.7	59.4	66.8	13.9	48.7
LightDiff [43]	90.1	58.5	71.5	29.9	31.2	42.8	45.2	27.5	65.1	10.1	74.1	25.4	55.1	70.6	11.6	47.3
DarkIR [42]	89.8	58.6	72.1	29.8	30.2	42.8	47.5	26.9	66.1	8.5	77.9	25.7	56.8	69.3	12.2	47.6
Multinex [44]	90.8	60.5	73.0	32.8	30.8	43.9	46.9	25.0	67.1	13.5	79.4	24.3	59.2	66.4	14.9	48.6
SSIENet [46]	89.6	59.3	72.5	29.9	31.7	45.4	43.9	24.5	66.7	10.6	78.3	22.8	52.6	71.1	5.4	46.9
CAL [48]	93.2	72.6	78.4	43.8	**46.5**	48.1	51.1	38.8	68.6	**14.9**	79.1	21.9	61.6	85.2	36.1	56.0
FeatEnHancer [38]	93.5	70.6	75.6	41.8	33.4	51.3	55.2	35.9	68.5	13.4	80.6	27.6	61.8	80.0	**51.2**	56.0
YOLA [18]	93.2	72.1	79.3	41.1	39.1	**53.1**	60.4	44.4	71.5	4.7	83.2	37.8	66.8	85.0	49.2	58.7
DADF-Seg (Ours)	**94.8**	**74.5**	**79.4**	**44.8**	43.2	52.9	**64.2**	**47.8**	**71.9**	10.8	**83.4**	**39.8**	**69.2**	**86.4**	49.4	**60.8**



Fig. 12 Qualitative semantic segmentation comparison of representative low-light vision methods on the ACDC-Night dataset.

Table 3 compares the semantic segmentation performance of different low-light vision methods on the ACDC-Night dataset [37]. Unlike the object detection tasks on ExDark and DarkFace, which rely on bounding-box supervision, ACDC-Night focuses on pixel-level semantic understanding in real nighttime driving scenes. It is used to

evaluate the cross-task applicability of the proposed representations after task-specific adaptation and training. Following the experimental protocol adopted by previous low-light semantic segmentation studies [38], DeepLabV3+ [33] is adopted as the segmentation baseline. The input images are resized to 2048×1024, and all methods are trained for 20K iterations under the same training and evaluation protocol. RO, SI, BU, WA, FE, PO, TL, TS, VE, TE, SK, PE, CA, TR, and BI denote road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, car, train, and bicycle, respectively. Both category-wise IoU and mIoU are computed over these 15 categories; truck, bus, rider, and motorcycle are excluded from both metrics.

To adapt the proposed framework from object detection to semantic segmentation, we retain the detection network's dual-branch feature extraction structure and degradation-aware fusion mechanisms, replacing only the detection head with a DeepLabV3+ decoder. ETB processes the enhanced input, while LEB extracts structural features from the original low-light image. The multi-scale branch features are projected to the required channel dimensions and fused through DDCF. The fused high-level features are passed to the decoder for pixel-level prediction, with low-level LEB features providing boundary information. This task-adapted variant is denoted as **DADF-Seg**.

The GDD controller is also retained to provide degradation-dependent reliability estimation for the fusion process, but its Stage-II supervision is adapted to the semantic segmentation objective. In the original detection task, the Stage-II target is determined according to the downstream detection utility based on bounding-box prediction. Since such detection-specific statistics are not directly applicable to semantic segmentation, we replace the detection-oriented utility objective with per-image segmentation performance measured by mIoU. Specifically, after Stage-I training, an eleven-point grid search is performed separately for each original–enhanced image pair in the training set by fixing the GDD gating value to $q \in \mathcal{Q}$, where

$$\mathcal{Q} = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}. \tag{34}$$

For the n-th training image pair, the optimal task-specific gating target is determined as

$$q^*_{\text{seg},n} = \arg\max_{q \in \mathcal{Q}} \text{mIoU}^{(n)}_{\text{train}}(q), \tag{35}$$

where $\text{mIoU}^{(n)}_{\text{train}}(q)$ denotes the per-image mIoU computed from the prediction for the n-th training image and its corresponding ground-truth segmentation mask when the gate is fixed to q. The search therefore yields a sample-specific target rather than a single target obtained from dataset-level mIoU. These targets are subsequently used to fine-tune the GDD prediction head with the same squared L_2 regression loss used in Stage II of the detection task. The loss for each training image pair is

$$\mathcal{L}^{\text{seg}}_{\text{GDD}} = \left\| q_\theta(\mathbf{x}^{(n)}_{\text{ori}}, \mathbf{x}^{(n)}_{\text{enh}}) - q^*_{\text{seg},n} \right\|_2^2, \tag{36}$$

where q_θ denotes the gating value predicted by GDD, $(\mathbf{x}_{\text{ori}}^{(n)}, \mathbf{x}_{\text{enh}}^{(n)})$ is the n-th training image pair, and $q_{\text{seg},n}^{*}$ is its segmentation-oriented utility target. The lightweight convolutional feature extractor of GDD is kept fixed during this stage, while its prediction head is optimized to regress these sample-specific targets. Both target generation and GDD refinement use only the training set; the validation set is not used to generate utility targets or fit the GDD and is retained for segmentation evaluation. Therefore, the proposed reliability-guided fusion mechanism can adapt its supervision according to the downstream task while preserving the degradation-aware representation learned from the original and enhanced inputs.

As illustrated in Fig. 12, DADF-Seg produces more continuous road boundaries and cleaner semantic regions under uneven nighttime illumination. These qualitative observations are consistent with the quantitative results reported in Table 3. DADF-Seg achieves the highest overall mIoU of 60.8%, outperforming YOLA and FeatEnHancer by 2.1 and 4.8 percentage points, respectively. It also maintains strong segmentation performance on important categories such as road (RO), traffic light (TL), traffic sign (TS), vegetation (VE), person (PE), and car (CA). These results indicate that the proposed degradation-aware dual-branch fusion strategy can effectively preserve structural boundaries while maintaining discriminative semantic information under challenging nighttime illumination.

Traditional enhancement methods such as RetinexNet, EnGAN, and ZeroDCE improve image brightness but do not outperform the segmentation baseline: their mIoU scores of 46.6%, 48.6%, and 48.7%, respectively, remain below the baseline's 50.9%. This observation indicates that perceptual enhancement alone is insufficient for reliable downstream semantic prediction. In contrast, DADF-Seg jointly exploits enhanced semantic information and original structural information, while GDD provides task-aware regulation of their relative contributions.

4.3 Controlled Comparison With Conventional Enhance-Then-Detect Methods

Publicly reported results are jointly affected by detector architectures, training strategies, and input resolutions, making it difficult to isolate the actual contribution of low-light enhancement to detection. To further evaluate the effectiveness of DADF-Det under different enhancement conditions, we establish a unified protocol on ExDark. YOLOv13 [39] is fixed as the baseline detector, and the data split, training procedure, and detection configuration remain unchanged across enhancers. Eight representative low-light enhancement methods [2, 4, 5, 40–44] are evaluated using both the conventional enhance-then-detect pipeline and the proposed DADF-Det dual-branch framework. For each enhancer, enhanced training inputs are generated separately, and both detection pipelines are trained under the same protocol; DADF-Det includes enhancer-specific GDD training on the corresponding original–enhanced pairs.

In Fig. 13, *Before* denotes the detection results obtained by directly feeding each enhanced image into the baseline detector. *After* denotes the results produced by DADF-Det using the same enhanced image together with its original low-light counterpart. In the displayed examples, DADF-Det detects more small and crowded targets, illustrating the potential benefit of degradation-aware dual-branch collaboration.



Fig. 13 Qualitative comparison between the conventional enhance-then-detect pipeline (*Before*) and the proposed DADF-Det framework (*After*) under different low-light enhancement methods.

Table 4 Detection Performance of Conventional Enhance-Then-Detect and DADF-Det With Different Enhancers on ExDark (%). All rows are trained and evaluated by us under the same protocol.

	Enhance-Then-Detect		DADF-Det	
Enhancer	mAP_{50}	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$
PairLIE [40]	76.7	50.9	79.7	51.8
Zero-DCE [4]	77.6	51.5	80.6	53.3
MBLLEN [41]	77.1	51.2	80.4	53.1
KinD [2]	76.4	50.8	79.2	51.9
LightDiff [43]	76.5	50.9	78.9	52.1
DarkIR [42]	76.8	51.2	79.5	52.6
Multinex [44]	76.8	51.4	80.0	52.9
EnlightenGAN [5]	77.4	51.7	**80.8**	**53.6**

Table 4 compares the two detection paradigms under different enhanced inputs. DADF-Det consistently improves performance for all eight enhancement methods. Relative to the conventional pipeline, the gains range from 2.4 to 3.4 percentage points in mAP_{50} and from 0.9 to 1.9 percentage points in $mAP_{50:95}$. With EnlightenGAN, Zero-DCE, and MBLLEN, DADF-Det achieves mAP_{50} values of 80.8%, 80.6%, and 80.4%, respectively. These results demonstrate that the improvement is not tied to a specific enhancer but arises from the collaborative use of enhanced semantic information and structural cues from the original low-light image.

Across the eight enhancement methods, DADF-Det yields average improvements of 3.0 percentage points in mAP_{50} and 1.5 percentage points in $mAP_{50:95}$. These results show gains across the evaluated enhancers under enhancer-specific training.

They do not establish transfer of a fixed model to unseen enhancers, which is evaluated separately in Table 5.

4.3.1 Cross-Enhancer Generalization

To evaluate whether DADF-Det generalizes to enhancement methods not observed during training, we train the model using only the image pairs generated by EnlightenGAN and evaluate the resulting checkpoint on the ExDark test split using enhanced inputs produced by EnlightenGAN, Zero-DCE, Multinex, and PairLIE. No fine-tuning is performed when changing the test-time enhancer, and the original low-light input remains unchanged.

Table 5 Cross-Enhancer Generalization on ExDark (%). The model is trained using EnlightenGAN and directly evaluated with different test-time enhancers without fine-tuning.

Test Enhancer	EnlightenGAN [5]	Zero-DCE [4]	Multinex [44]	PairLIE [40]
mAP_{50}	**80.8**	80.1	79.7	79.5

As shown in Table 5, the model achieves 80.8% mAP_{50} when the training and test enhancers are both EnlightenGAN. Replacing the test-time enhancer with Zero-DCE, Multinex, or PairLIE yields 80.1%, 79.7%, and 79.5%, respectively. The performance range across all four enhancers is 1.3 percentage points. These results support the applicability of the frozen checkpoint to the three evaluated unseen enhancers without fine-tuning.

4.4 Ablation Studies

To further analyze the contribution of each proposed component, we conduct ablation experiments and parameter sensitivity analyses on ExDark. All experiments use YOLOv13 as the baseline detector and maintain identical data splits, input resolutions, and training procedures. We compare the detection metrics of different component configurations and auxiliary loss settings. Throughout the ablation experiments, all configurations without GDD use a fixed gating value of $q = 0.5$, corresponding to $\alpha_g = \beta_g = 0.5$. In configurations that retain DDCF, this setting gives $s = 2q - 1 = 0$, so the discrepancy-dependent residual modulation reduces to the identity, $\widetilde{F}^i_{fuse} = F^i_g$. DDCF is not entirely disabled: equal-weight cross-branch fusion remains active, and the subsequent hierarchical attention mechanisms (CBAM and AIFI) continue to refine the fused features when included in the configuration.

4.4.1 Ablation Study of the GDD Module

The objective of GDD is not to assess whether an enhanced image is perceptually more appealing, but to estimate whether it provides useful information for downstream detection. Its training strategy therefore directly affects the consistency between the predicted gating score and the resulting detection performance. We compare four configurations: fixed branch weights without GDD ($q = 0.5$), degradation-statistics

pretraining only, detection-utility refinement only, and the proposed two-stage training strategy.

Table 6 Ablation Study of GDD Training Strategies on ExDark (%). All rows are trained and evaluated by us under the same protocol.

Training Strategy	mAP_{50}	$mAP_{50:95}$
Without GDD (fixed $q = 0.5$)	78.7	52.1
Degradation pretraining only	79.3	52.4
Detection-utility refinement only	79.8	53.2
Two-stage training	**80.8**	**53.6**

As shown in Table 6, the proposed two-stage strategy achieves the highest scores among the evaluated configurations, improving mAP_{50} and $mAP_{50:95}$ by 2.1 and 1.5 percentage points, respectively, over the fixed-weight configuration. Degradation-statistics pretraining initializes GDD using global degradation characteristics, while detection-utility refinement introduces downstream task supervision. The reported results support using both stages rather than either stage alone.

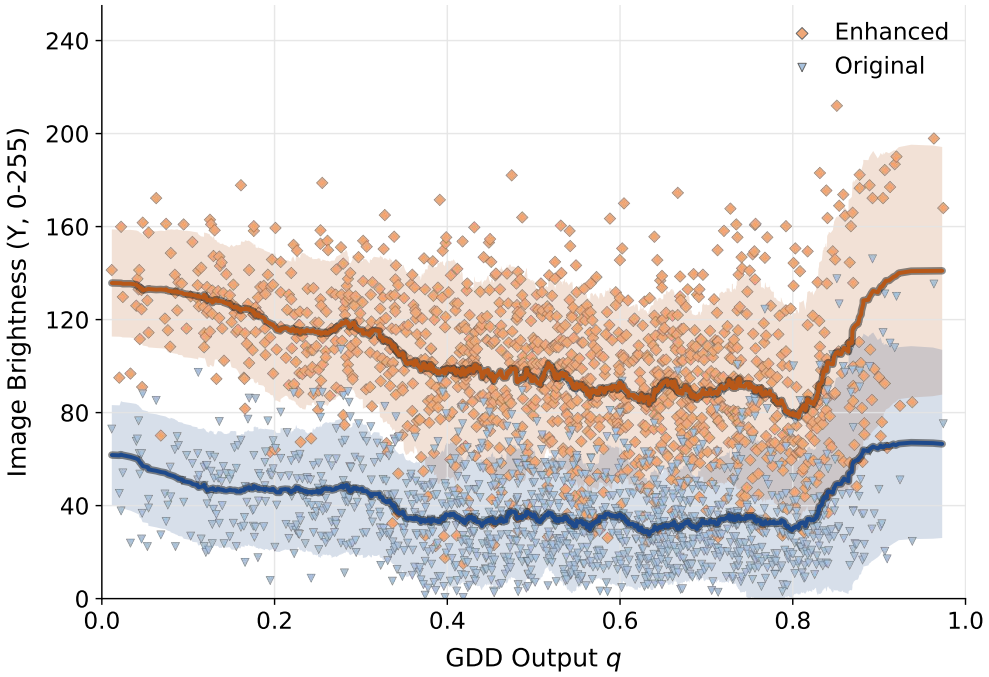


Fig. 14 Auxiliary analysis of the GDD utility score (q) and image brightness on representative ExDark samples. Blue and orange denote the original and EnlightenGAN-enhanced images, respectively; the curves represent moving averages.

Figure 14 provides an auxiliary analysis of the predicted utility score (q), its distribution, and its relationship with image brightness on representative ExDark samples. The solid curves are moving averages after sorting samples by q, and the shaded regions indicate the corresponding 1.4σ local variation. Most displayed low-brightness samples have q values within 0.35–0.80, rather than near the endpoints. This plot describes the observed score distribution and brightness relationship; it does not directly establish correspondence between q and actual detection utility.

4.4.2 Ablation Study of ETB

Table 7 Ablation Study of LTM, GTM, and GDD on ExDark (%). Each configuration is obtained from the complete model by removing the indicated ETB modules and/or disabling GDD, while retaining the remaining components. "–" denotes an omitted component; configurations without GDD use $q = 0.5$. All rows are trained and evaluated by us under the same protocol.

LTM	GTM	GDD	mAP_{50}	$mAP_{50:95}$
✓	–	–	78.2	51.8
–	✓	–	78.5	52.0
✓	–	✓	79.0	52.8
–	✓	✓	79.5	53.1
✓	✓	–	78.7	52.1
✓	✓	✓	**80.8**	**53.6**

ETB consists of the Local Texture Module (LTM), the Global Texture Module (GTM), and degradation-aware gating provided by GDD. LTM preserves fine-grained local details, whereas GTM enlarges the effective receptive field to capture global semantic context. GDD regulates both modeling paths according to the estimated detection utility of the enhanced image, thereby limiting the unconditional propagation of unreliable enhanced textures.

Table 7 reports ablations obtained by starting from the complete DADF-Det model and removing the LTM and/or GTM modules within ETB, and disabling GDD where indicated. The remaining model components, including LEB, DDCF, and boundary consistency supervision, are retained. Thus, the rows evaluate reduced versions of the complete model rather than standalone LTM or GTM networks. When GDD is omitted, its predicted gate is replaced by the fixed value $q = 0.5$.

Without GDD, retaining GTM alone within ETB achieves 78.5% mAP_{50} and 52.0% $mAP_{50:95}$, exceeding the corresponding LTM configuration by 0.3 and 0.2 percentage points, respectively. Enabling GDD in the configuration retaining LTM but omitting GTM improves the two metrics by 0.8 and 1.0 percentage points, respectively. For the configuration retaining GTM but omitting LTM, the corresponding gains are 1.0 and 1.1 percentage points. These results indicate that both local texture selection and global semantic modeling benefit from degradation-aware regulation.

The complete configuration achieves the highest reported scores, with an mAP_{50} of 80.8% and an $mAP_{50:95}$ of 53.6%. Compared with combining LTM and GTM without GDD, it yields improvements of 2.1 and 1.5 percentage points, respectively.

These results support combining LTM and GTM with GDD-guided modulation in the complete framework.

4.4.3 Ablation Study of Attention Mechanisms in DDCF

To evaluate the contributions of the attention mechanisms in DDCF, we compare four configurations on ExDark: no attention, CBAM only, AIFI only, and CBAM combined with AIFI, as shown in Table 8. All configurations retain global reliability fusion and discrepancy calibration, with the remaining model components and training settings unchanged. Here, "No attention" specifically denotes the model without CBAM and AIFI in DDCF, rather than a model without DDCF or without attention mechanisms elsewhere in the network. When enabled, CBAM refines the P3 and P4 features, while AIFI models global dependencies in the P5 feature.

Table 8 Ablation Study of Attention Mechanisms in DDCF on ExDark (%). All configurations retain global reliability fusion and discrepancy calibration. "No attention" denotes omission of CBAM and AIFI from DDCF only. All rows are trained and evaluated by us under the same protocol.

Attention Configuration	mAP_{50}	$mAP_{50:95}$
No attention	80.0	52.8
CBAM only	80.4	53.2
AIFI only	80.5	53.3
CBAM + AIFI	**80.8**	**53.6**

The configuration without attention in DDCF achieves 80.0% mAP_{50} and 52.8% $mAP_{50:95}$. Adding CBAM improves both metrics by 0.4 percentage points, while adding AIFI improves both by 0.5 percentage points. Combining CBAM and AIFI achieves the highest reported scores of 80.8% and 53.6%, yielding gains of 0.8 percentage points in both metrics over the no-attention configuration. These results support the benefit of attention-based refinement at different feature levels within DDCF.

4.4.4 Ablation Study of the Loss Function

The boundary consistency loss combines a feature-gradient consistency term and a gradient-matching term. The former emphasizes the alignment of local boundary responses, whereas the latter penalizes large residual gradient discrepancies. We separately evaluate the two constraints and their combined formulation to determine their contributions to detection and localization performance.

Table 9 shows that both individual constraints improve the reported detection metrics over the detection-only objective, which achieves 79.7% mAP_{50} and 52.7% $mAP_{50:95}$. Feature-gradient consistency increases the two metrics by 0.2 and 0.1 percentage points, respectively, while gradient matching yields gains of 0.5 and 0.4 points. Combining the two terms as $\mathcal{L}_{bound}$ achieves 80.8% and 53.6%, improving the two metrics by 1.1 and 0.9 percentage points over the detection-only objective. This result supports the benefit of combining the two constraints.

Table 9 Ablation Study of Boundary Consistency Loss on ExDark (%). Loss configurations are abbreviated: each auxiliary loss term is weighted by $\lambda_{grad} = 0.10$, including both terms in $\mathcal{L}_{bound} = \mathcal{L}_{feat} + \mathcal{L}_{match}$. All rows are trained and evaluated by us under the same protocol.

Loss Configuration	mAP_{50}	$mAP_{50:95}$
$\mathcal{L}_{det}$	79.7	52.7
$\mathcal{L}_{det} + \mathcal{L}_{feat}$	79.9	52.8
$\mathcal{L}_{det} + \mathcal{L}_{match}$	80.2	53.1
$\mathcal{L}_{det} + \mathcal{L}_{bound}$	**80.8**	**53.6**

Table 10 Sensitivity Analysis of λ_{grad} on ExDark (%). All rows are trained and evaluated by us under the same protocol.

λ_{grad}	mAP_{50}	$mAP_{50:95}$
0.01	79.7	52.8
0.05	80.3	53.2
0.10	**80.8**	**53.6**
0.20	80.2	53.1
0.50	79.5	52.7

Table 10 analyzes the sensitivity to λ_{grad}. Increasing the weight from 0.01 to 0.10 improves mAP_{50} from 79.7% to 80.8% and $mAP_{50:95}$ from 52.8% to 53.6%. A small weight may provide insufficient boundary supervision, whereas an excessively large weight can overemphasize gradient consistency and suppress semantic discrimination. At $\lambda_{grad} = 0.50$, the two metrics decrease to 79.5% and 52.7%, respectively. The best performance is obtained at $\lambda_{grad} = 0.10$, which provides a suitable balance between localization constraints and semantic representation. We therefore adopt 0.10 as the default setting in all remaining experiments.

4.4.5 Ablation Study of the Core Modules

We further evaluate the main components by progressively extending the baseline detector. In Table 11, GDD is enabled together with DDCF. All configurations without DDCF use the Without-GDD setting with fixed $q = 0.5$, so GDD does not adaptively modulate ETB; when both branches are present, they are fused with fixed equal weights. Consequently, the transition to DDCF in this table evaluates the joint addition of GDD and DDCF, rather than DDCF alone. The Boundary Consistency Loss is varied separately as indicated.

As illustrated in Fig. 15, LEB emphasizes compact structural regions around object boundaries, whereas ETB captures broader semantic responses from enhanced textures. DDCF consolidates these complementary activations and produces more concentrated responses on target regions while suppressing irrelevant background areas.

As shown in Table 11, adding LEB to the baseline improves mAP_{50} and $mAP_{50:95}$ by 1.4 and 0.7 percentage points, respectively. ETB alone yields smaller gains of 0.4 and 0.2 points. Combining the two branches with fixed equal weights results in an mAP_{50} of 77.1%, below the LEB-only result. This observation shows that the two inputs do not provide additive gains under fixed-weight fusion.

Table 11 Ablation Study of the Core Components on ExDark (%). GDD is enabled with DDCF; rows without DDCF use fixed $q = 0.5$ without GDD. All rows are trained and evaluated by us under the same protocol.

Baseline	LEB	ETB	DDCF	Boundary Consistency Loss	mAP_{50}	$mAP_{50:95}$
✓	–	–	–	–	77.0	51.6
✓	✓	–	–	–	78.4	52.3
✓	–	✓	–	–	77.4	51.8
✓	✓	✓	–	–	77.1	52.2
✓	✓	✓	–	✓	78.2	52.4
✓	✓	✓	✓	–	79.7	52.7
✓	✓	✓	✓	✓	**80.8**	**53.6**



Fig. 15 Qualitative visualization of detection results and response heatmaps produced by LEB, Dynamic Discrepancy Calibration Fusion (DDCF), and ETB in representative low-light scenes.

Enabling GDD together with DDCF raises mAP_{50} from 77.1% to 79.7%. Adding the Boundary Consistency Loss improves mAP_{50} by 1.1 points both without and with GDD/DDCF, while its gain in $mAP_{50:95}$ increases from 0.2 to 0.9 points in the latter configuration. The complete model reaches 80.8% mAP_{50} and 53.6% $mAP_{50:95}$, outperforming the GDD/DDCF configuration without the Boundary Consistency Loss by 1.1 and 0.9 percentage points, respectively. These results support the benefit of boundary supervision within the complete framework.

The complete framework achieves the highest scores among the configurations in Table 11, improving the baseline by 3.8 percentage points in mAP_{50} and 2.0 percentage points in $mAP_{50:95}$.

4.4.6 Computational cost and inference speed

Enhanced images are generated in advance. FPS is measured with a batch size of 8 when the enhanced inputs are already available, and therefore reports detection

throughput rather than end-to-end throughput including enhancement. For DADF-Det, FPS, Params, and FLOPs include GDD but exclude the external enhancement model.

Table 12 Model Complexity, Inference Speed, and Detection Accuracy on ExDark. Params and FLOPs are reported in millions (M) and billions (G), respectively. FPS is measured at batch size 8 with pre-generated enhanced inputs. DADF-Det measurements include GDD and exclude the external enhancer.

Model	Params (M)	FLOPs (G)	FPS	mAP_{50} (%)
Baseline	27.6	89.3	78.3	77.0
DADF-Det	40.6	164.2	21.4	**80.8**

Table 12 compares the model complexity, inference speed, and detection accuracy of the baseline and DADF-Det. DADF-Det improves mAP_{50} from 77.0% to 80.8%, a gain of 3.8 percentage points, while increasing the parameter count from 27.6M to 40.6M and the computational cost from 89.3G to 164.2G FLOPs. The corresponding increases are 13.0M parameters and 74.9G FLOPs. Meanwhile, inference speed decreases from 78.3 to 21.4 FPS.

These results highlight the accuracy–efficiency trade-off of the proposed framework: the complementary branches and adaptive fusion improve detection accuracy but require additional computation and reduce throughput. DADF-Det therefore prioritizes detection accuracy over inference speed. For deployment under strict computational or latency constraints, reducing the cost of dual-branch feature extraction and fusion remains an important direction for future optimization.

4.5 Failure Case Analysis

Figure 16 presents several representative failure cases of DADF-Det on the ExDark dataset. Each case contains the original low-light image (*Original*), the enhanced image (*Enhance*), the ground-truth bounding boxes (*Ground Truth*), and the detection results produced by the model (*Failure Detection*).

Low-light and fog-induced local degradation: The first scene contains two bicycles under low-light and foggy conditions. DADF-Det detects the bicycle with clearer shape cues at a confidence of 0.85 but misses the one in the darker, more blurred region. Reduced local contrast and obscured boundaries may explain the missed detection despite the improved overall visibility after enhancement.

Large-scale color shifts introduced by enhancement: The second scene exhibits a large-area purple color shift after enhancement. Of the two bottles, the one in the brighter region is detected at a confidence of 0.90, while the darker bottle is missed. The combination of altered appearance in the enhanced image and weak structural cues in the original image may make local discrepancy calibration less effective.

Small-scale crowded objects with background interference: The third scene contains overlapping pedestrians and local reflections. The model misses some targets, and several low-confidence bounding boxes deviate from the annotated locations. Occlusion



Fig. 16 Representative failure cases of DADF-Det on the ExDark dataset. From top to bottom, the rows show the original low-light images, enhanced images, ground-truth annotations, and failure detection results, respectively.

and reflective background responses may obscure object contours and make it difficult to distinguish meaningful edges from noise, potentially limiting the benefit of discrepancy calibration and boundary supervision.

Extremely low illumination with severely degraded object features: The fourth and remaining cases show extremely dark surveillance scenes with few visible object details. DADF-Det detects one chair at a confidence of 0.46 in the fourth case, while no effective detections are obtained in the remaining cases. Residual noise after enhancement may overwhelm the weak target responses in these scenes.

Together, these examples illustrate a limitation of utility-guided fusion: estimating branch reliability and calibrating discrepancies may be insufficient when both inputs contain weak or corrupted target evidence. The explanations above are qualitative interpretations of the displayed cases, rather than isolated causal tests of individual modules. These observations motivate the more fine-grained enhancement-utility modeling discussed in Section 5.

5 Discussion

DADF-Det treats enhancement utility, rather than visual quality alone, as the criterion for dual-branch fusion. GDD provides sample-level reliability guidance, while DDCF calibrates local cross-branch discrepancies, enabling more selective use of enhanced information.

The framework nevertheless has practical limitations. First, the dual-branch architecture with GDD increases model size and computational demand relative to the baseline, reducing inference speed. Dependence on an external enhancement model further adds to the processing requirements, limiting deployment in resource-constrained or latency-sensitive applications.

Second, adaptation involves a relatively involved training workflow. Under dataset- or enhancer-specific retraining, changing the dataset or enhancement method requires generating new enhanced training inputs for ETB and retraining GDD on the resulting original–enhanced pairs with task annotations. Thus, adaptation requires repeating both enhanced-data preparation and GDD training, rather than merely replacing the

input images. This training burden is distinct from the frozen-checkpoint transfer evaluated in Table 5.

Third, GDD provides only a global utility score, not a pixel-wise utility map. Local discrepancy calibration remains insufficient when enhancement severely distorts object features and the original image also lacks reliable structural cues. Broader validation beyond the three evaluated datasets is also needed.

Future work will focus on lighter architectures, simpler adaptation procedures, and spatially resolved utility estimation, with temporal consistency modeling for video-based low-light vision.

6 Conclusion

This paper proposes DADF-Det, a degradation-aware dual-branch fusion framework for low-light object detection. To address the inconsistency between the visual quality of enhanced images and their utility for object detection, a Global Degradation Discriminator (GDD) is introduced to estimate enhancement utility and adaptively regulate the global contributions of ETB and LEB. To handle local artifacts and cross-domain feature discrepancies, a Dynamic Discrepancy Calibration Fusion (DDCF) module performs pixel-wise discrepancy calibration between the two branches. In addition, a gradient-domain boundary consistency loss is incorporated to preserve structural continuity and improve boundary stability during feature optimization.

Experimental results demonstrate that DADF-Det achieves consistent improvements on ExDark, DarkFace, and ACDC-Night. The proposed framework also maintains reliable generalization performance across different enhancement methods and detection settings.

References

[1] Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. In: British Machine Vision Conference (2018)

[2] Zhang, Y., Zhang, J., Guo, X.: Kindling the darkness: A practical low-light image enhancer. In: Proceedings of the ACM International Conference on Multimedia, pp. 1632–1640 (2019). https://doi.org/10.1145/3343031.3350926

[3] Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., Zhang, Y.: Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023). https://doi.org/10.1109/ICCV51070.2023.01149

[4] Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: 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). https://doi.org/10.1109/CVPR42600.2020.00185

[5] Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: EnlightenGAN: Deep light enhancement without paired supervision. IEEE Transactions on Image Processing **30**, 2340–2349 (2021) https://doi.org/10.1109/TIP.2021.3051462

[6] Ji, J., Zhao, Y., Zhang, Y., Zuo, X., Wang, C., Shi, F.: FCMA-Det: Low-light image object detection based on feature complementarity and multicontent aggregation. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–14 (2025) https://doi.org/10.1109/TGRS.2025.3596581

[7] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2017) https://doi.org/10.1109/TPAMI.2016.2577031

[8] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91

[9] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37 (2016). https://doi.org/10.1007/978-3-319-46448-0_2

[10] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 (2020). https://doi.org/10.1007/978-3-030-58452-8_13

[11] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)

[12] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.-Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: International Conference on Learning Representations (2023)

[13] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: DETRs beat YOLOs on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024). https://doi.org/10.1109/CVPR52733.2024.01605

[14] Cui, Z., Qi, G.-J., Gu, L., You, S., Zhang, Z., Harada, T.: Multitask AET with orthogonal tangent regularity for dark object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021). https://doi.org/10.1109/ICCV48922.2021.00255

[15] Kalwar, S., Patel, D., Aanegola, A., Konda, K.R., Garg, S., Krishna, K.M.: GDIP:

Gated differentiable image processing for object detection in adverse conditions. In: IEEE International Conference on Robotics and Automation (2023). https://doi.org/10.1109/ICRA48891.2023.10160356

[16] Qin, Q., Chang, K., Huang, M., Li, G.: DENet: Detection-driven enhancement network for object detection under adverse weather conditions. In: Asian Conference on Computer Vision, pp. 491–507 (2022). https://doi.org/10.1007/978-3-031-26313-2_30

[17] Yin, X., Yu, Z., Fei, Z., Lv, W., Gao, X.: PE-YOLO: Pyramid enhancement network for dark object detection. In: International Conference on Artificial Neural Networks, pp. 163–174 (2023). https://doi.org/10.1007/978-3-031-44195-0_14

[18] Hong, M., Cheng, S., Huang, H., Fan, H., Liu, S.: You only look around: Learning illumination-invariant feature for low-light object detection. In: Advances in Neural Information Processing Systems (2024)

[19] Sun, F., Li, C., Yang, K., Pan, Y., Yu, H., Zhang, X., Li, Y.: FRBNet: Revisiting low-light vision through frequency-domain radial basis network. In: Advances in Neural Information Processing Systems (2025)

[20] Du, Z., Shi, M., Deng, J.: 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). https://doi.org/10.1109/CVPR52733.2024.01204

[21] Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1395–1403 (2015). https://doi.org/10.1109/ICCV.2015.164

[22] Li, M., Liu, J., Yang, W., Sun, X., Guo, Z.: Structure-revealing low-light image enhancement via robust retinex model. IEEE Transactions on Image Processing **27**(6), 2828–2841 (2018) https://doi.org/10.1109/TIP.2018.2810539

[23] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594

[24] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(4), 834–848 (2018) https://doi.org/10.1109/TPAMI.2017.2699184

[25] Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W.: Multi-level wavelet-CNN for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision

and Pattern Recognition Workshops (2018). https://doi.org/10.1109/CVPRW.2018.00121

[26] Sobel, I., Feldman, G.: A 3x3 Isotropic Gradient Operator for Image Processing. Stanford Artificial Intelligence Project (1968)

[27] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745

[28] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: European Conference on Computer Vision, pp. 3–19 (2018). https://doi.org/10.1007/978-3-030-01234-2_1

[29] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[30] Yang, W., Yuan, Y., Ren, W., Liu, J., Scheirer, W.J., Wang, Z., Zhang, T., Zhong, Q., Xie, D., Pu, S., Zheng, Y., Qu, Y., Xie, Y., Chen, L., Li, Z., Hong, C., Jiang, H., Yang, S., Liu, Y., Qu, X., Wan, P., Yan, S., Feng, J., Shao, L., Kuo, C.C.J., Li, Q.: Advancing image understanding in poor visibility environments: A collective benchmark study. IEEE Transactions on Image Processing **29**, 5737–5752 (2020) https://doi.org/10.1109/TIP.2020.2981922

[31] Zhang, L., Zhang, L., Mou, X., Zhang, D.: FSIM: A feature similarity index for image quality assessment. IEEE Transactions on Image Processing **20**(8), 2378–2386 (2011) https://doi.org/10.1109/TIP.2011.2109730

[32] Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018). https://doi.org/10.1109/CVPR.2018.00813

[33] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision, pp. 801–818 (2018). https://doi.org/10.1007/978-3-030-01234-2_49

[34] Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Barnard, K.: Attentional feature fusion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3559–3568 (2021)

[35] Loh, Y.P., Chan, C.S.: Getting to know low-light images with the exclusively dark dataset. Computer Vision and Image Understanding **178**, 30–42 (2019) https://doi.org/10.1016/j.cviu.2018.10.010

[36] Yang, W., et al.: DARK FACE: Face Detection in Low-Light Conditions.

IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)

[37] Sakaridis, C., Dai, D., Gool, L.V.: ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021). https://doi.org/10.1109/ICCV48922.2021.01059

[38] Hashmi, K.A., Kallempudi, G., Stricker, D., Afzal, M.Z.: 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). https://doi.org/10.1109/ICCV51070.2023.00619

[39] Lei, M., Li, S., Wu, Y., Hu, H., Zhou, Y., Zheng, X., Ding, G., Du, S., Wu, Z., Gao, Y.: YOLOv13: Real-time object detection with hypergraph-enhanced adaptive visual perception. arXiv preprint arXiv:2506.17733 (2025) https://doi.org/10.48550/arXiv.2506.17733

[40] Fu, Z., Yang, Y., Tu, X., Huang, Y., Ding, X., Ma, K.-K.: Learning a simple low-light image enhancer from paired low-light instances. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023). https://doi.org/10.1109/CVPR52729.2023.02131

[41] Lv, F., Lu, F., Wu, J., Lim, C.: MBLLEN: Low-light image/video enhancement using CNNs. In: British Machine Vision Conference (2018)

[42] Feijoo, D., Benito, J.C., Garcia, A., Conde, M.V.: DarkIR: Robust low-light image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10879–10889 (2025). https://doi.org/10.1109/CVPR52734.2025.01016

[43] Li, J., Li, B., Tu, Z., Liu, X., Guo, Q., Juefei-Xu, F., Xu, R., Yu, H.: Light the night: A multi-condition diffusion framework for unpaired low-light enhancement in autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15205–15215 (2024). https://doi.org/10.1109/CVPR52733.2024.01440

[44] Brateanu, A., Mu, T., Ancuti, C., Ancuti, C.: Multinex: Lightweight low-light image enhancement via multi-prior retinex. arXiv preprint arXiv:2604.10359 (2026) https://doi.org/10.48550/arXiv.2604.10359

[45] Cui, Z., Li, K., Gu, L., Su, S., Gao, P., Jiang, Z., Qiao, Y., Harada, T.: You only need 90k parameters to adapt light: A lightweight transformer for image enhancement and exposure correction. In: British Machine Vision Conference (2022)

[46] Zhang, Y., Di, X., Zhang, B., Wang, C.: Self-supervised image enhancement network: Training with low-light images only. arXiv preprint arXiv:2002.11300 (2020) https://doi.org/10.48550/arXiv.2002.11300

[47] Feng, X., Zeng, J., Wang, S., He, Z.: Toward highly efficient semantic-guided machine vision for low-light object detection. In: British Machine Vision Conference (2024)

[48] Xue, X., He, J., Ma, L., Wang, Y., Fan, X., Liu, R.: Best of both worlds: See and understand clearly in the dark. In: Proceedings of the ACM International Conference on Multimedia, pp. 2154–2162 (2022). https://doi.org/10.1145/3503161.3548259

[49] Cao, W., Yang, P., Lyu, W.: Explicit illumination modeling for object detection in low-light environments. Electronics **15**(10), 2057 (2026) https://doi.org/10.3390/electronics15102057

[50] Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J.: From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3063–3072 (2020). https://doi.org/10.1109/CVPR42600.2020.00313