Expert Systems With Applications 297 (2026) 129476

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Retinex-guided illumination recovery and progressive feature adaptation for real-world nighttime UAV-based vehicle detection

Li Chen [a], Hongbin Deng [a], Guanghong Liu [b], Rob Law [c], Dongfang Li [d,*], Edmond Q. Wu [e], Limin Zhu [f]

[a] School of Mechatronical Engineering, Beijing Institute of Technology, Beijing, 100081, China
[b] School of information Science and Technology, University of Science and Technology of China, Anhui, 230026, China
[c] University of Macau, Macao, 999078, China
[d] School of Electrical Engineering and Automation, Fuzhou University, Fuzhou, 350108, China
[e] Department of Computer Science and Engineering, Shanghai JiaoTong University, Shanghai, 200240, China
[f] School of Mechanical Engineering, Shanghai JiaoTong University, Shanghai, 200240, China

ARTICLE INFO

Keywords:
Differential transformer
Nighttime image enhancement
UAV-based vehicle detection
Progressive feature adaptation

ABSTRACT

Nighttime vehicle detection in Unmanned Aerial Vehicle (UAV) imagery is critical for intelligent transportation systems yet faces significant challenges due to low signal-to-noise ratios and pervasive small and medium-scale objects. Existing methods suffer from two critical limitations: (1) conventional low-light enhancement approaches prioritize human perception over downstream detection tasks, and (2) feature fusion frameworks exhibit inadequate cross-level interactions and computational inefficiency for UAV platforms. To address these gaps, we propose the Retinex-guided illumination Differential Transformer Detection network (ReDT-Det), which integrates a nighttime image enhancer with a robust vehicle detection module. Our approach leverages Retinex principles to design an illumination-fused differential transformer block for preliminary image enhancement and illumination recovery, which can effectively improve the quality of nighttime images while preserving critical details. To address the issue of small and medium-scale objects, we introduce a dilation-wise residual cross-stage partial module to enhance the ability to capture fine-grained features. During the feature fusion stage, we propose two key modules: a cross-level feature adaptive adjustment module for the effective integration of multi-scale features and a small object auxiliary feature module specifically designed to enhance the representation of small-scale objects. To validate our method, we curated a comprehensive benchmark dataset for real-world nighttime UAV-based vehicle detection, named NightDrone-Mix. Extensive comparative experiments demonstrate that ReDT-Det outperforms various state-of-the-art image enhancement and detection methods, highlighting its advantages in both accuracy and effectiveness. Additionally, we evaluated ReDT-Det on the DroneVehicle(Night) and ExDark datasets to assess its performance in detecting dark objects, achieving equally promising results.

1. Introduction

Vehicle detection, a critical technology in urban intelligent transportation systems, is widely used for traffic flow monitoring, traffic violation detection, and accident response (Telikani et al., 2024). Unmanned Aerial Vehicles (UAVs) equipped with cameras offer a unique advantage in intelligent urban management due to their wide field of view and ability to collect vehicle data from high-altitude perspectives. While UAV-based vehicle detection has achieved significant success under normal lighting conditions (Hoanh & Vu Pham, 2024; Ying et al., 2024; Zhu et al., 2024), its effectiveness is limited when applied directly to nighttime imagery, as shown in Fig. 1(b). This limitation arises from two key challenges in nighttime UAV-collected datasets: (1) Low visibility and high noise: Nighttime imaging typically suffers from poor illumination and low signal-to-noise ratios, making it difficult for existing methods to accurately detect and analyze objects, as illustrated in Fig. 1(a). (2) Small and medium scale object detection: Imagery captured by UAVs at high altitudes often results in blurred or low-resolution object representations, as shown in Table 1. This creates challenges in distinguishing critical features necessary for effective detection. This outcome highlights the need for specialized techniques to address these challenges in nighttime UAV-based vehicle detection.

* Corresponding author.
E-mail addresses: chenli_bit@bit.edu.cn (L. Chen), denghongbin@bit.edu.cn (H. Deng), lghong@mail.ustc.edu.cn (G. Liu), roblaw@um.edu.mo (R. Law), lidongfang@fzu.edu.cn (D. Li), edmondqwu@sjtu.edu.cn (E.Q. Wu), zhulm@sjtu.edu.cn (L. Zhu).

https://doi.org/10.1016/j.eswa.2025.129476
Received 14 April 2025; Received in revised form 11 August 2025; Accepted 20 August 2025
Available online 24 August 2025

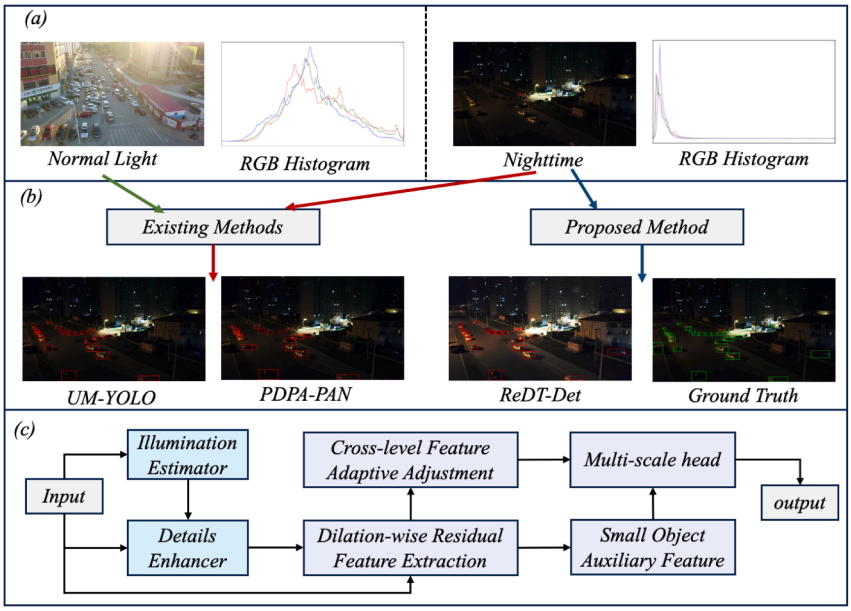

Fig. 1. (a) The image and its red, green and blue channel histogram of normal light and nighttime. (b) Vehicle detection results obtained by UM-YOLO (Zhu et al., 2024), PDPA-PAN (Ying et al., 2024), and the proposed method and the ground truth of the image. (c) The overall flowchart of our ReDT-Det.For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

Table 1
The object scale analysis of typical UAV-based datasets.

samples	small	medium	large	total
VisDrone-DET	292,065 (61.97 %)	155,487 (32.99 %)	23,714 (5.04 %)	471266
UAVDT	493,860 (61.83 %)	291,081 (36.44 %)	13,853 (1.73 %)	798795

Previous research on nighttime image enhancement has predominantly focused on improving human perception (Brateanu et al., 2024; Cai et al., 2023a; Cui et al., 2024; Ma et al., 2022a; Xu et al., 2020), which may not be effective for downstream vision tasks such as instance segmentation, object tracking, and object detection. For example, LYT-Net (Brateanu et al., 2024) optimize illumination adjustment and noise suppression for visual appeal, but fail to preserve low-contrast structural features essential for object boundaries in detection. These methods treat enhancement as an isolated image-to-image translation task rather than a feature-preserving preprocessing stage for high-level vision. This explains their suboptimal performance when integrated into detection pipelines despite strong perceptual metrics. Some studies have explored the integration of image enhancement with object detection or tracking tasks (Qin et al., 2022; Wang et al., 2020b; Ye et al., 2022; Yin et al., 2023; Zhang et al., 2024). For example, DE-Net (Qin et al., 2022) uses a Laplacian pyramid to decompose the input image into low-frequency and high-frequency components, enabling global enhancement and cross-level guidance, and then employs a YOLOv3 detector (Farhadi & Redmon, 2018) for object detection. Similarly, PE-YOLO (Yin et al., 2023) combines a detail processing module and a low-frequency enhancement filter for image enhancement, followed by object detection using a YOLOv3 detector. CPA-Enhancer (Zhang et al., 2024) first introduced a chain-of-thought mechanism to progressively adapt its enhancement strategy under different image degradations, achieving adaptively recognized image degradation-related information. While these methods have shown certain improvements in detection performance, they often involve high computational complexity, making them unsuitable for resource-constrained UAV platforms. Thus, enhancing detection performance during nighttime while maintaining low computational costs remains a significant challenge.

To evaluate scale distribution in UAV-based detection tasks, we analysed two prominent benchmarks, VisDrone-DET (Du et al., 2019) and UAVDT (Du et al., 2018). As shown in Table 1, UAV-captured imagery exhibits a pronounced dominance of small and medium scale objects, with VisDrone containing 61.97 % small, 32.99 % medium, and only 5.04 % large objects, while UAVDT comprises 61.83 % small, 36.44 % medium, and 1.73 % large objects. Additionally, we conducted a statistical analysis of the object scales in the images from the VisDrone and UAVDT datasets. The results, as shown in Table 2, indicate that in the VisDrone dataset, over 93 % of the images contain small and medium scale objects. Meanwhile, in the UAVDT dataset, more than 87 % of the images include objects of these scales. This statistical evidence underscores the critical importance of small and medium scale target detection for UAV-based vehicle systems. Conventional approaches like Feature Pyramid Networks (FPN) (Lin et al., 2017) and its variants (Liu et al., 2018; Tan et al., 2020) aim to address multi-scale detection but suffer from limited cross-level interaction, as they primarily fuse adjacent-level features while neglecting explicit information exchange across non-adjacent layers, leading to suboptimal feature representations for small objects. Advanced frameworks such as GoldYOLO (Wang et al., 2024b), which aligns four-level backbone features, and AFPN-head (Yang et al., 2023), which progressively fuses backbone outputs, improve cross-layer communication but incur prohibitive computational costs (e.g., GoldYOLO requires 62.7 GFLOPs), rendering them impractical for resource-constrained UAV platforms. These limitations highlight the urgent need for lightweight, efficient architectures tailored to UAV-specific scale distributions.

Table 2
The rate of images including different scale objects.

rate of images	small	medium	large
VisDrone-DET	93.4 %	97.8 %	60.6 %
UAVDT	88.4 %	87.3 %	22.6 %

To overcome the dual challenges of high computational complexity and insufficient accuracy for small and medium targets in UAV-based detection during nighttime, we propose ReDT-Det, a Retinex-guided illumination Differential Transformer Detection network. ReDT-Det is designed to enhance nighttime imaging and improve detection performance in real-world UAV-based vehicle detection scenarios. The architecture of ReDT-Det is illustrated in Fig. 1(c). Initially, we apply Retinex theory (Land & McCann, 1971) to estimate lighting information and introduce an Illumination-Fused Differential Transformer Block (IFDTB) to enhance images. This block effectively restores low-light image details while reducing computational overhead. Next, we design a Dilation-wise Residual Cross Stage Partial (DR-CSP) module to expand the receptive field and improve feature extraction capabilities, particularly for small-scale objects. Finally, a Cross-level Feature Adaptive Adjustment (CFAA) module and a Small Object Auxiliary Feature (SOAF) are employed to facilitate progressive cross-level information exchange and enhance feature representation for small objects. Our contributions can be summarized in three key aspects:

- Illumination-fused differential transformer block (IFDTB): A Retinex-based enhancement module that integrates differential attention mechanisms to suppress noise and focus on dark regions. This approach achieves a 15.24 % reduction in computational overhead compared to RetinexFormer, while providing low-light images with greater detail to improve object detection performance.
- Progressive feature fusion architecture: A lightweight framework designed to enhance the detection of small and medium-scale objects. It combines Dilation-wise Residual Cross Stage Partial (DR-CSP) module to expand the receptive field, Cross-level Feature Adaptive Adjustment (CFAA) for multi-scale feature fusion, and Small Object Auxiliary Feature (SOAF) to boost small-scale object detection by 3.6 % mAP.
- NightDrone-Mix benchmark: We introduce a comprehensive benchmark dataset named NightDrone-Mix for real-world nighttime UAV-based vehicle detection. It consists of 15,144 annotated images across diverse urban scenarios. Extensive experiments on NightDrone-Mix, DroneVehicle(Night) and ExDark datasets demonstrate state-of-the-art performance, with ReDT-Det achieving 65.2 % mAP while reducing FLOPs by 4.8G compared to the baseline on NightDrone-Mix.

The remainder of this paper is organized as follows: Following this introduction, Section 2 comprehensively reviews the evolution of object detection methodologies tailored to UAV imagery, emphasizing advancements and persistent gaps in low-light scenarios. Section 3 elaborates on the architectural details of the proposed ReDT-Det framework, including its Retinex-guided illumination enhancement module, progressive feature fusion strategy. Section 4 systematically evaluates the method through benchmark experiments on the NightDrone-Mix and ExDark datasets, providing quantitative comparisons of detection accuracy, computational efficiency, and robustness across varying illumination conditions. Finally, Section 5 synthesizes the key contributions of this work and outlines future research directions.

2. Related work

The rapid advancement of object detection techniques in aerial images has profoundly transformed intelligent urban transportation systems, enabling real-time traffic monitoring, incident response, and infrastructure management. This section systematically examines four critical dimensions of this technological evolution. First, we trace the historical progression of UAV-oriented detection frameworks, analyzing their adaptation from conventional ground-based systems to aerial-view optimized architectures. Subsequently, we critically evaluate state-of-the-art low-light enhancement methodologies, emphasizing their limitations in balancing perceptual quality and computational efficiency for UAV deployment. Then, we explore the development of low-light object detection. Finally, we explore the paradigm shift brought by visual Transformers in object detection, discussing their capabilities and challenges in addressing UAV-specific requirements like multi-scale target recognition and illumination invariance.

2.1. Object detection on UAV images

Technological advances in UAV-based object detection have driven significant progress and expanded its applications. Much of the research has focused on adapting existing detectors, originally designed for natural scenes, to the unique characteristics of drone imagery. For instance, YOLO variants (Ying et al., 2024; Zhu et al., 2024) and Faster R-CNN variants (Cao et al., 2020; Lin et al., 2020; Yang et al., 2019) have been widely studied and optimized for this purpose. PDPA-PAN (Ying et al., 2024) enhances object detection by integrating spatial and channel attention during multi-scale feature fusion. CRPN-SFNet (Yang et al., 2019) introduces a semantic region proposal network to quickly identify relevant regions while filtering out irrelevant ones. UAV imaging often involves varying angles of view, which complicates the use of traditional rectangular bounding boxes for accurate object localization. To address this, several studies (Cheng et al., 2022; Ding et al., 2019; Li et al., 2022c; Xie et al., 2021) have explored oriented bounding boxes for more precise object localization. Oriented R-CNN (Li et al., 2022c) employs an oriented region proposal network to generate high-quality oriented proposals, which are then refined and recognized. Oriented RepPoints (Li et al., 2022c) further improves this by introducing an adaptive point representation that captures the geometric information of arbitrarily oriented instances, enabling effective representation of non-axis-aligned features. Another area of focus has been addressing class and label imbalance during detection model training. DSHNet (Yu et al., 2021) improves tail-class detection performance by introducing dual samplers and heads for long-tail and head classes. A class-aware dynamic label assignment strategy (Feng et al., 2024) ensures consistent training between classification and position regression. While these methods have shown strong performance on datasets like DOTA (Xia et al., 2018), VisDrone (Du et al., 2019), and UAVDT (Du et al., 2018), they are primarily designed for normal lighting conditions and are less effective in nighttime scenarios.

Despite these advancements, existing methods still face significant challenges in nighttime applications due to low signal-to-noise ratios and the prevalence of small and medium scale objects. This highlights the need for robust solutions that can enhance detection performance under low-light conditions while maintaining computational efficiency for deployment on resource-constrained UAV platforms. Thus, we introduce progressive feature adaptation to optimize multi-scale representation learning, with emphasis on small and medium-scale objects.

2.2. Nighttime image enhancement

Night image enhancement aims to improve human visual perception by restoring image details and correcting color distortion. It employs various methods to provide images with enhanced details suitable for complex tasks such as object detection and segmentation. Traditional approaches include gamma correction (Rahman et al., 2016; Wang et al., 2009) and cognition-based methods (Fu et al., 2016; Wang et al., 2013). With the rapid advancement of deep learning, a growing number of methods (Bai et al., 2024; Cai et al., 2023a; Cui et al., 2022; Ma et al., 2023, 2022a; Xing et al., 2023) have adopted CNNs and Transformers for illumination estimation and detail restoration based on Retinex theory (Land & McCann, 1971). Retinexformer (Cai et al., 2023a) introduces a one-stage Retinex-based framework with an illumination-guided transformer for simultaneous brightness adjustment and detail restoration. Retinexmamba (Ma et al., 2023) improves computational efficiency by replacing attention mechanisms with state space models while maintaining Retinex decomposition principles. The Illumination

Adaptive Transformer (Cui et al., 2022) utilizes attention queries to dynamically optimize parameters of the image signal processor, including color correction matrices and gamma values. While these methods have achieved notable improvements in nighttime image enhancement, they often involve high computational complexity, making them unsuitable for resource-constrained UAV platforms.

These advancements in nighttime image enhancement have significantly improved the restoration of image details and visual perception. However, most existing methods are primarily designed to enhance human visual perception and may not be optimized for downstream vision tasks such as object detection. First, many of these methods prioritize enhancement quality over computational efficiency, resulting in high computational complexity. This makes them impractical for deployment on UAV platforms with limited computational resources. Second, these methods are often not specifically designed for object detection, which requires not only visual enhancement but also the preservation of critical features necessary for accurate detection. To address these gaps, we propose a lightweight architecture employing dynamic convolutions and depthwise separable convolutions to replace standard convolutions, alongside a differential transformer for enhancement and efficient task-oriented feature extraction for detection.

2.3. Object detection during nighttime

Object detection during nighttime presents unique challenges due to low illumination, high noise, and uneven lighting. Recent research addresses these through four dominant paradigms.

1) Image enhancement-driven and detection: This paradigm improves input quality via images preprocessing and joint object detectors to boost detection performance in degraded conditions. PE-YOLO (Yin et al., 2023) combines a detail processing module and a low-frequency enhancement filter for image enhancement, followed by object detection using a YOLOv3 detector. DEDet (Xi et al., 2024) jointly optimizes exposure correction and detection via a fine-grained parameter predictor. 2PCNet (Kennerley et al., 2023) employs dual-phase pseudo-label refinement with nighttime-specific augmentation for unsupervised domain adaptation, enhancing small-object detection while suppressing error propagation.

2) Infrared-centric approaches: This paradigm leverages thermal radiation properties to achieve illumination-invariant detection, primarily through specialized network architectures and optimisation strategies. DDCNN (Patel et al., 2022) incorporates novel loss functions to enhance accuracy and achieve real-time application potential with notable mAP and car detection accuracy. Thermal-enhanced YOLOv5 (Vo & Quach, 2023) utilizes hybrid SGD-Adam training and activation function optimization to boost nighttime detection in driver-assistance systems. MSH-Net (Liu et al., 2024) introduces a scale-and-location-sensitive loss and multi-scale head to improve infrared small target detection, dynamically weighting IoU by target size and adding polar-coordinate penalty terms for precise localization.

3) Unified framework for enhancement and detection: This paradigm represents an end-to-end optimized paradigm that jointly learns task-shared representations to overcome error propagation in cascaded pipelines. PIA (Ma et al., 2022b) integrates a parallel architecture with a decomposition-type warm-start and illumination-aware feature allocator to jointly perform coarse-to-fine enhancement and decomposed-to-integrated detection in low-light conditions. Reference (Xue et al., 2022) introduces a cascaded framework with contrastive-alternative learning to jointly optimize low-light image enhancement and semantic perception tasks, enabling task entanglement and mutual guidance via inter-task contrastive mechanisms.

Enhancement-driven detection improves input quality but suffers from task misalignment and error accumulation between separated modules. Infrared-Centric Approaches offer illumination invariance but require expensive hardware and struggle with textureless targets. Unified Frameworks jointly optimize tasks but often incur high computational complexity for UAV deployment. We propose an end-to-end network that hierarchically links image processing and object detection via feature passing, optimized solely by detection losses. This forces enhancement modules to preserve detection-critical features while maintaining UAV-compatible efficiency.

2.4. Vision transformer

The Transformer, originally designed for machine translation tasks, incorporates a multi-head self-attention mechanism and a feedforward network. Its ability to capture long-range dependencies has effectively addressed the limitations of CNNs, driving its expanding applications in high-level vision tasks such as semantic segmentation (Wu et al., 2021; Zheng et al., 2021) and object detection (Cai et al., 2023b; Carion et al., 2020; Liu et al., 2022; Zhang et al., 2022b; Zhao et al., 2024). For instance, DETR (Carion et al., 2020) and its variants redefine object detection as a set prediction problem, moving away from traditional anchor-based designs. This approach efficiently manages duplicate predictions and eliminates the need for Non-Maximum Suppression (NMS) during post-processing. However, DETR-like methods require substantial data for training and exhibit weaker performance in detecting small-scale objects, which restricts their use in UAV-based object detection tasks. Beyond high-level tasks, Transformers have also shown remarkable performance in low-level vision tasks, including image restoration (Cai et al., 2022; Zamir et al., 2022) and image synthesis (Jiang et al., 2021; Zhang et al., 2022a). One example is the Retinexformer (Cai et al., 2023a), which applies Retinex theory for brightness estimation and introduces illumination-guided multi-head self-attention for restoring image details. Despite its contributions, the Retinexformer does not account for noise from self-attention calculations and may allocate attention to irrelevant information. Meanwhile, the DIFF Transformer (Ye et al., 2024) introduces a differential attention mechanism in language modelling, using two distinct softmax attention maps to calculate attention scores, thereby promoting the emergence of sparse attention patterns. Drawing inspiration from these studies, we have designed a Retinex-guided illumination-fused differential transformer for nighttime image enhancement, achieving image detail restoration through the calculation of differential attention in image information.

3. Proposed method

To enable robust vehicle detection in nighttime scenarios on UAV-based images, we propose ReDT-Det, an integrated framework combining nighttime image enhancement with efficient object detection. As illustrated in Fig. 2, our architecture operates through two sequential stages: 1) Retinex-guided illumination-fused differential Transformer for image enhancement (IFDT-Enhancer): The input nighttime image undergoes Retinex-based decomposition through a dedicated illumination estimator and detail enhancer. We develop an Illumination-Fused Differential Transformer Block (IFDTB) that integrates frequency-domain differential operators with illumination guidance to recover texture details while suppressing noise. 2) Progressive feature learning for nighttime UAV-based vehicle detection (NightDrone-YOLO): The enhanced image is processed by our Dilation-wise Residual Cross Stage Partial (DR-CSP) module, which employs hybrid dilated convolutions for multi-receptive-field feature extraction. Cross-level Feature Adaptive Adjustment (CFAA) dynamically fuses multi-scale features through spatial-channel attention, while the Small Object Auxiliary Feature (SOAF) component enhances hard-to-detect targets through dedicated high-resolution feature preservation.

3.1. IFDT-enhancer For low-light image recovery

Our image enhancement method, grounded in Retinex theory (Land & McCann, 1971), comprises two key components: the Illumination Estimator IE and the Details Enhancer DE. In IE stage, two 1×1

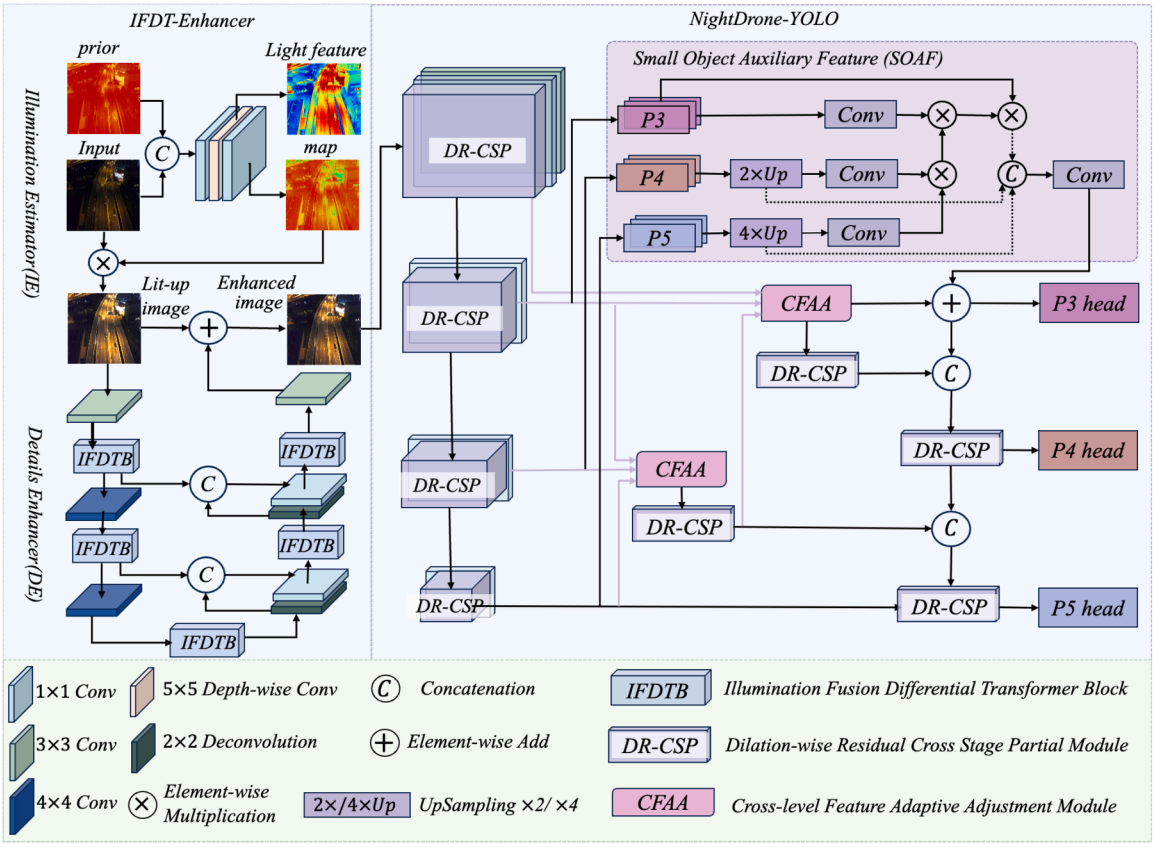

Fig. 2. Architecture of ReDT-Det. The light blue section represents the nighttime image enhancement of ReDT-Det, the light purple section corresponds to the object detection, and the light green section provides the legend explanation. The nighttime image enhancement is composed of illumination estimator (IE) and details enhancer (DE), while the details enhancer includes Illumination-Fused Differential Transformer Block (IFDTB) and a U-shape encoder and decoder. NightDrone-YOLO includes Dilation-wise Residual Cross Stage Partial (DR-CSP) module for feature extraction, Cross-level Feature Adaptive Adjustment (CFAA) module and the Object Auxiliary Feature (SOAF) for progressive multi-scale feature fusion.For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

convolutions and one 5×5 depth-wise convolution are employed to generate the light feature and illuminance map. DE is designed as a U-shaped network, utilizing 4×4 convolution for downsampling and 2×2 deconvolution for upsampling. The Illumination-Fused Differential Transformer Block (IFDTB) is used for illuminance feature extraction, while skip connections are implemented to prevent information loss.

3.1.1. Retinex-guided framework

Retinex theory is widly used on low-light image enhancement networks. Traditional methods decompose an image $I \in \mathbb{R}^{H\times W\times 3}$ into two components:

$$I = L \otimes R \tag{1}$$

where $L \in \mathbb{R}^{H\times W}$ denotes he illumination map, $R \in \mathbb{R}^{H\times W\times 3}$ represents the reflectance component, and $\otimes$ donates element-wise multiplication. While this formulation assumes noise-free imaging conditions, real-world low-light images suffer from two key degradations: 1) Noise Amplification: Sensor noise $\tilde{R}$ and illumination estimation errors $\tilde{L}$ escalate during enhancement; 2) Non-linear Artifacts: Cross-terms between perturbations create irreversible information loss (Cai et al., 2023a). To address these issues, we propose a degradation-aware Retinex model:

$$I = (R + \tilde{R}) \otimes (L + \tilde{L}) \tag{2}$$

Expanding this formulation reveals four distinct components:

$$\begin{aligned} I = & \underbrace{R \otimes L}_{\text{Ideal Component}} + \underbrace{R \otimes \tilde{L}}_{\text{Illumination Error}} + \underbrace{\tilde{R} \otimes L}_{\text{Reflectance Noise}} + \underbrace{\tilde{R} \otimes \tilde{L}}_{\text{Cross Corruption}} \\ & = R \otimes L + R \otimes \tilde{L} + \tilde{R} \otimes (L + \tilde{L}) \end{aligned} \tag{3}$$

where $\tilde{R} \in \mathbb{R}^{H\times W\times 3}$ and $\tilde{L} \in \mathbb{R}^{H\times W}$ model reflectance noise and illumination perturbations, respectively. We introduce a light-up map $\bar{L}$ satisfying $L \otimes \bar{L} = 1$ applied via element-wise multiplication:

$$I \otimes \bar{L} = R + R \otimes (\tilde{L} \otimes \bar{L}) + (\tilde{R} \otimes (L + \tilde{L})) \otimes \bar{L}, \tag{4}$$

where the term $\tilde{R} \otimes (L + \tilde{L})$ captures the inherent noise and compression artifacts latent in low-light conditions, which undergoes amplification through the light-up mapping $\bar{L}$. Meanwhile, $R \otimes (\tilde{L} \otimes \bar{L})$ characterizes the non-linear interactions leading to exposure imbalance (under/over-exposure) and chromatic aberration, both introduced during the illumination correction phase. After simplification, we can express the illuminated image I_{lu} as follows:

$$I_{\text{lu}} = I \otimes \bar{L} = R + C, \tag{5}$$

where $I_{\text{lu}} \in \mathbb{R}^{H\times W\times 3}$ notes the lit-up image and $C \in \mathbb{R}^{H\times W\times 3}$ represents the information regarding overall damage in the image. Therefore, the process of our image enhancement can be expressed as:

$$\left(I_{\text{lu}}, F_{\text{lu}}\right) = \text{IE}\left(I, L_{\text{p}}\right) \tag{6}$$

$$I_{\text{en}} = \text{DE}\left(I_{\text{lu}}, F_{\text{lu}}\right) \tag{7}$$

$$I_{\text{final}} = \text{IE}\left(I_{\text{lu}}, L_{\text{p}}\right) \oplus \text{DE}\left(I_{\text{lu}}, F_{\text{lu}}\right) \tag{8}$$

$$L_{\text{p}} = \frac{1}{3}\sum_{c=1}^{3} I^{(c)} \tag{9}$$

where IE donates Illumination Estimator and DE donates Details Enhancer. The inputs IE are the RGB images $I \in \mathbb{R}^{H\times W\times 3}$ and their prior $L_{\text{p}} \in \mathbb{R}^{H\times W}$, which is derived by calculating the mean value of each

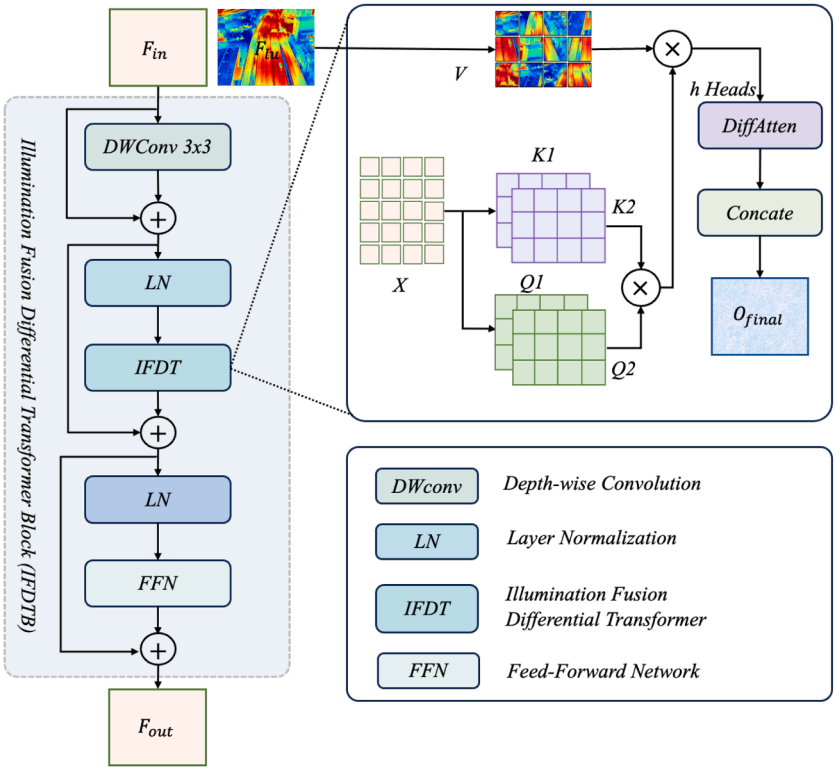

Fig. 3. The overview of the illumination-fused differential transformer block (IFDTB).

pixel of input I across all channels. This value is utilized to evaluate the overall brightness or illumination level of the image. I_{lu} and F_{lu} represent lit-up image and illumination map independently. Then I_{lu} and F_{lu} are fed into DE for details enhancement. DE is a U-shaped network that incorporates IFDTB and skip connections to facilitate corruption restoration. I_{final} means the final enhanced images.

3.1.2. *Illumination-fused differential transformer block (IFDTB)*

To address the limitations of conventional attention mechanisms in low-light enhancement, specifically noise amplification and attention dispersion, we propose a Retinex-guided differential attention mechanism, called Illumination-Fused Differential Transformer Block (IFDTB). After obtained I_{lu} and F_{lu}, they are transported to DE for corruption restoration. IFDTB is the core of DE . As depicted in Fig. 3, IFDTB includes a 3×3 depth-wise convolution, Layer Normanization (LN), Illumination-Fused Differential Transformer (IFDT), Feed-Forward Network (FFN) and skip connection.

As illustrated in Fig. 1, the illumination-corrected feature I_{lu} is processed by a 3×3 convolution to generate the base feature F_{in}, while the illumination feature F_{lu} is incorporated as a physical prior. F_{in} is then reshaped into tokenized features $X \in \mathbb{R}^{HW \times C}$ for multi-head self-attention score computation. Unlike RetinexFormer (Cai et al., 2023a), which projects X to query (Q) , key (K) and (V) vectors and further fuses F_{lu} through token-wise multiplication with V, Our method directly leverages X and F_{lu} constructs multi-head differential attention score, effectively suppressing noise propagation while preserving computational efficiency. To address the limitations of traditional Transformers tendency to allocate excessive attention to irrelevant regions, we decompose X and F_{lu} into k parallel heads:

$$X = \left[X_1, X_2, \ldots, X_k\right], \quad F_{lu} = \left[F_{lu1}, F_{lu2}, \ldots, F_{luk}\right] \tag{10}$$

where $X_i \in \mathbb{R}^{HW \times d_k}, d_k = \frac{C}{k}$, and $i = 1, 2, \cdots, k$. k is the number of the heads and C is the channel of the input feature.

We note that using the light feature F_{lu} as the value vector V for final multiplication allows the model to concentrate more effectively on the darker areas of the image. Hence, we project the input X to the $Q_1, Q_2, K_1, K_2 \in \mathbb{R}^{HW \times d_k}$ and the F_{lu} to $V \in \mathbb{R}^{HW \times d_k}$:

$$\left[Q_1; Q_2\right] = XW^Q, \quad \left[K_1; K_2\right] = XW^K, \quad V = F_{lu}W^V \tag{11}$$

where W^Q, W^K and W^V are learnable parameters. Thus the differential attention operator DiffAttn is calculated as follows:

$$S_{Diff} = S\left(\frac{Q_1 K_1^T}{\alpha}\right) - \lambda S\left(\frac{Q_2 K_2^T}{\alpha}\right) \tag{12}$$

where S denotes $Softmax$ function, T represents matrix transpose oprater. α denotes the learnable parameter that acts as a scaling factor for adjusting the attention scores. $\otimes$ donates element-wise multiplication. The value of λ is computed as follows:

$$\lambda = \exp\left(\lambda_{\mathbf{q}_1} \cdot \lambda_{\mathbf{k}_1}\right) - \exp\left(\lambda_{\mathbf{q}_2} \cdot \lambda_{\mathbf{k}_2}\right) + \lambda_{\text{init}} \tag{13}$$

where $\lambda_{\mathbf{q}_1}$, $\lambda_{\mathbf{k}_1}$, $\lambda_{\mathbf{q}_2}$ and $\lambda_{\mathbf{q}_2}$ are learnable vectors that regulate the sharpness of the attention weights. The term $\lambda_{\text{init}} \in (0, 1)$ serves as the initial value for λ. In our experiments, λ_{init} is empirically set to $0.8 - 0.6 \times \exp(-0.3 \cdot (l-1))$, where $l \in [l, L]$ denotes the layer index. This dynamic initialization strategy helps balance the attention weights and reduces noise. We also explored using a fixed λ_{init} (e.g., 0.6) across all layers, and the performance remains robust to different initialization strategies, as demonstrated in the ablation studies. This design effectively minimizes attention noise and enhances the model's efficiency.

Therefore, the final output of IFDT can be represented as:

$$\begin{aligned} O_i &= S_{Diff} \times V \\ O_{final} &= Concat(O_i) \end{aligned} \tag{14}$$

where $O_i \in \mathbb{R}^{HW \times d_k}$ means the output of per head and $O_{final} \in \mathbb{R}^{HW \times C}$ presents the final output. IFDTB reduces computational overhead by using X and F_{lu} directly for multi-head attention, enhancing focus on dark regions with F_{lu} as the value vector. Computing the differential attention scores of input X can reduce the attention noise and allocate more attention to relevant areas. IFDTB allows for a greater emphasis on areas

that require enhancement while simultaneously reducing the model's computational complexity. This results in images that are better suited for downstream object detection by effectively highlighting relevant regions without unnecessary computational overhead.

3.2. Nightdrone-YOLO for UAV-based nighttime vehicle detection

After obtaining the illumination-enhanced image, we feed it into the object detection network for further processing. The YOLO series of detectors is widely used in real-world applications, such as urban intelligent transportation systems, due to its balance between speed and accuracy. However, it struggles with vehicle detection in drone imagery, particularly under nighttime conditions. To overcome these limitations, we introduce enhancements to the YOLOv11 framework (Jocher et al., 2023), specifically optimizing it for nighttime UAV-based vehicle detection through progressive feature extraction and fusion.

First, we incorporate the Dilation-wise Residual Cross Stage Partial (DR-CSP) Module in the backbone feature extraction phase. This module expands the network's receptive field, enabling more effective contextual information capture while minimizing information loss. Second, we design the Cross-level Feature Adaptive Adjustment (CFAA) module to enhance cross-level feature fusion, facilitating better information exchange across different feature hierarchies and improving the representation of multi-scale objects. Finally, we introduce the Small Object Auxiliary Feature (SOAF) module to specifically address the challenge of detecting small-scale objects. This module strengthens the feature representation of small objects, significantly improving detection accuracy. With these advancements, our improved object detection network, tailored for nighttime UAV-based vehicle detection, is named NightDrone-YOLO.

3.2.1. Dilation-wise residual cross stage partial (DR-CSP) module

CSPNet (Wang et al., 2020a) introduces an innovative architecture that enhances gradient flow by splitting gradients across different network paths, effectively integrating feature maps from various stages. This approach reduces computational complexity while preserving accuracy, making it highly efficient for resource-constrained environments. Due to its advantages, CSPNet has been widely adopted in numerous object detectors (Qin et al., 2022; Ying et al., 2024; Zhu et al., 2024). However, UAV-captured images often contain a large number of small-scale objects with limited distinguishable features. Directly applying the CSPNet architecture may not fully capture and utilize this critical information. Moreover, DWRSeg (Wei et al., 2022) suggests that multi-scale contextual information can be more effectively extracted through a two-step process of region residualization and semantic residualization. Inspired by this, we propose the Dilation-wise Residual Cross Stage Partial (DR-CSP) module at the backbone stage. The DR-CSP module is designed to replace the C3k2 block in YOLOv11, enhancing the model's receptive field for improved small-object representation. As illustrated in Fig. 5(a), the input feature $X \in \mathbb{R}^{H\times W\times C}$ is divided into two parts:

$$X_1, X_2 = \text{split}(X) \tag{15}$$

where $X_1, X_2 \in \mathbb{R}^{H\times W\times C/2}$, X_1 is directly used for the final feature concatenation, while X_2 undergoes dilation-wise residualization. Three parallel dilated depth-wise convolutions extract multi-scale features:

$$Y_k = DWConv_{d_k}(X_2) \tag{16}$$

where $k \in \{1,3,5\}$, $DWConv_{d_k}$ denotes depth-wise convolution with dilation rate d_k. These outputs are then concatenated along the channel dimension, followed by a 1×1 convolution to refine and integrate channel information:

$$Y' = Conv_{1\times 1}\big(\text{Concat}(Y_1, Y_2, Y_3)\big) \tag{17}$$

Then, the output Y' is added to the X_2 and passed through a Batch Normalization (BN) layer followed by a Gaussian Error Linear Unit (GELU) activation function, as shown in the following equation:

$$Y = \text{BN}\big(\text{GELU}(Y' + X_2)\big) \tag{18}$$

Finally, the other part X_1 is concatenated with the output Y:

$$Z = \text{Concat}\big(X_1, Y\big) \tag{19}$$

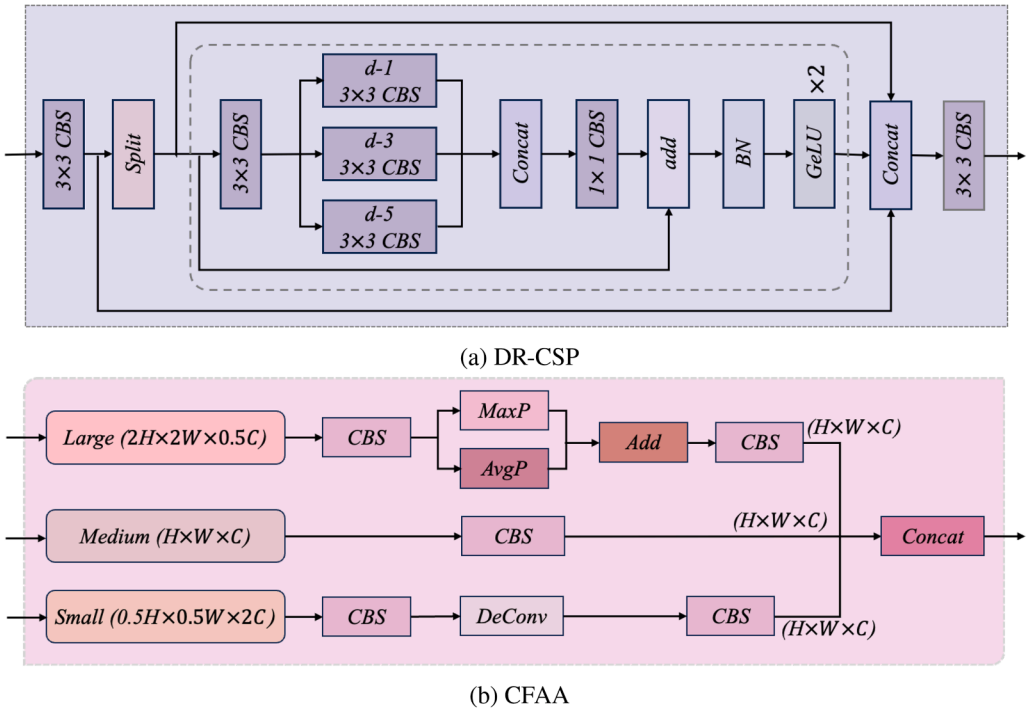

Fig. 4. The detail of Dilation-wise Residual Cross Stage Partial (DR-CSP) and Cross-level Feature Adaptive Adjustment (CFAA) module. (a) DR-CSP. (b) CFAA. CBS means Convolution + BatchNormal(BN) + SiLU. Add means element-wise addition. $d-i, (i=1,3,5)$ means the dilated rate of convolution. $MaxP$ and $AvgP$ are represent the max pooling and average pooling independently. $DeConv$ is the deconvolution. ⊗ means element-wise multiplication.

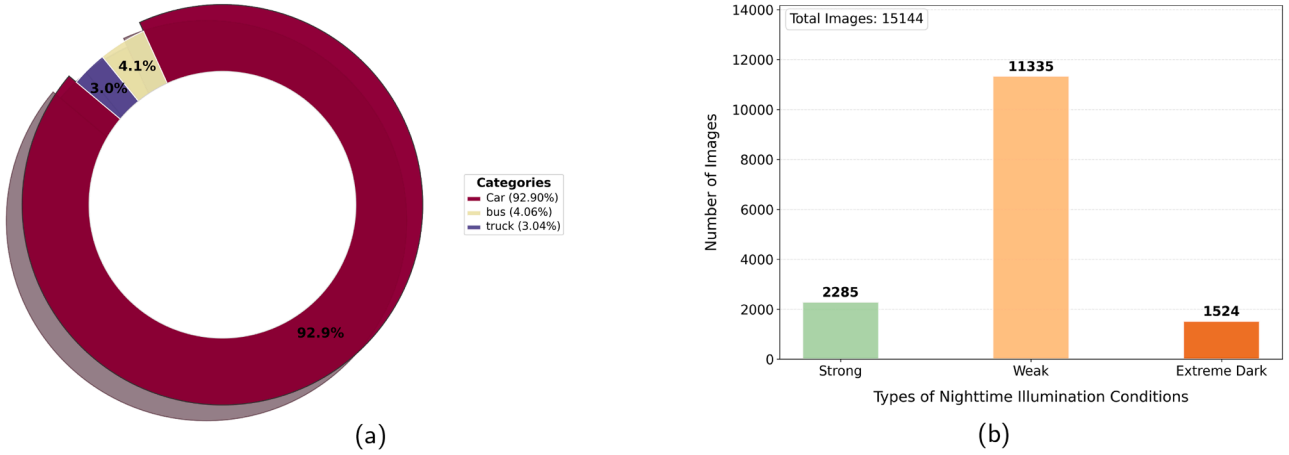

Fig. 5. Analysis of NightDrone-Mix on category and illumination. (a) Category analysis; (b) Illumination analysis.

The final representation Z integrates multi-scale contextual features through parallel dilated convolutions and residual learning.

The DR-CSP module significantly enhances small-object detection in UAV imagery by expanding receptive fields through dilated depth-wise convolutions to capture multi-scale morphological features, while maintaining robust gradient flow via identity shortcut connections and enriching semantics through residual feature fusion.

3.2.2. Cross-level feature adaptive adjustment (CFAA)

Traditional YOLO-series detectors rely on PAFPN (Liu et al., 2018) to enhance feature fusion through bottom-up path augmentation. However, PAFPN is limited to facilitating information exchange between adjacent feature levels, which restricts direct communication across different layers and limits the effective representation of multi-scale features. ASF-YOLO (Kang et al., 2024) attempts to address this by encoding features from the backbone through simple downsampling or upsampling, but this approach remains insufficient for nighttime UAV-based detection tasks. To overcome these limitations, we propose Cross-level Feature Adaptive Adjustment (CFAA), an innovative mechanism designed for efficient cross-level feature exchange. Unlike previous methods that rely on local feature interactions, CFAA leverages global attention to enable direct communication across different feature levels. As shown in Fig. 4(b), large scale feature $F_L \in \mathbb{R}^{2H\times 2W\times 0.5C}$ comes from $P3$, medium feature $F_M \in \mathbb{R}^{H\times W\times C}$ from $P4$, and small scale feature $F_S \in \mathbb{R}^{0.5H\times 0.5W\times 2C}$ from $P5$, respectively. Small-scale features are upsampled to the medium scale via deconvolution:

$$F_s' = DeConv(CBS(F_s)) \tag{20}$$

where $F_S' \in \mathbb{R}^{H\times W\times 2C}$. CBS stands for $Convolution + BN + SiLU$. And large-scale features are downsampled to the medium scale using a combination of max pooling and average pooling:

$$F_L' = \mathrm{Add}\big(\mathrm{MaxPool}(CBS(F_L)), \mathrm{AvgPool}(CBS(F_L))\big) \tag{21}$$

where $F_L' \in \mathbb{R}^{H\times W\times 0.5C}$. Attention weights are computed via 1×1 $Conv$ and applied element-wise to each feature.

Finally, the features from the three branches are concatenated along the channel dimension:

$$F = Concat(CBS(F_S'), CBS(F_M), CBS(F_L')) \tag{22}$$

This design enables comprehensive cross-layer information interaction, improving the representation of features at different scales and significantly boosting detection accuracy, especially in challenging multi-scale scenarios.

3.2.3. Small object auxiliary feature (SOAF)

In most commonly used UAV-based datasets, over 61 % of detected objects are small-scale, making their accurate detection critical for overall performance. To specifically enhance the detection capability for small objects, we propose the Small Object Auxiliary Feature (SOAF) module, designed to refine and strengthen small-object representation. The detailed architecture of SOAF is illustrated in Fig. 3. SOAF takes multi-scale features $P3 \in \mathbb{R}^{H\times W\times C_3}$, $P4 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C_4}$, and $P5 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_5}$ from the backbone. To capture fine-grained details, $P5$ and $P4$ are upsampled to match the resolution of $P3$. However, rather than directly fusing these features, SOAF selectively refines and integrates information from $P4$ and $P5$, ensuring that only the most relevant small-object details contribute to the final feature representation.

$$\widetilde{P5} = Conv(\mathrm{Up}_{\times 4}(P5)), \widetilde{P4} = Conv(\mathrm{Up}_{\times 2}(P4)), \widetilde{P3} = Conv(P3) \tag{23}$$

where $\mathrm{Up}_{\times k}$ denotes $k\times$ upsampling via nearest-neighbor interpolation. This alignment enables unified feature processing while preserving small-object details.

Instead of direct fusion, SOAF performs cross-scale feature modulation to emphasize small-object characteristics. First, multiplicative interaction highlights co-activated patterns:

$$P' = \widetilde{P5} \odot \widetilde{P4} \odot \widetilde{P3} \tag{24}$$

where $\odot$ denotes element-wise multiplication. This operation amplifies features consistently activated across scales - crucial for small objects. The modulated features are then recursively refined with the high-resolution $\widetilde{P3}$ stream.

$$P'' = P3 \odot P' \tag{25}$$

This step further enhances spatial precision by preserving high-frequency details from shallow layers. The final representation combines original and modulated features:

$$P_{SOAF} = Conv(Concat(\widetilde{P5}, \widetilde{P4}, P'')) \tag{26}$$

Concatenation retains multi-scale context while convolution adaptively reweights channels. P_{SOAF} is combined with the CFAA-enhanced **P3** representation, yielding a discriminative feature map optimized for small-scale UAV object detection. By employing adaptive feature refinement and targeted fusion, SOAF significantly enhances small-object visibility, leading to improved classification and localization accuracy in UAV-based detection scenarios, particularly under challenging conditions.

4. Experiments

4.1. Datasets

Most existing UAV object detection datasets contain limited nighttime samples, resulting in a lack of a dedicated benchmark for systematically evaluating the performance of different approaches in nighttime

Table 3
The data resources of NightDrone-Mixdataset.

Data source	train	validation	test	total
VisDrone-DET (Du et al., 2019)	1695	346	371	2412
UAVDT (Du et al., 2018)	8905	1925	1902	12732

Table 4
The object scale analysis of NightDrone-Mixdataset.

small	medium	large	total
117,993 (53.38 %)	94,114 (42.57 %)	8956 (4.05 %)	221063

UAV object detection. The NightDrone (Xi et al., 2024) dataset, captured using a DJI MINI 2 drone for nighttime drone object detection, is not publicly available. To address this, we introduce NightDrone-Mix, a specialized dataset for UAV-based vehicle detection at night. NightDrone-Mix comprises low-light images curated from the VisDrone-DET (Du et al., 2019) and UAVDT (Du et al., 2018) datasets. Therefore, NightDrone-Mix is a cross-dataset dataset specifically designed for nighttime UAV-based vehicle detection.

NightDrone-mix. Since VisDrone-DET and UAVDT datasets contain different object categories, we selected nighttime images that share common vehicle classes, specifically Car, Truck, and Bus. This process resulted in a collection of 15,144 nighttime images, comprising 221,063 instances. Table 3 provides a breakdown of the dataset, detailing the distribution of samples across training, validation, and testing sets. We conducted an analysis of object sizes within NightDrone-Mix, as presented in Table 4. The dataset is dominated by small-scale objects, which constitute 53.38 % of all annotations, followed by medium-scale objects at 42.57 %, and large-scale objects at only 4.05 %. This distribution highlights the challenge of detecting small objects in UAV imagery, reinforcing the need for advanced detection techniques. We performed a statistical analysis of the instance categories in the dataset, with the results shown in Fig. 5(a). The distribution is as follows: Car account for 92.9 %, Bus for 4.06 %, and Truck for 3.04 %. Additionally, following the methodology of the NightDrone dataset Xi et al. (2024), we conducted statistical analyses of the lighting conditions, categorizing them into Strong light, weak light, and extreme dark. The results are presented in Fig. 5(b). The majority of images (11,335) fall under low-light conditions, while bright conditions comprise 2285 images, and extremely dark scenes account for 1524 images.

To ensure consistency and usability, we standardized the annotations from both datasets into a unified COCO format, facilitating seamless model training and evaluation. The image resolutions in NightDrone-Mix vary, with the majority measuring approximately $1,360 \times 750$, 960×540, or $1,080 \times 540$ pixels. By consolidating and refining nighttime UAV imagery, NightDrone-Mix serves as a crucial benchmark dataset, enabling the development and evaluation of robust object detection models tailored for UAV-based nighttime scenarios.

DroneVehicle (Night): DroneVehicle is a multimodal aerial dataset featuring paired RGB and infrared images captured from drone perspectives, including substantial nighttime scenes annotated with five vehicle categories: Car, Bus, Truck, Van, and Freight Car. To validate our method's cross-dataset performance, we constructed the DroneVehicle (Night) subset by selecting nighttime RGB images from the original collection based on luminance thresholds and dark pixel ratios. This specialized subset maintains the native 840×712 pixel resolution and contains 9562 training images, 4956 testing images, and 729 validation images. We performed comprehensive label standardization to rectify annotation inconsistencies, including mapping truvk to Truck and consolidating variants like feright car and feright_car into standardized Freight Car. The category analysis is shown in Fig. 6. We convert all annotations into both COCO and YOLO formats to facilitate model training pipelines. The category taxonomy remains consistent with the original DroneVehicle benchmark as established in prior work.

ExDark. To validate the generalization capability of our methods, we conducted experiments on the nighttime dataset ExDark (Loh & Chan, 2019), a benchmark specifically designed for low-light image enhancement and object detection. ExDark consists of 7363 images captured under ten different low-light conditions, ranging from extremely dark environments to twilight scenarios. The dataset provides meticulously annotated images with twelve object categories, including both image-level class labels and localized bounding boxes. Additionally, ExDark

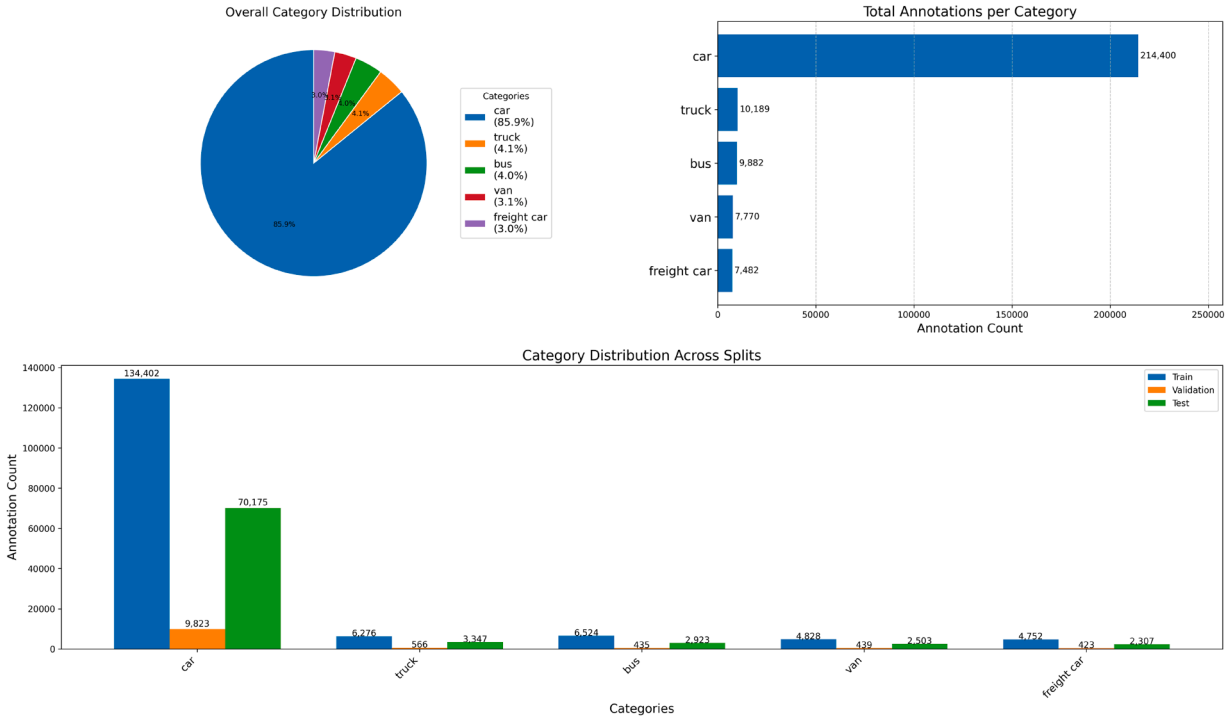

Fig. 6. The category analysis of DroneVehicle(Night), including train, validation and test datasets.

Table 5
Parameters for model training.

Training Parameters	Setting
Loss Function	Focal, Distribution Focal and CIoU loss
Loss weight	0.5 cls, 1.5 dfl, 7.5 box
Image resolution	640×640
Number of epochs	500
Batch size	8
Optimizer	SGD
initial learning rate	0.01
Learning rate decay	0.0005
Momentum	0.937
Data Preprocessing	Mosaic & Mixup

features images of varying resolutions, further increasing the complexity of the detection task. For our experiments, we split the dataset into training, validation, and testing sets using an 8:1:1 ratio, ensuring a balanced evaluation of model performance across different lighting conditions.

4.2. Experimental setting

Experimental details. Our ReDT-Det is implemented using Ultralytics (Jocher et al., 2023), with all comparable methods retrained on NightDrone-Mixusing either Ultralytics or MMDetection (Chen et al., 2019) for a fair comparison. All experiments are conducted on a single RTX 3090 GPU. The details of experimental setting are shown on Table 5. By default, the input image resolution is set to 640×640 during both training and testing. The classification loss is computed using Focal Loss, while box regression loss incorporates Distribution Focal Loss (DFL) and Complete IoU (CIoU) Loss for more precise localization. For training, we set the batch size to 8 and use Stochastic Gradient Descent (SGD) as the optimizer. The initial learning rate is configured to 0.01, with a weight decay of 0.0005. Additionally, the IoU threshold for Non-Maximum Suppression (NMS) is set to 0.7, ensuring optimal filtering of redundant detections.

Evaluation metrics. To validate the performance of UAV-based nighttime object detection, we employ several metrics similar to those in the COCO dataset (Lin et al., 2014). The overall detection accuracy is assessed using mean average precision (mAP), AP_{50}, and AP_{75}. The precision for detecting small, medium, and large-scale objects is measured by AP_s, AP_m , and AP_l, respectively. To evaluate the model's lightness from multiple perspectives, we utilize parameter size $Params(M)$ and $FLOPs(G)$ to estimate computational complexity, while inference time measured in milliseconds ms indicates the speed and efficiency in real-time applications.

4.3. Comparison experiments

4.3.1. Comparison experiments with start-of-the-art object detector

We conducted a comprehensive evaluation of state-of-the-art object detection methods on the NightDrone-Mix dataset and DroneVehicle(Night). Tables 6 and 7 present the performance of various approaches on the test set, including two-stage detectors such as Faster R-CNN (Ren et al., 2016) and one-stage methods like ATSS (Zhang et al., 2020), GFL (Li et al., 2020), and RetinaNet (Lin, 2017). Additionally, we assessed DETR-like architectures, including DiNO (Zhang et al., 2022b), RT-DETR (Zhao et al., 2024), and AlignDETR (Cai et al., 2023b). Meanwhile, All these methods were implemented using MMDetection, while YOLO-series models were retrained on NightDrone-YOLO using Ultralytics. Concurrently, we benchmarked recent drone-based object detection methods, including CEASC (Du et al., 2023), UM-YOLO (Zhu et al., 2024), PDPA-PAN (Ying et al., 2024), HiCAL (Zhang et al., 2025), and CFIA (Bi et al., 2025). Among these, CEASC and HiCAL were trained using officially released source code. The other three methods were reimplemented based on their respective publications or provided core code repositories. The best and second-best results are indicated by single underline and double underlines, respectively.

NightDrone-Mix. From Table 6, several key observations can be made: DETR-like methods demonstrate strong performance in detecting large objects but struggle significantly with small object detection. Their reliance on global attention mechanisms makes them less effective in capturing fine-grained details of small targets, leading to overall subpar performance on NightDrone-Mix. Traditional one-stage and two-stage detectors fail to match the efficiency of YOLO-based methods in UAV-based nighttime detection. The two-stage Faster R-CNN suffers from slower inference speeds and limited adaptability to UAV-acquired images with varying resolutions and illumination conditions. When comparing specialized UAV detection methods, CEASC has the lowest accuracy (26.8 % mAP) and high computational demands (41.3M Params/57.7G FLOPs). In contrast, UM-YOLO offers competitive accuracy (62.6 % mAP) with efficient computation (8.1M Params/19.8G FLOPs). YOLOv11 shows stronger performance, especially in detection accuracy and small object recognition, making it suitable for

Table 6
The detection precision (%) of various detectors on NightDrone-Mix testing dataset.

Method	mAP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	$Params(M)$	$FLOPs(G)$
Faster R-CNN (Ren et al., 2016)	47.8	65.0	56.2	41.7	55.8	33.3	41.358	90.908
ATSS (Zhang et al., 2020)	57.1	71.1	64.1	52.3	63.3	41.8	32.118	80.515
GFL (Li et al., 2020)	23.2	56.9	12.8	13.6	31.1	23.6	32.263	81.752
RetinaNet (Lin, 2017)	42.3	61.1	47.9	32.9	51.3	22.1	36.275	81.095
DETR (Carion et al., 2020)	30.8	65.2	22.7	19.5	37.0	48.6	41.555	38.815
DiNO (Zhang et al., 2022b)	40.1	75.3	37.6	28.2	49.0	55.2	47.544	121.856
RT-DETR-l (Zhao et al., 2024)	42.7	69.3	47.8	28.3	51.6	55.8	31.989	103.402
AlignDETR (Cai et al., 2023b)	44.5	84.1	36.3	36.8	48.1	76.3	47.492	121.884
GoldYOLO (Wang et al., 2024b)	62.6	80.3	72.0	50.0	69.0	73.0	29.973	62.716
YOLOv5 (Jocher et al., 2022)	61.7	80.0	71.1	49.9	67.9	70.1	9.112	23.813
YOLOv6 (Li et al., 2022b)	57.9	78.0	68.7	47.9	66.2	68.9	16.312	44.051
YOLOv8 (Jocher et al., 2023)	62.2	80.2	72.0	50.6	68.6	70.4	11.132	28.612
YOLOv9 (Wang et al., 2024c)	61.3	78.6	70.0	48.6	67.6	70.9	6.191	22.103
YOLOv10 (Wang et al., 2024a)	61.7	79.6	71.1	50.5	67.9	69.2	8.041	24.511
YOLOv11 (Jocher et al., 2023)	62.3	80.4	71.6	50.5	68.6	70.5	9.412	21.306
CEASC (Du et al., 2023)	26.8	49.4	26.7	15.4	35.8	36.1	41.341	57.67
UM-YOLO (Zhu et al., 2024)	62.6	79.8	71.5	52.5	68.7	69.8	8.079	19.801
PDPA-PAN (Ying et al., 2024)	39.2	75.2	36.5	30.5	25.8	48.4	31.941	223.352
HiCAL (Zhang et al., 2025)	63.2	81.5	72.4	49.4	68.9	73.6	51.004	238.883
CFIA (Bi et al., 2025)	63.4	80.7	73.1	53.5	68.9	68.4	15.814	110.052
NightDrone-YOLO	63.8	81.1	72.6	53.2	70.1	73.4	21.398	24.065
ReDT-Det	65.2	83.4	74.8	54.1	71.6	75.2	21.453	32.706

Table 7
The detection precision (%) of various detectors on DroneVehicle(Night) testing dataset.

Method	mAP	AP_{50}	$mAP@car$	$mAP@bus$	$mAP@truck$	$mAP@van$	$mAP@Freightcar$
Faster R-CNN (Ren et al., 2016)	33.9	60.1	47.4	54.3	22.6	23.3	22.0
ATSS (Zhang et al., 2020)	33.2	58.5	48.5	53.4	19.8	22.1	21.9
GFL (Li et al., 2020)	39.0	66.4	50.9	58.7	28.8	30.0	27.2
RetinaNet (Lin, 2017)	26.3	50.1	48.1	45.2	13.0	15.8	10.2
DETR (Carion et al., 2020)	31.0	57.4	40.2	53.4	20.7	20.6	20.4
DiNO (Zhang et al., 2022b)	42.9	70.4	53.3	62.0	34.6	34.1	31.1
RT-DETR-l (Zhao et al., 2024)	39.1	62.6	46.5	54.1	24.8	25.4	24.1
AlignDETR (Cai et al., 2023b)	44.6	70.7	54.7	64.3	40.2	42.1	30.2
YOLOv5 (Jocher et al., 2022)	44.4	69.6	57.5	65.0	34.8	33.4	31.1
YOLOv8 (Jocher et al., 2023)	45.2	70.2	58.0	65.6	37.3	33.3	31.8
YOLOv11 (Jocher et al., 2023)	43.5	69.2	56.8	63.7	33.0	32.8	30.9
CEASC (Du et al., 2023)	39.8	59.1	52.4	50.3	29.1	30.5	28.4
UM-YOLO (Zhu et al., 2024)	44.0	69.2	57.7	64.4	34.6	32.3	31.1
PDPA-PAN (Ying et al., 2024)	37.0	62.9	50.6	57.1	25.9	28.0	24.1
HiCAL (Zhang et al., 2025)	44.2	69.8	57.8	62.1	39.8	38.1	29.6
CFIA (Bi et al., 2025)	43.9	68.0	58.3	64.4	34.6	33.5	28.7
ReDT-Det	46.8	72.1	59.8	64.7	41.5	41.9	32.3

real-world UAV applications. Our ReDT-Det approach surpasses all competing methods by combining NightDrone-YOLO with IFDT-Enhancer, a low-light image enhancement model.

Compared to YOLOv11 (62.3 % mAP, 9.4M Params, 21.3G FLOPs), NightDrone-YOLO achieves higher accuracy (63.8 % mAP) with a moderate parameter increase (21.4M) and efficient computation (24.1G FLOPs). ReDT-Det delivers the highest detection performance (65.2 % mAP) with only a 0.7 % parameter increase over NightDrone-YOLO. By integrating IFDT-Enhancer, ReDT-Det improves nighttime detection results, achieving a 2.9 % increase in mAP and a 3.6 % improvement in small object detection precision, while maintaining computational efficiency (32.7G FLOPs), which is much lower than specialized UAV detectors like HiCAL (238.9G FLOPs) and CFIA (110.1G FLOPs). This highlights the importance of low-light image enhancement in improving object visibility and detection accuracy in UAV-based nighttime scenarios.

By effectively fusing low-light enhancement, progressive feature extraction, and multi-scale information interaction, ReDT-Det significantly outperforms all existing methods. Its superior results in both overall detection and small object recognition validate its robustness and practical applicability for UAV-based nighttime surveillance and monitoring.

DroneVehicle(Night). Table 7 presents experimental results on the DroneVehicle(Night) test dataset. The analysis on the DroneVehicle(Night) test dataset echoes the observations from NightDrone-Mix. ReDT-Det leads in six of seven key metrics, achieving an mAP of 46.8 % and AP_{50} of 72.1 %, and excels in category-specific precision for Car, Truck, and Freight car, while ranking second in Bus and Van detection. When matched against specialized UAV detectors, ReDT-Det outperforms CEASC by 7.0 % in mAP, UM-YOLO by 2.8 %, PDPA-PAN by 9.8 %, HiCAL by 2.6 %, and CFIA by 2.9 %, showing ReDT-Det is more suitable for UAV-based object detection during nighttime. Traditional detectors fall short-AlignDETR, the best conventional method with 44.6 % mAP, lags behind ReDT-Det by 2.2 %, and YOLO variants like YOLOv11 (43.5 % mAP) show limited improvement. Notably, ReDT-Det adeptly handles challenging nighttime categories, outperforming YOLOv8 in Freight car detection by 0.5 % and AlignDETR in Truck detection by 1.3 %. These results confirm ReDT-Det's specialized low-light UAV scenario optimizations, especially its superior resistance to artificial illumination-induced metallic reflections on vehicle surfaces.

4.3.2. Comparison experiments on advanced image enhancer with YOLO detection

Our proposed ReDT-Det framework integrates IFDT-Enhancer for low-light image enhancement and NightDrone-YOLO for object detection. To assess the impact of IFDT-Enhancer on nighttime UAV-based vehicle detection, we conducted experiments on the NightDrone-Mixdataset, comparing our approach with several state-of-the-art low-light enhancement methods, including SCINet (Ma et al., 2022a), RetinexFormer (Cai et al., 2023a), IAT (Cui et al., 2024), LYTNet Brateanu et al. (2024), DENet (Qin et al., 2022), PENet (Yin et al., 2023), CPAEnhancer (Zhang et al., 2024), and AirNet (Li et al., 2022a). YOLOv11 was used as the object detector across all experiments. Additionally, to verify the generalization capability of IFDT-Enhancer, we conducted experiments on the ExDark dataset, which contains diverse low-light conditions, including indoor and outdoor scenes.

NightDrone-Mix. Table 8 presents the detection performance of various image enhancement methods on the NightDrone-Mixdataset along with an analysis of their computational efficiency. The experimental results indicate that although SCINet and DENet achieve low FLOPs and fast inference speeds, their detection performance is not optimal. This suggests that relying solely on lightweight models does not guarantee high accuracy. In contrast, AirNet and CPAEnhancer are designed to restore images affected by unknown degradations and they demonstrate superior enhancement quality. However, their contribution to improving the precision of UAV-based vehicle detection is limited. In addition, many existing methods incur high FLOPs, which makes them unsuitable for real-time deployment on UAV platforms. In comparison with the baseline RetinexFormer, our IFDT-Enhancer achieves a notable improvement of over 1 % in both detection mAP and AP_s. At the same time, it reduces model complexity by 4.8 GFLOPs and decreases inference time by 1.2 ms. These results demonstrate that the Retinex-guided illumination strategy used in IFDT-Enhancer, together with the introduction of a multi-head differential attention mechanism, effectively balances detection accuracy with computational efficiency. This balance makes IFDT-Enhancer a practical choice for nighttime vehicle detection from UAV imagery.

ExDark. To further assess the robustness of IFDT-Enhancer, we evaluated its performance on the ExDark dataset, which features a wide range of indoor and outdoor nighttime scenes with varying illumination levels. All methods were retrained and evaluated using the YOLOv11 detector. Table 9 presents the experimental results, which include per-class AP scores and the overall mAP calculated at an IoU threshold of 0.5. Our IFDT-Enhancer achieved an overall mAP of 73.6 %, outperforming the previous best Retinex-based method, RetinexFormer (Cai et al., 2023a), by 1.5 % and surpassing the chain-of-thought enhancement model, CPAEnhancer (Zhang et al., 2024), by 1.4 %. These results confirm that IFDT-Enhancer not only enhances visibility in low-light conditions but also significantly improves object detection accuracy across different domains. Moreover, IFDT-Enhancer demonstrated the best performance on five object categories, namely bottle, bus, chair,

Table 8
The detection and lightweight performance of various methods on NightDrone-Mix testing dataset.

Method	mAP	AP_s	AP_m	AP_l	$FLOPs(G)$	$Params(M)$	ms
SCINet (Ma et al., 2022a)	61.6	51.5	68.8	70.5	22.4	9.42	2.7
IAT (Cui et al., 2024)	61.1	49.8	67.6	68.8	39.2	9.51	15.2
CPAEnhancer (Zhang et al., 2024)	61.9	49.8	67.6	68.8	34.1	9.93	11.9
PENet (Yin et al., 2023)	61.2	50.0	67.3	69.3	43.7	9.51	18.2
DENet (Qin et al., 2022)	61.8	50.4	68.1	70.8	25.3	9.47	6.1
LYTNet (Brateanu et al., 2024)	61.5	50.4	67.1	70.3	42.6	9.46	17.1
AirNet (Li et al., 2022a)	62.1	51.3	68.1	70.2	349.7	12.39	23.6
RetinexFormer (Cai et al., 2023a)	62.7	51.5	68.8	70.5	31.5	9.49	13.8
IFDT-Enhancer	63.9	52.6	70.4	71.6	26.7	9.47	12.6

Table 9
The performance of various methods for object detection on the ExDark testing dataset.

Methods	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motor	People	Table	Total
SCINet (Ma et al., 2022a)	87.9	63.6	62.4	93.1	79.5	69.4	58.4	61.4	70.3	71.0	75.4	61.6	71.2
IAT (Cui et al., 2024)	85.0	69.3	64.2	93.4	80.7	73.9	60.2	61.7	70.3	68.5	74.9	57.1	71.6
CPAEnhancer Zhang et al. (2024)	85.4	65.8	68.0	94.3	79.2	74.2	58.1	63.2	66.8	74.2	78.3	59.2	72.2
PENet (Yin et al., 2023)	86.1	64.5	67.6	93.5	78.6	71.5	60.1	61.3	70.1	69.5	76.8	60.1	71.5
DENet (Qin et al., 2022)	85.3	64.1	67.4	93.1	78.1	71.8	60.7	61.6	69.1	69.8	77.3	60.7	71.3
LYTNET (Brateanu et al., 2024)	83.8	63.5	64.6	93.4	78.3	71.6	56.6	56.9	72.0	70.1	72.5	59.4	70.2
AirNet (Li et al., 2022a)	84.4	64.2	63.9	93.6	78.5	71.1	56.2	56.1	71.2	69.2	71.8	60.1	70.9
RetinexFormer (Cai et al., 2023a)	89.1	64.2	67.5	93.9	81.2	72.6	59.5	63.2	71.1	70.8	76.2	69.4	72.1
IFDT-Enhancer	88.8	66.4	69.5	95.2	81.1	73.1	62.3	64.1	71.9	71.7	76.3	70.1	73.6

Table 10
The overall ablation study on NightDrone-Mixtesting dataset.

Model	Image Enhancer	Detection	Metrics			
			mAP	AP_s	$FLOPs(G)$	ms
$Model1$	-	YOLOv11	62.3	50.5	21.3	2.1
$Model2$	RetinexFormer	YOLOv11	62.7	51.5	31.5	13.8
$Model3$	IFDT-Enhancer	YOLOv11	63.9	52.6	26.7	12.6
$Model4$	-	NightDrone-YOLO	63.8	53.2	22.9	2.9
$Model5$	RetinexFormer	NightDrone-YOLO	64.1	53.4	32.2	14.6
$Model6$	IFDT-Enhancer	NightDrone-YOLO	65.2	54.1	27.1	13.5

cup, and table, while it obtained the second-best results on five other categories, including bicycle, boat, car, dog, and motor. This performance underscores its capability to preserve fine details and improve feature extraction for small and complex objects in nighttime environments.

Extensive evaluations on both the NightDrone-Mixand ExDark datasets demonstrate that our IFDT-Enhancer is highly effective for UAV-based nighttime object detection. It significantly improves detection precision compared to existing enhancement models while maintaining balanced computational efficiency, making it well suited for deployment on UAV platforms. In addition, IFDT-Enhancer exhibits strong generalization capabilities across a wide range of nighttime scenes. By integrating IFDT-Enhancer with the NightDrone-YOLO detector, our ReDT-Det framework achieves state-of-the-art performance and sets a new benchmark for nighttime UAV vehicle detection.

4.4. Ablation experiments

We first validated the efficiency of our IFDT-Enhancer and the object detection model NightDrone-YOLO. Next, we performed ablation experiments on the DR-CSP, CFAA, and SOAF modules to assess the contribution of each module to the overall detection mAP. The results of these ablation studies are summarized in Tables 10 and 11. In addition, we also discuss the initial value of lambda λ_{init} in the multi-head differential self-attention module of IFDTB. The results are shown in Fig. 7.

Analysis of IFDT-Enhancer and NightDrone-YOLO. To validate the impact of the IFDT-Enhancer module on UAV-based nighttime vehicle detection, we employed a combination of RetinexFormer and YOLOv11 as the baseline. The experimental results are presented in Table 10. From the comparisons between $Model2$ and $Model3$, it is evident that using IFDT-Enhancer as the image enhancement module results in a detection mAP that is 1.2 % higher than that of RetinexFormer. At the same time, the model complexity is reduced by 15.24 %, and the inference time is decreased by 1.2 ms. In addition, the comparison between $Model1$ and $Model4$ shows that NightDrone-YOLO enhances detection precision by 1.5 %. Similar conclusions can be drawn from comparing $Model2$ with $Model5$ and $Model3$ with $Model6$, further demonstrating that NightDrone-YOLO outperforms YOLOv11 in both detection mAP and APs. Moreover, an analysis of GFLOPs and inference time in milliseconds confirms that our IFDT-Enhancer has lower model complexity and reduced inference time compared to RetinexFormer. Taken together, these results indicate that the overall performance of ReDT-Det is superior to the baseline.

Table 11
The object detection ablation study on NightDrone-Mixtesting dataset.

DR-CSP	CFAA	SOAF	mAP	AP_s	AP_m	AP_l
			62.3	50.5	67.9	70.5
✓			62.9	51.2	68.1	72.3
✓	✓		63.2	52.3	69.6	72.5
✓	✓	✓	63.8	53.2	70.5	72.8

Fig. 7. Analysis of λ_{init} on the NightDrone-Mixdataset.

Analysis of DR-CSP. In the ablation experiments for object detection, we used YOLOv11 as the baseline. To expand the model's receptive field, we enhanced the original C3K2 module by incorporating DR-CSP for improved object feature extraction. The experimental results indicate that the application of DR-CSP not only boosts the detection mAP but also improves precision across different object scales, with a particularly notable 1.8 % increase in large-scale object detection accuracy. These findings confirm that DR-CSP facilitates more accurate feature representation, thereby enhancing overall object detection performance.

Analysis of CFAA. In the neck section of the object detection model, most approaches employ PAFPN to achieve multi-scale feature fusion. However, this method only allows for the fusion of features between adjacent layers, which limits cross-layer information exchange and hampers the effective use of high-resolution features for small object detection. To address this limitation, we integrated CFAA for cross-layer feature fusion to enhance small object feature representation. Experimental results show that incorporating CFAA leads to a 0.9 % improvement in detection mAP and a 1.8 % increase in detection precision for small-scale objects. These findings demonstrate that CFAA plays a significant role in improving the detection accuracy of small objects.

Analysis of SOAF. Since more than half of the targets in aerial images are small-scale objects, this proportion notably surpasses that of images in natural scenes. To enhance the feature representation of these small-scale objects, we not only employed CFAA for cross-layer feature fusion but also introduced the SOAF module to reinforce the high-resolution detection head ($P3$). Our approach incrementally applies multi-layer feature attention, which enables the final detection layer to prioritize small object detection. Experimental results show that incorporating the SOAF module further enhances the model's capability to detect small objects, achieving an additional increase of 0.9 % in detection performance beyond the improvements provided by CFAA alone and a total improvement of 2.7 % compared to the baseline YOLOv11.

Analysis of λ_{init}. To mitigate the issue of traditional multi-head self-attention mechanisms allocating excessive focus to irrelevant regions, we propose a differential attention computation strategy. As formalized in Eq. (12), this approach subtracts two distinct softmax attention distributions to suppress noise. A hyperparameter λ, derived from Eq. (13), balances the contributions of these distributions. We empirically evaluate $\lambda_{init} \in (0, 1)$ across a range of values $[0.2, 0.8]$ with a stride of 0.2. Experimental results in Fig. 7 demonstrate that the detection performance is robust to variations in λ_{init}, with mAP fluctuations remaining below 0.5 %, indicating stable optimization characteristics. This stability suggests that the differential mechanism inherently regularizes attention patterns rather than relying heavily on hyperparameter tuning.

4.5. Qualitative analysis

To qualitatively assess nighttime vehicle detection performance, we present visual comparisons of detection outputs from various models in Fig. 8. Ground truth bounding boxes are overlaid on the original images to serve as reference standards. We compare the baseline detector YOLOv11, along with ATSS, and UAV-based detector including CEASC, PDPA-PAN, HiCAL, CFIA. In the visualizations, only detection boxes with confidence scores above 0.5 are displayed. Considering the prevalence of small objects in UAV-captured images, we omit category labels and confidence scores from the visuals to better highlight the detection results for small objects; only bounding boxes are shown on the images.

Fig. 8 presents a qualitative comparison between SOTA methods and our ReDT-Det on the NightDrone-Mix dataset. Bounding boxes are color-coded to represent detection results: blue indicates objects predicted by the different methods, red signifies missed targets (false negatives), and yellow denotes false alarms (false positives). The first three columns present results captured under extreme dark conditions, while the last column displays outcomes from low-illumination scenarios. As observed, CEASC yields the least satisfactory detection performance, exhibiting significant false positives or target misses across all scenarios. Compared to all compared methods, our approach demonstrates minimal missed detections and false alarms, with failures primarily occurring on distant small targets in the imagery. The most pronounced detection misses emerge in Scenario 3 (third column), where even our ReDT-Det method shows limitations, indicating that current methodologies remain insufficient for fully addressing object detection under extreme darkness. This represents a critical direction for future research.

Visualization results in Fig. 9 demonstrate significant performance gains achieved through our image enhancement module. The comparison reveals substantially improved detection precision in illumination-enhanced subfigures (b) and (d), where ReDT-Det successfully identifies critical targets previously obscured in original low-light imagery (a) and (c). Our enhancement approach improves bounding box accuracy for low-contrast objects. It also reduces false positives in complex backgrounds. These improvements indicate partial mitigation of low-light challenges in drone detection.

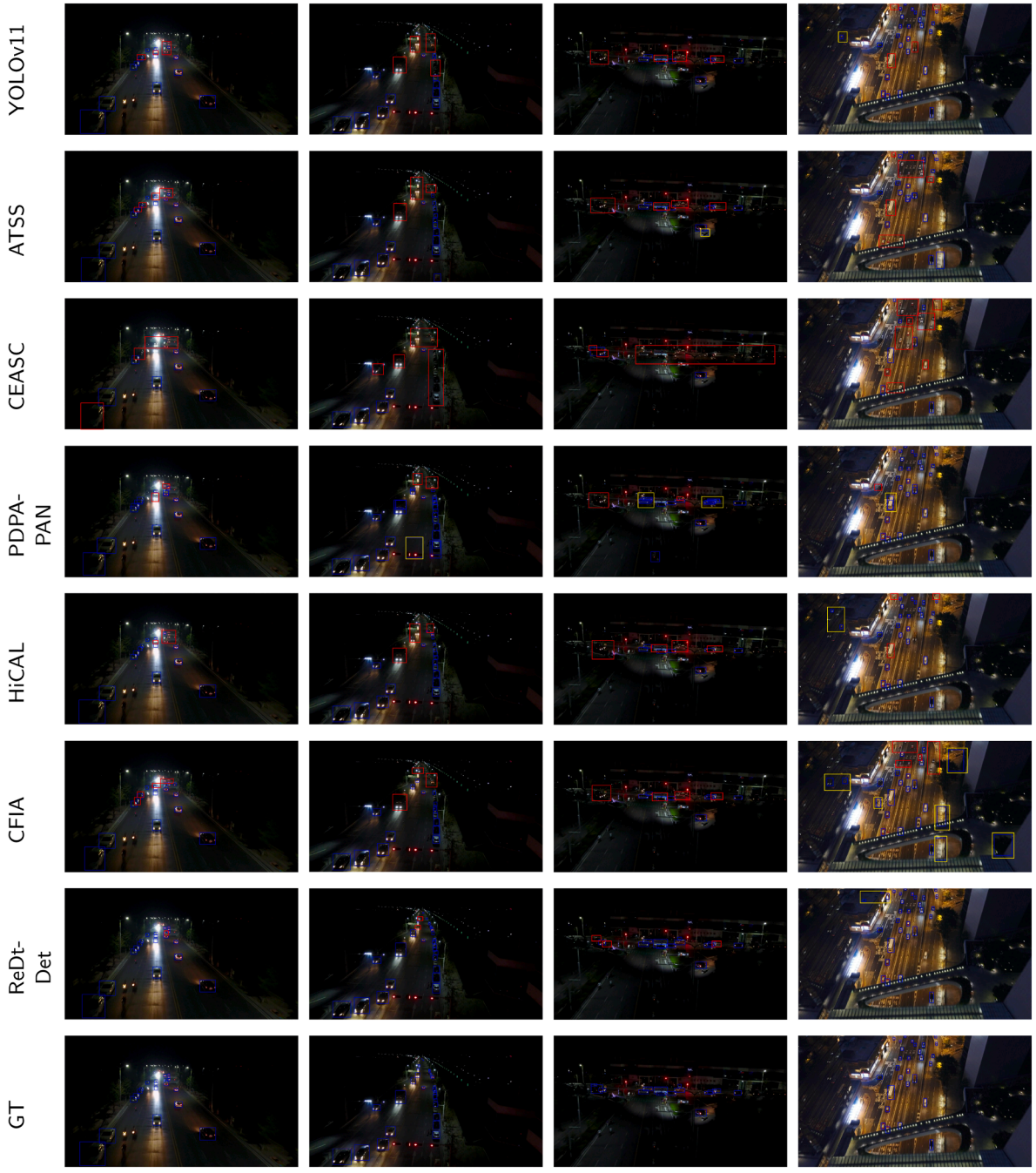

Fig. 8. Qualitative comparison of SOTA methods and our ReDT-Det on NightDrone-Mix. The BBoxes, distinguished by different colors, represent various detection outputs: blue for the prediction results of different methods, red for missed detections, and yellow for false alarms.

4.6. Discussion

ReDT-Det, which integrates low-light enhancement and detection techniques, can effectively identify vehicles in nighttime UAV images. However, given the limited hardware resources available on drones, there is still potential to further reduce FLOPs and improve inference speed when compared to lightweight models such as SCINet. Due to the challenge of collecting paired images of low-light and normal lighting conditions from identical scenes for training, the current version of ReDT-Det employs the low-light enhancement module as a preprocessing step and utilizes detection loss functions to optimize the overall model. As the next step, we plan to simulate low-light conditions using images captured by drones under normal lighting, thereby creating a dataset that includes both normal-light and low-light images. After obtaining image pairs, low-light image enhancement tasks based on UAV images can be performed to further jointly optimize the detection model and improve nighttime vehicle recognition performance.

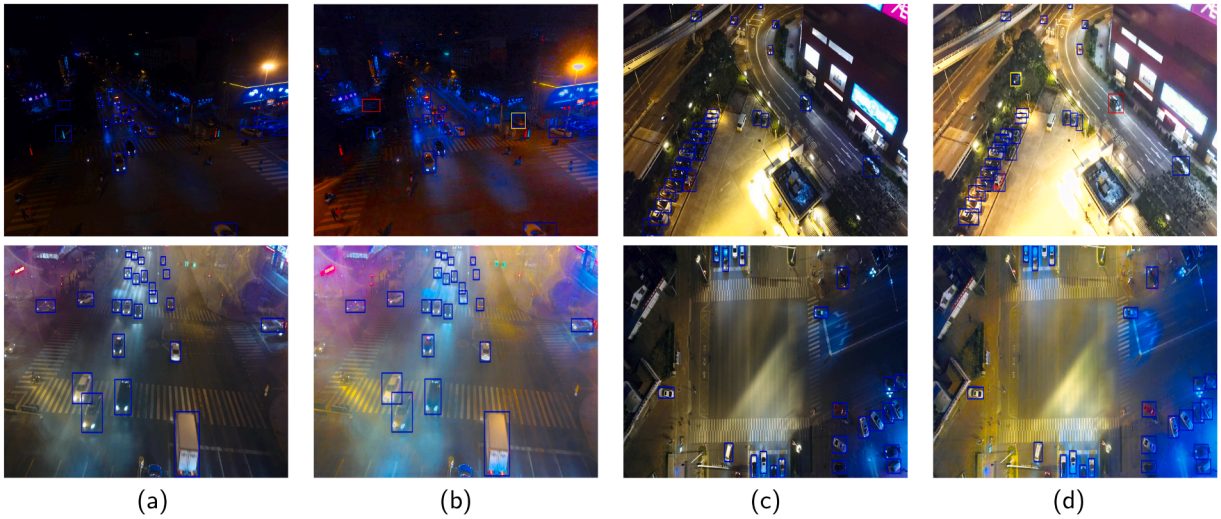

Fig. 9. Visualization of detection results by ReDT-Det on NightDrone-Mix. Subfigures (a) and (c) display original imagery with annotated detection labels, while (b) and (d) present the corresponding illumination-enhanced images generated by ReDT-Det, overlaid with predicted bounding boxes.

5. Conclusion

In this paper, we propose a novel method, ReDT-Det, for vehicle detection in UAV imagery captured at nighttime. Our approach begins by applying a Retinex-guided low-light image enhancement module to mitigate the challenges posed by poor illumination. Specifically, our IFDTB module is designed to compute differential multi-head self-attention while effectively filtering out irrelevant noise, thereby refining the enhanced images. Furthermore, to improve detection accuracy, particularly for small and medium-scale objects, we employ a progressive feature adaptation strategy that comprises three key modules. The DR-CSP module is developed to expand the detector's receptive field and enhance object feature extraction. The CFAA module facilitates cross-layer feature fusion to better aggregate multi-scale information. Finally, the SOAF module is introduced to strengthen the representation of small-scale objects at the high-resolution detection head. Although overall detection accuracy still has room for improvement, the combination of these modules leads to a notable enhancement in detection performance. We further validate our method on NightDrone-Mixdataset, a comprehensive benchmark dataset for nighttime UAV-based vehicle detection. Extensive quantitative and qualitative evaluations on NightDrone-Mixdemonstrate that ReDT-Det outperforms state-of-the-art methods, establishing it as a promising solution for challenging nighttime detection scenarios. In future work, we plan to develop a unified framework that further integrates the image enhancement and detection components, thereby optimizing performance in both low-light enhancement and detection precision.

CRediT authorship contribution statement

Li Chen: Conceptualization, Methodology, Software, Writing - original draft; **Hongbin Deng:** Writing – review & editing, Project administration; **Guanghong Liu:** Data curation, Formal analysis; **Rob Law:** Writing – review & editing; **Dongfang Li:** Software, Validation; **Edmond Q. Wu:** Writing – review & editing; **Limin Zhu:** Writing – review & editing, Software.

Data availability

Data will be made available on request.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest. **Ethical approvals** This article does not contain any studies with human participants or animals performed by any of the authors.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Li Chen reports financial support was provided by Beijing Institute of Technology. Reports a relationship with that includes:. Has patent pending to. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge the financial support provided by the National Natural Science Foundation of China (Grant 62303117) and the Natural Science Foundation of Fujian Province of China (Grant 2025J09022, 2024J01278) and the Military Science and Technology Commission Science and Technology Innovation project under Grant C1692.

References

Bai, J., Yin, Y., & He, Q. (2024). Retinexmamba: Retinex-based mamba for low-light image enhancement. *arXiv preprint arXiv: 2405.03349*.

Bi, Y., Ning, Y., & Nie, X. (2025). Delving into coarse-fine feature interaction alignment for UAV object detection. In *Icassp 2025 - 2025 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Brateanu, A., Balmez, R., Avram, A., & Orhei, C. C. (2024). Lyt-net: Lightweight yuv transformer-based network for low-light image enhancement. *arXiv preprint arXiv: 2401.15204*.

Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., & Zhang, Y. (2023a). Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 12504–12513).

Cai, Y., Lin, J., Wang, H., Yuan, X., Ding, H., Zhang, Y., Timofte, R., & Gool, L. V. (2022). Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. *Advances in Neural Information Processing Systems, 35*, 37749–37761.

Cai, Z., Liu, S., Wang, G., Ge, Z., Zhang, X., & Huang, D. (2023b). Align-DETR: Improving DETR with simple iou-aware BCE loss. *arXiv preprint arXiv: 2304.07527*.

Cao, J., Cholakkal, H., Anwer, R. M., Khan, F. S., Pang, Y., & Shao, L. (2020). D2det: Towards high quality object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11485–11494).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Springer.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., & Lin, D. (2019). MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv: 1906.07155*.

Cheng, G., Wang, J., Li, K., Xie, X., Lang, C., Yao, Y., & Han, J. (2022). Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–11.

Cui, Z., Gu, L., Sun, X., Ma, X., Qiao, Y., & Harada, T. (2024). Aleth-nerf: Illumination adaptive nerf with concealing field assumption. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1435–1444). (*vol. 38*).

Cui, Z., Li, K., Gu, L., Su, S., Gao, P., Jiang, Z., Qiao, Y., & Harada, T. (2022). You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction. *arXiv preprint arXiv: 2205.14871*.

Ding, J., Xue, N., Long, Y., Xia, G.-S., & Lu, Q. (2019). Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2849–2858).

Du, B., Huang, Y., Chen, J., & Huang, D. (2023). Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13435–13444).

Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., & Tian, Q. (2018). The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision* (pp. 370–386).

Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., Peng, T., Zheng, J., Wang, X., Zhang, Y. et al. (2019). Visdrone-DET2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops* (pp. 1–13).

Farhadi, Ali, & Redmon, Joseph (2018). Yolov3: An incremental improvement. In *Computer vision and pattern recognition* (pp. 1–6). Springer Berlin/Heidelberg, Germany (*vol. 1804*).

Feng, M., Yu, H., Dang, X., & Zhou, M. (2024). Category-aware dynamic label assignment with high-quality oriented proposal. *arXiv preprint arXiv: 2407.03205*.

Fu, X., Zeng, D., Huang, Y., Zhang, X.-P., & Ding, X. (2016). A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2782–2790).

Hoanh, N., & Vu Pham, T. (2024). A multi-task framework for car detection from high-resolution UAV imagery focusing on road regions. *IEEE Transactions on Intelligent Transportation Systems*, *25*(11), 17160–17173.

Jiang, Y., Chang, S., & Wang, Z. (2021). Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, *34*, 14745–14758.

Jocher, Glenn, Chaurasia, Ayush et al. (2022). ultralytics/yolov5: V6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. *Zenodo*.

Jocher, G., Qiu, J., Chaurasia, A., 2023. Ultralytics YOLO. https://github.com/ultralytics/ultralytics.

Kang, M., Ting, C.-M., Ting, F. F., & Phan, R. C.-W. (2024). Asf-yolo: A novel yolo model with attentional scale sequence fusion for cell instance segmentation. *Image and Vision Computing*, *147*, 105057.

Kennerley, M., Wang, J.-G., Veeravalli, B., & Tan, R. T. (2023). 2PCNet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11484–11493).

Land, E. H., & McCann, J. J. (1971). Lightness and retinex theory. *Josa*, *61*(1), 1–11.

Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., & Peng, X. (2022a). All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17452–17462).

Li, C., Li, L., Jiang, H., & Weng, o. (2022b). Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv: 2209.02976*.

Li, W., Chen, Y., Hu, K., & Zhu, J. (2022c). Oriented reppoints for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1829–1838).

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., & Yang, J. (2020). Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, *33*, 21002–21012.

Lin, Q., Zhao, J., Fu, G., & Yuan, Z. (2020). Crpn-sfnet: A high-performance object detector on large-scale remote sensing images. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(1), 416–429.

Lin, T. (2017). Focal loss for dense object detection. *arXiv preprint arXiv: 1708.02002*.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2125).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, Zurich, Switzerland, september 6–12, 2014, proceedings, Part V 13* (pp. 740–755). Springer.

Liu, Q., Liu, R., Zheng, B., Wang, H., & Fu, Y. (2024). Infrared small target detection with scale and location sensitivity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17490–17499).

Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., & Zhang, L. (2022). Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv: 2201.12329*.

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8759–8768).

Loh, Y. P., & Chan, C. S. (2019). Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, *178*, 30–42.

Ma, C., Xu, C., Zhou, P., & Zhang, Y. (2023). Low-light aerial image enhancement algorithm based on retinex theory. In *2023 international conference on image processing, computer vision and machine learning (ICICML)* (pp. 138–142). IEEE.

Ma, L., Ma, T., Liu, R., Fan, X., & Luo, Z. (2022a). Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5637–5646).

Ma, T., Ma, L., Fan, X., Luo, Z., & Liu, R. (2022b). Pia: Parallel architecture with illumination allocator for joint enhancement and detection in low-light. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 2070–2078).

Patel, H., Prajapati, K., Sarvaiya, A., Upla, K., Raja, K., Ramachandra, R., & Busch, C. (2022). Depthwise convolution for compact object detector in nighttime images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 379–389).

Qin, Q., Chang, K., Huang, M., & Li, G. (2022). Denet: Detection-driven enhancement network for object detection under adverse weather conditions. In *Proceedings of the Asian conference on computer vision* (pp. 2813–2829).

Rahman, S., Rahman, M. M., Abdullah-Al-Wadud, M., Al-Quaderi, G. D., & Shoyaib, M. (2016). An adaptive gamma correction for image enhancement. *EURASIP Journal on Image and Video Processing*, *2016*, 1–13.

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster r-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149.

Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10781–10790).

Telikani, A., Sarkar, A., Du, B., & Shen, J. (2024). Machine learning for UAV-aided ITS: A review with comparative study. *IEEE Transactions on Intelligent Transportation Systems*, *25*(11), 15388–15406.

Vo, H.-T., & Quach, L.-D. (2023). Advanced night time object detection in driver-assistance systems using thermal vision and yolov5. *International Journal of Advanced Computer Science and Applications*, *14*(6).

Wang, Ao, Chen, Hui, Liu, Lihao, Chen, Kai et al. (2024a). Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, *37*, 107984–108011.

Wang, Chengcheng, He, Wei, Nie, Ying, Guo, Jianyuan, Liu, Chuanjian, Wang, Yunhe, & Han, Kai (2024b). Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. *Advances in Neural Information Processing Systems*, *36*.

Wang, C.-Y., Liao, H.-Y. M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., & Yeh, I.-H. (2020a). Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 390–391).

Wang, C.-Y., Yeh, I.-H., & Mark Liao, H.-Y. (2024c). Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision* (pp. 1–21). Springer.

Wang, S., Zheng, J., Hu, H.-M., & Li, B. (2013). Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, *22*(9), 3538–3548.

Wang, W., Peng, Y., Cao, G., Guo, X., & Kwok, N. (2020b). Low-illumination image enhancement for night-time UAV pedestrian detection. *IEEE Transactions on Industrial Informatics*, *17*(8), 5208–5217.

Wang, Z.-G., Liang, Z.-H., & Liu, C.-L. (2009). A real-time image processor with combining dynamic contrast ratio enhancement and inverse gamma correction for PDP. *Displays*, *30*(3), 133–139.

Wei, H., Liu, X., Xu, S., Dai, Z., Dai, Y., & Xu, X. (2022). Dwrseg: Rethinking efficient acquisition of multi-scale contextual information for real-time semantic segmentation. *arXiv preprint arXiv: 2212.01173*.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J. E., Keutzer, K., & Vajda, P. (2021). Visual transformers: Where do transformers really belong in vision models? In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 599–609).

Xi, Y., Jia, W., Miao, Q., Feng, J., Ren, J., & Luo, H. (2024). Detection-driven exposure-correction network for nighttime drone-view object detection. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–14.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., & Zhang, L. (2018). Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3974–3983).

Xie, X., Cheng, G., Wang, J., Yao, X., & Han, J. (2021). Oriented r-CNN for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3520–3529).

Xing, L., Qu, H., Xu, S., & Tian, Y. (2023). Clegan: Toward low-light image enhancement for uavs via self-similarity exploitation. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–14.

Xu, K., Yang, X., Yin, B., & Lau, R. W. H. (2020). Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2281–2290).

Xue, X., He, J., Ma, L., Wang, Y., Fan, X., & Liu, R. (2022). Best of both worlds: See and understand clearly in the dark. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 2154–2162).

Yang, F., Fan, H., Chu, P., Blasch, E., & Ling, H. (2019). Clustered object detection in aerial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8311–8320).

Yang, G., Lei, J., Zhu, Z., Cheng, S., Feng, Z., & Liang, R. (2023). Afpn: Asymptotic feature pyramid network for object detection. In *2023 IEEE international conference on systems, man, and cybernetics (SMC)* (pp. 2184–2189). IEEE.

Ye, J., Fu, C., Cao, Z., An, S., Zheng, G., & Li, B. (2022). Tracker meets night: A transformer enhancer for UAV tracking. *IEEE Robotics and Automation Letters*, *7*(2), 3866–3873.

Ye, T., Dong, L., Xia, Y., Sun, Y., Zhu, Y., Huang, G., & Wei, F. (2024). Differential transformer. *arXiv preprint arXiv: 2410.05258*.

Yin, X., Yu, Z., Fei, Z., Lv, W., & Gao, X. (2023). Pe-yolo: Pyramid enhancement network for dark object detection. In *International conference on artificial neural networks* (pp. 163–174). Springer.

Ying, Z., Zhou, J., Zhai, Y., Quan, H., Li, W., Genovese, A., Piuri, V., & Scotti, F. (2024). Large-scale high-altitude UAV-based vehicle detection via pyramid dual pooling attention path aggregation network. *IEEE Transactions on Intelligent Transportation Systems*, *25*(10), 14426–14444.

Yu, W., Yang, T., & Chen, C. (2021). Towards resolving the challenge of long-tail distribution in UAV images for object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3258–3267).

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M.-H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5728–5739).

Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., Wang, Y., & Guo, B. (2022a). Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11304–11314).

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., & Shum, H.-Y. (2022b). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv: 2203.03605*.

Zhang, S., Chi, C., Yao, Y., Lei, Z., & Li, S. Z. (2020). Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9759–9768).

Zhang, X., Liu, Y., Li, A., Han, K., Zhao, Q., & Wu, J. (2025). HiCAL: Hierarchical consistency-based active learning for drone-view object detection. *IEEE Transactions on Geoscience and Remote Sensing*, *63*, 1–14.

Zhang, Y., Wu, Y., Liu, Y., & Peng, X. (2024). Cpa-enhancer: Chain-of-thought prompted adaptive enhancer for object detection under unknown degradations. *arXiv preprint arXiv: 2403.11220*.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. (2024). Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16965–16974).

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. S. et al. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6881–6890).

Zhu, Junqing, Wu, Yuxuan, & Ma, Tao (2024). Multi-object detection for daily road maintenance inspection with UAV based on improved YOLOv8. *IEEE Transactions on Intelligent Transportation Systems*, *25*(11), 16548–16560.