# GDEIM-SF: A Lightweight UAV Detection Framework Coupling Dehazing and Low-Light Enhancement

Jihong Zheng and Leqi Li *

Posted Date: 16 December 2025

doi: 10.20944/preprints202512.1468.v1

Keywords: drone vision; small object detection; DETR; multi-scale fusion; lightweight design

*Article*

# GDEIM-SF: A Lightweight UAV Detection Framework Coupling Dehazing and Low-Light Enhancement

**Jihong Zheng [1] and Leqi Li [2,*]**

[1] College of Urban Construction, Yangtze University, Jingzhou 434100, China

[2] College of Electronic Information and Electrical Engineering, Yangtze University, Jingzhou 434100, China

* Correspondence: 18164134806@163.com

**Abstract**

In complex traffic environments, image degradation caused by haze, low illumination, and occlusion significantly undermines the reliability of vehicle and pedestrian detection. To address these challenges, this paper proposes an aerial vision framework that tightly couples multi-level image enhancement with a lightweight detection architecture. At the image preprocessing stage, a cascaded "dehazing + illumination" module is constructed. Specifically, a learning-based dehazing method, Learning Hazing to Dehazing, is employed to restore long-range details affected by scattering artifacts. Additionally, HVI-CIDNet is introduced to decouple luminance and chrominance in the Horizontal/Vertical Intensity (HVI) color space, thereby simultaneously enhancing structural fidelity in low-light regions and achieving global brightness consistency. On the detection side, a lightweight yet robust detection architecture, termed GDEIM-SF, is designed. It adopts GoldYOLO as the lightweight backbone and integrates D-FINE as an anchor-free decoder. Furthermore, two key modules, CAPR and ASF, are incorporated to enhance high-frequency edge modeling and multi-scale semantic alignment, respectively. Evaluated on the VisDrone dataset, the proposed method achieves improvements of approximately 2.5–2.7 percentage points in core metrics such as mAP@50–90 compared to similar lightweight models (e.g., the DEIM baseline and YOLOv12s), while maintaining low parameter count and computational overhead. This ensures a balanced trade-off among detection accuracy, inference efficiency, and deployment adaptability, providing a practical and efficient solution for UAV-based visual perception tasks under challenging imaging conditions.

**Keywords:** drone vision; small object detection; DETR; multi-scale fusion; lightweight design

## 1. Introduction

With the rapid development of smart cities and the emerging low-altitude economy, unmanned aerial vehicles (UAVs) have seen widespread applications in urban traffic surveillance, public safety inspection, and emergency response. UAV-based aerial visual analysis offers advantages such as wide-area coverage and high flexibility. However, due to complex imaging environments, UAV vision perception still faces significant challenges. Particularly under adverse conditions such as haze and low illumination, image quality is severely degraded. Combined with issues such as small target sizes, high target density, and frequent occlusions, traditional object detection algorithms often perform suboptimally on UAV imagery, falling short of meeting the dual demands for accuracy and efficiency in real-world deployments.

In recent years, the YOLO (You Only Look Once) series has emerged as the mainstream framework for UAV-based object detection, owing to its exceptionally high inference efficiency. Since the introduction of YOLOv1 by Joseph Redmon et al. [1], the architecture has undergone continuous iterations—from single-scale detection to multi-scale feature fusion and collaborative deep-shallow

 doi:10.20944/preprints202512.1468.v1



network designs—culminating in YOLOv8, which significantly enhances performance in small object detection and edge deployment scenarios. Building upon the YOLO framework, several recent studies have further improved lightweight designs and small object detection capabilities. For instance, Yin et al. [2] proposed a method that integrates image enhancement with a lightweight YOLOv5 framework to significantly boost detection accuracy for small targets in UAV imagery. Chen et al. [3] designed a multi-scale feature extraction mechanism that enhances detection efficiency and robustness in aerial images. To address real-time constraints on edge devices, Wang et al. [4] developed a lightweight YOLOv5-based framework for real-time forest smoke detection, achieving a favorable balance between real-time performance and accuracy.

Despite the clear advantages of the YOLO series in speed and deployment feasibility, its detection process still relies on predefined anchor boxes and Non-Maximum Suppression (NMS). This architecture struggles in scenarios involving overlapping, occlusion, and scale variations common in dense urban aerial scenes, leading to false positives and missed detections. Nikouei et al. [5] highlighted the limited adaptability of anchor-based mechanisms in complex dense scenes, while Li et al. [6] demonstrated YOLO's sensitivity to image degradation under low-light aerial conditions, often resulting in localization errors and reduced detection precision.

To address YOLO's inherent structural limitations, Carion et al. [7] introduced the DETR (DEtection TRansformer) model, incorporating a Transformer-based architecture to achieve end-to-end object detection. DETR replaces traditional anchors and NMS with a self-attention mechanism, significantly simplifying the detection pipeline. However, the original DETR suffers from slow convergence and insensitivity to small objects during training.

Subsequent studies have improved upon the DETR framework. Zhu et al. [8] proposed Deformable DETR, which enhances small object detection and complex background modeling via multi-scale deformable attention. Han et al. [9] adapted this model for UAV aerial imagery, achieving improved localization performance for small targets. To further balance accuracy and efficiency, Shufang et al. [10] introduced RT-DETR (Real-Time DETR) by redesigning the decoder to accelerate inference. Chen et al. [11] developed Freq-DETR, which introduces frequency-domain modeling to improve the recovery of high-frequency details. Additionally, Han et al. [12] proposed CAEM-DETR, which enhances detection performance in complex environments through contrastive attention and multi-domain feature fusion.

While the YOLO and DETR families have made significant strides in UAV image detection, several critical challenges remain unresolved under complex real-world conditions:Severe image degradation under haze and low-light environments blurs semantic features, which current models struggle to recover effectively;Dense small-object distributions and large scale variations due to the complex aerial perspective of UAVs make small targets prone to omission;Frequent occlusions and background clutter limit robustness, especially for YOLO's anchor-based matching;Limited feature fusion granularity, where Transformer-based models, despite their global modeling capabilities, lack mechanisms to recover fine-grained high-frequency textures.To address these limitations, we propose a novel visual detection framework that integrates multi-stage image enhancement with an enhanced DETR architecture, termed GDEIM-SF. The proposed framework provides comprehensive optimization for degraded input conditions, small-object detection, and efficient deployment. Specifically, a preprocessing stage is introduced that combines dehazing and low-light enhancement modules to significantly improve image quality. This is followed by the construction of a lightweight and robust GDEIM backbone network, where a multi-dimensional guided enhancement module improves cross-scale feature fusion and high-frequency modeling of small targets. The primary contributions of this study are summarized as follows:

1. To address the common degradation in UAV imagery captured under haze and low-light conditions, we design a cascaded preprocessing framework that integrates Learning Hazing to Dehazing (LHD) and HVI-CIDNet. This framework provides more stable, low-noise, and highly discriminative feature inputs for downstream detection modules, thereby enhancing detection robustness and reliability under adverse imaging conditions.
2. To balance small-object detection and resource efficiency in UAV applications, we propose a lightweight and robust backbone–neck architecture, termed GDEIM-SF. This is based on a

 doi:10.20944/preprints202512.1468.v1



modified GoldYOLO framework with an aggregate-distribute feature flow strategy for efficient feature integration and semantic alignment. Features are first uniformly aggregated across multiple levels and then distributed to different branches as needed, preserving both local details and global context. Additionally, the proposed CAPR (C2-Aware P2 Restoration) module leverages shallow detail features as priors to upsample and reconstruct high-resolution low-level pyramid features, which are fused with the main features via gated residual connections. This significantly enhances edge modeling capabilities for small targets.

3. To tackle the challenge of multi-scale target distributions in complex aerial scenes, we introduce the Scale-Adaptive Fusion (SAF) module, which improves semantic consistency and perceptual granularity across pyramid features. During fusion, a frequency-domain prior is incorporated via spectral enhancement mechanisms to strengthen the expression of high-frequency details, thereby improving the model's ability to preserve fine edge and texture information in small objects.

## 2. Methodology

To address the frequent image degradation encountered by UAVs operating in adverse conditions such as haze and low illumination, we propose an integrated visual perception framework that combines multi-stage image enhancement with a lightweight detection network. The overall system pipeline is illustrated in Figure 1.

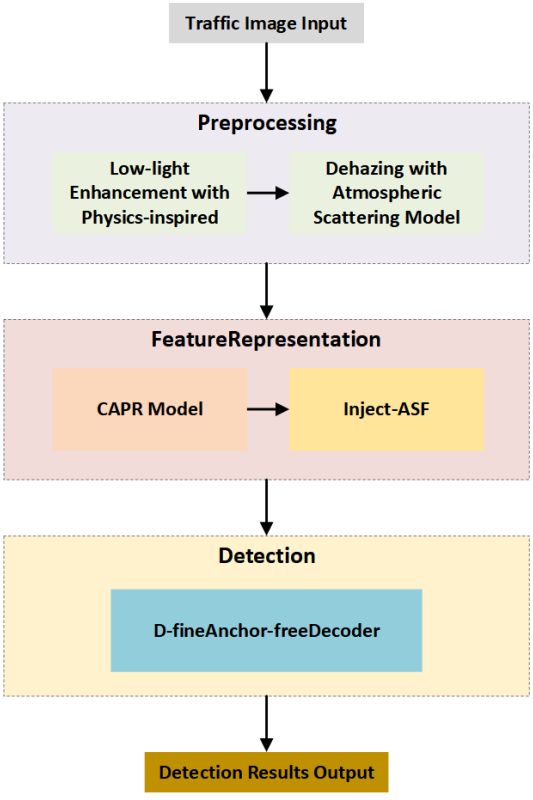


**Figure 1.** Overview of the proposed detection pipeline.

The system begins by acquiring raw aerial images through a UAV platform, which are then processed by an image quality enhancement module to improve visual discriminability. The preprocessing stage consists of two key submodules targeting degradation in aerial imaging:1.Dehazing Module: Based on Learning Hazing to Dehazing (LHD) [13], this module learns the mapping between atmospheric scattering and image degradation. It effectively suppresses contrast loss and edge blurring caused by haze, thereby restoring visual clarity.2.Illumination Enhancement Module: Built upon HVI-CIDNet [14], this module operates in the Horizontal–Vertical Intensity (HVI) color space to decouple luminance and chrominance. By learning structural priors in low-light regions, it achieves both detail restoration and global brightness correction.This multi-stage

 doi:10.20944/preprints202512.1468.v1



image enhancement strategy significantly improves structural fidelity and illumination balance in the input images, thus providing a more stable and discriminative feature foundation for subsequent detection tasks.

After enhancement, the images are passed into a lightweight backbone network based on GoldYOLO [15] for feature extraction. GoldYOLO employs a gather-and-distribute mechanism to perform centralized integration and directional redistribution of multi-scale semantic information, maintaining a balance between semantic abstraction and spatial resolution.To further enhance the network's ability to model small object boundaries and texture details, the backbone is integrated with the CAPR (C2-Aware P2 Restoration) module. By introducing shallow detail priors and adopting a gated residual mechanism, CAPR enables fine-grained reconstruction of high-resolution pyramid layers, thereby improving robustness in the perception of small-scale structures.

At the feature fusion stage, we introduce a Scale-Adaptive Fusion (SAF) module to improve multi-scale semantic alignment and the expression of high-frequency details. This module incorporates frequency-domain priors to preserve and enhance textures and edge features in the feature maps. Notably, it achieves this without significantly increasing model complexity (GFLOPs), effectively improving fine-grained fusion capability.

For the detection head, we adopt a DFINE decoder designed in the style of a Real-Time DETR decoder. This decoder employs a multi-scale memory mechanism and a deformable cross-layer attention strategy to dynamically model contextual dependencies within candidate regions. It ultimately outputs object bounding boxes and class confidence scores in an end-to-end fashion.

The proposed detection framework is a synergistic integration of:Image Quality Enhancement via LHD + HVI-CIDNet; Lightweight Backbone with GoldYOLO + CAPR;Frequency-Domain Guided Scale-Adaptive Fusion (SAF);Transformer-Based Decoder (DFINE).This design significantly enhances detection precision and recall for small, dense objects in complex low-altitude urban scenarios involving haze, low light, and occlusion. Furthermore, the lightweight design of each submodule ensures strong inference efficiency and adaptability to edge deployment.Experimental results demonstrate that, in challenging urban aerial environments such as nighttime streets, complex intersections, and low-light occlusion conditions, the proposed method outperforms existing mainstream detection architectures in both accuracy and stability. Notably, it shows outstanding performance in detecting small targets such as vehicles and pedestrians, confirming its practical value and potential for deployment in real-world UAV-based visual perception applications.

### *2.1. Data Collection*

To validate the adaptability and robustness of the proposed detection framework under complex imaging environments, this study selects the widely used VisDrone dataset [16] as the fundamental experimental platform. This dataset, collected and constructed by the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, contains 263 video sequences and 10,209 static images, with more than 2.6 million annotated targets in total. VisDrone presents significant challenges in multiple dimensions, including: (1) small targets densely distributed; (2) complex imaging backgrounds, involving highly interfering elements such as buildings, trees, and vehicles; and (3) diverse viewpoints and uneven illumination. Therefore, the dataset has become a key benchmark for current UAV object detection tasks under harsh environmental conditions.

To enhance the robustness of the model under haze-degraded conditions, this study introduces a synthetic image degradation method based on the Atmospheric Scattering Model. This model effectively simulates the scattering and attenuation of light by atmospheric particles. The basic imaging model is expressed as:

$$I(x) = J(x) \cdot t(x) + A \cdot \big(1 - t(x)\big), \tag{1}$$

$I(x)$ denotes the intensity of the hazy image received by the camera, $J(x)$is the radiance of the clear scene, A represents the global atmospheric light, and $t(x)$ is the medium transmission rate, defined as:

$$t(x) = e^{-\beta d(x)}, \tag{2}$$

 doi:10.20944/preprints202512.1468.v1



where $\beta$ is the scattering coefficient and $d(x)$ denotes the distance between the scene point and the imaging device. By adjusting the values of $\beta$ and A, image degradation under various haze concentrations can be simulated. This method generates image samples with multiple haze levels while maintaining spatial structural consistency, thereby enhancing the model's adaptability to atmospheric scattering.

Meanwhile, to evaluate the perception capability and stability of the detection framework under low-light conditions, a physically inspired low-light image synthesis pipeline is designed. In this process, the standard sRGB image is first approximately linearized to simulate the inverse gamma process of a camera:

$$I_{lin} = I^{\gamma_e}, \quad \gamma_e \approx 2.2 \tag{3}$$

Then, Exposure Value (EV) is introduced to simulate global illumination attenuation, resulting in a low-light image:

$$J = 2^{-EV}, \quad EV \in [1,4] \tag{4}$$

where EV values from 1 to 4 correspond to increasing levels of darkening, from mild to severe. To simulate non-uniform lighting commonly seen in nighttime or urban street scenes, a normalized variable illumination field $L(x) \in (0,1]$ is further introduced, defined as:

$$L(x) = a + (1-a)(1 - vr(x)^2)\big(1 + \delta(x)\big) \tag{5}$$

Here, $a \in [0.15,0.35]$ represents the base illumination, $v \in [0,0.6]$ controls the vignette intensity, $r(x)^2$ denotes the normalized radius from pixel x to the image center, and $\delta(x)$ is a large-scale Gaussian-smoothed low-frequency random field used to simulate gradual illumination variations caused by streetlights or shadows. The final illumination-adjusted image is represented as:

$$J'(x) = L(x) \cdot J(x) \tag{6}$$

To further enhance the realism of the synthesized images, signal-to-noise modeling is introduced. Signal-dependent shot noise $\epsilon s(x)$ and signal-independent readout noise $\epsilon_r(x)$ are added to $J'(x)$, resulting in the synthesized image:

$$\tilde{J}(x) = \text{clip}(J'(x) + \epsilon_s(x) + \epsilon_r(x), 0,1) \tag{7}$$

Finally, the image is restored to sRGB space through gamma correction:

$$I_{dark} = \tilde{J}^{1/\gamma_e} \tag{8}$$

In summary, based on the VisDrone dataset, this study constructs an extended training and testing dataset by integrating the atmospheric scattering model and a physically inspired low-light model, covering multiple real-world degradation conditions (e.g., haze, low illumination, non-uniform lighting, and imaging noise). This extended dataset significantly increases the complexity and challenge of training samples, providing a solid foundation for verifying the robustness of the proposed detection framework under complex conditions. It also offers reproducibility and scalability.

### *2.2. Dehazing*

Under adverse weather conditions, UAV-captured images often suffer from low contrast, blurred edges, and structural degradation, which severely impacts the detection accuracy of critical targets such as pedestrians and vehicles. To address this challenge, various image dehazing methods have been proposed.

He et al. [17] proposed the Dark Channel Prior (DCP) method, which estimates the transmission map by statistically analyzing local minimum channels and inverts the atmospheric scattering model. It is one of the most representative physics-based prior methods. Zhu et al. [18] further introduced the Color Attenuation Prior (CAP), modeling image brightness and saturation as a linear combination to improve the stability of transmission map estimation. Berman et al. [19] proposed the Non-Local Prior (NLP), which utilizes repeated color structures within the image for non-local clustering, thereby guiding the dehazing process.

 doi:10.20944/preprints202512.1468.v1



With the development of deep learning, data-driven dehazing methods have gained increasing attention. DehazeNet, proposed by Cai et al. [20], is one of the earliest convolutional neural network-based models, capable of directly learning transmission maps from images. DCPDN, introduced by Qu et al. [21], combines physical modeling and image enhancement to estimate transmission maps and atmospheric light in an end-to-end fashion, achieving more refined image restoration. Gao et al. [22] designed a novel Multi-Scale Density-Aware Network (MSDAN) using a multi-scale architecture to capture haze structures at different scales, thus improving overall restoration performance.

In recent years, generative models have shown impressive performance in dehazing tasks. Cycle-Dehaze, proposed by Engin et al. [23], utilizes CycleGAN to perform unpaired image-to-image translation for haze removal. Trident Dehazing Network (TDN), proposed by Liu et al. [24], adopts a triple-branch architecture to reconstruct image content from coarse to fine, thereby adapting to regional variations in heavy and light haze.

Building upon these advances, Wang et al. recently proposed the Learning Hazing to Dehazing (LHD) model, which is the first to introduce Diffusion Probabilistic Models into image dehazing. It establishes an end-to-end "generation-to-restoration" dehazing pipeline composed of two key components:1.HazeGen Module: This module generates realistic hazy images from textual prompts. By leveraging the text-to-image generation capabilities of diffusion models, combined with hybrid objective modeling and stochastic sampling strategies, it produces diverse and authentic haze samples, greatly expanding the training data space.2.DiffDehaze Module: A diffusion-based image restoration network trained on large-scale synthetic and real hazy image datasets. This module incorporates a statistical alignment operation (AlignOp) and a haze-density-aware fidelity-guided mechanism to achieve an efficient trade-off between structural restoration and detail preservation. It enhances the dehazing effect while maintaining the natural appearance of the image.The method relies on the forward–reverse modeling process of diffusion models, with core formulations as follows:

Forward diffusion process of the diffusion model:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) \tag{9}$$

where $\{\beta_t\}$ denotes the noise variance schedule at timestep t.

Reverse denoising (training objective):

$$L_{denoise} = \mathbb{E}_{t,x_0,\epsilon}\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \tag{10}$$

The model learns to predict the added noise.

AlignOp operation: Within local patch ppp, hybrid statistical quantities are replaced, aligning the intermediate reconstruction statistics of the generated image with that of the clean reference. Mathematically, this is expressed as:

$$\tilde{J}_p = \sigma_p(I) \cdot \frac{J_{early,p} - \mu_p(J_{early,p})}{\sigma_p(J_{early,p})} + \mu_p(I) \tag{11}$$

where $\mu_p(\cdot), \sigma_p(\cdot)$ denote the mean and standard deviation within the patch. AlignOp transfers statistics from clean images to the early-stage generated structure, enabling better color and texture alignment and providing strong structural priors for refined dehazing in subsequent stages.

LHD achieves state-of-the-art image restoration quality and downstream task adaptability on several real-world haze datasets such as RESIDE and RTTS. Compared to traditional physics-based or adversarial training approaches, LHD preserves more structural and detailed information while restoring naturalness, making it particularly suitable for perception-critical tasks such as detection and segmentation.

In this study, LHD is employed to preprocess hazy images in the VisDrone dataset, generating dehazed inputs with rich details and structural consistency. This significantly improves the localization accuracy and recall of the downstream detection network under haze conditions. This module will be further evaluated in the following experimental section as a critical preprocessing mechanism in the performance validation under complex scenes.

### *2.3. Illumination Enhancement under Extreme Low-Light Conditions*

 doi:10.20944/preprints202512.1468.v1



In complex UAV-based surveillance and object detection scenarios, extreme low-light imaging often results in severely reduced brightness, insufficient contrast, and color distortion. These degradation effects lead to blurred target boundaries and the loss of texture details, directly impairing the performance of downstream detection models—especially in tasks involving small objects and low-contrast defects.

To tackle these challenges, various illumination enhancement techniques have been developed. Liu et al. [25] emphasized the importance of systematically evaluating different enhancement strategies to understand their respective strengths and limitations under low-light conditions. Wang et al. [26] proposed the Progressive Recursive Inference Network (PRIEN), which adopts a dual-attention mechanism for global feature extraction and enhances low-light images in an end-to-end fashion. Lu et al. [27] introduced a dual-branch exposure fusion network that simulates the degradation process of low-light images and effectively restores visibility by estimating illumination transfer functions at different brightness levels.

Based on structural priors, Guo et al. [28] proposed GLNet, which leverages grayscale-channel guidance and dense residual connections to recover fine-grained textures. Yang et al. [29] applied a deep color consistency network to address color fidelity issues, ensuring the enhanced output retains natural visual appearance. Yi et al. [30] introduced Diff-Retinex, which redefines the illumination enhancement task using a physically grounded Retinex decomposition framework, modeling it as a conditional diffusion process. Hou et al. [31] further incorporated global structure-aware diffusion dynamics and uncertainty-guided regularization mechanisms, enabling the enhancement model to achieve higher robustness in extreme lighting conditions. Wang et al. [32] proposed a joint optimization framework that integrates dehazing and illumination enhancement to improve detection performance in traffic-heavy scenarios involving haze and low-light conditions.

In this study, the HVI-CIDNet algorithm is introduced. The network consists of two core mechanisms:(1) A channel interaction and decomposition module, capable of distinguishing global illumination trends from local anomalies;(2) A high-variance compensation module, which adaptively suppresses overexposed and underexposed regions. The compensation process can be formalized as:

$$I_{enh}(x) = I(x) \cdot \alpha(x) + \beta(x) \tag{12}$$

Where $I(x)$ denotes the input degraded image, $\alpha(x)$ is the spatially varying gain factor obtained through channel interaction, and $\beta(x)$ represents the adaptive compensation term for high-variance illumination regions.

In addition, HVI-CIDNet adopts a multi-scale feature modeling strategy combined with a residual learning framework, further enhancing the preservation of high-frequency structures such as image edges and textures. During training, a joint loss function—comprising illumination balance and structural preservation—is employed to guide the network toward improving image clarity while minimizing detail loss and edge blurring.

In summary, HVI-CIDNet demonstrates strong stability and generalization ability in handling images with non-uniform illumination. The enhanced outputs effectively reduce the impact of glare and illumination-induced structural distortions, providing clearer and structurally consistent image inputs for downstream tasks such as object detection and defect recognition.

### *2.4. Defect Detection*

In the feature extraction stage, this paper adopts HgNetV2, a lightweight and efficient feature extraction network, as the backbone to obtain a multi-scale hierarchical semantic feature set {c_2, c_3, c_4, c_5}. To further enhance the representational capability of shallow features—especially c_2—for small object detection, we design a Channel-Aware Projection Refinement (CAPR) module. This module consists of a deep convolution (3×3) and pointwise convolution (1×1), with adaptive fusion of input and output achieved through residual connections. Its computation can be formalized as:

$$F_{CAPR} = F_{in} + PwConv\big(DwConv(F_{in})\big) \tag{13}$$

 doi:10.20944/preprints202512.1468.v1



Where $F_{in}$ denotes the input feature map. This design effectively enhances inter-channel correlation and suppresses redundant information in shallow features while maintaining low computational complexity. As a result, it preserves critical textures such as vehicle contours, laying a solid foundation for detecting small distant objects.

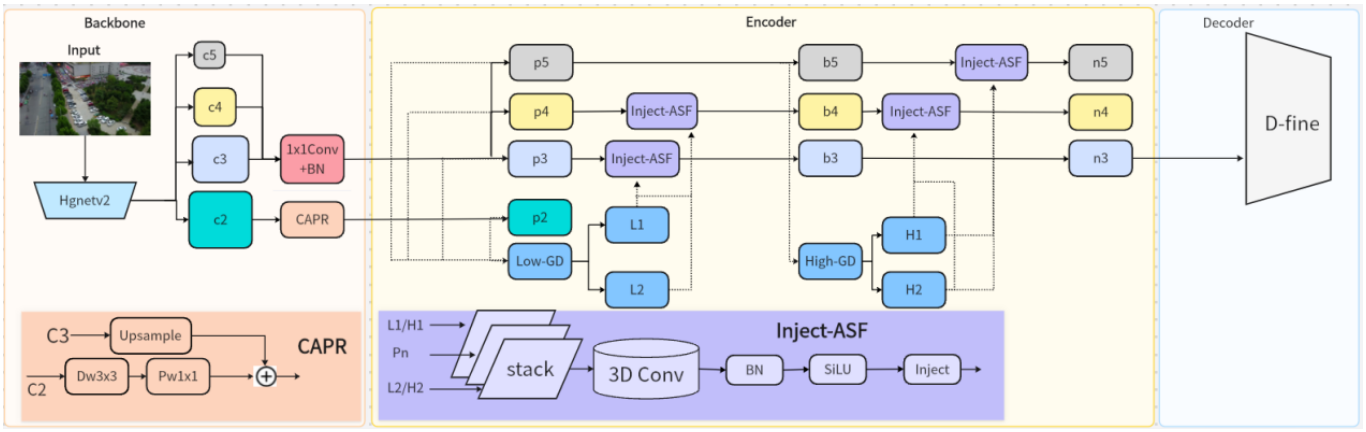


**Figure 2.** GDEIM-SF Network Architecture.

During the encoding stage, high-level semantic features from {c_3, c_4, c_5} are passed into a multi-branch fusion path, where feature enhancement operations are performed in different submodules. First, shallow feature p_2 and mid-level features {p_3, p_4, p_5} are fed into the Low-GD (Low-frequency Guided Decoder) module, which generates two auxiliary output branches, L_1 and L_2 ,to model the edge structure in low-texture regions. The computation is given by:

$$F_{L_i} = \mathrm{LowGD}\left(F_{p2},\ F_{p3},\ F_{p4},\ F_{p5}\right),\quad i \in \{1,2\} \tag{14}$$

At the same time, high-level features {b_3, b_4, b_5} are input to the High-GD (High-frequency Guided Decoder) module to capture global context and high-level semantic consistency, producing two output branches, H_1 and H_2:

$$F_{H_i} = \mathrm{HighGD}(F_{b3},\ F_{b4},\ F_{b5}),\quad i \in \{1,2\} \tag{15}$$

To improve hierarchical consistency during the feature fusion process, we design the Inject-ASF (Attention-guided Spatiotemporal Fusion Injection) module to inject cross-scale contextual dependencies between feature layers. This module performs stacking and 3D convolution operations on input feature sequences (e.g.,{p_3, p_4} or {b_4, b_5}), computed as:

$$\widetilde{B}_n = \mathrm{Inject}\left(\sigma\left(\mathrm{BN}\left(\mathrm{Conv3D}\left(\mathrm{stack}(P_n, L_1, L_2)\right)\right)\right)\right) \tag{16}$$

$$\widetilde{N}_n = \mathrm{Inject}\left(\sigma\left(\mathrm{BN}\left(\mathrm{Conv3D}\left(\mathrm{stack}(B_n, H_1, H_2)\right)\right)\right)\right) \tag{17}$$

Where $\sigma()$ denotes the SiLU activation function, and stack() represents cross-scale stacking. Inject-ASF effectively enhances contextual consistency across feature hierarchies, improving the cross-scale modeling of edge and texture regions.

In the decoding stage, we introduce the D-fine module to replace traditional anchor-based detection heads. Using {n_3, n_4, n_5} as the input feature set, D-fine achieves dynamic modeling and anchor-free bounding box regression. Its prediction mechanism is formulated as:

$$Y = \mathcal{D}(n3, n4, n5; \Theta_{D-\mathrm{fine}}) \tag{18}$$

where $\mathcal{D}()$ denotes the decoding function and $\Theta_{D-\mathrm{fine}}$ represents the learnable parameter set. D-fine incorporates global contextual information and multi-scale saliency responses to significantly improve classification and bounding box regression accuracy, especially in challenging traffic scenarios involving occlusion, distant small targets, and dense objects.

 doi:10.20944/preprints202512.1468.v1



In summary, the proposed architecture enhances channel representation in the Backbone through the CAPR module, establishes cross-layer semantic consistency in the Encoder via Inject-ASF, and performs anchor-free fine detection in the Decoder using D-fine. The synergy of these three components significantly improves the model's robustness and accuracy in detecting distant, occluded, and illumination-variant targets under complex traffic environments, while maintaining high inference efficiency.

## 3. Experiment

To comprehensively evaluate the effectiveness of the proposed "Image Enhancement + GDEIM-SF Detection Backbone" framework under complex imaging conditions, this study conducts a systematic analysis from two perspectives: image quality assessment and object detection performance.

The image quality evaluation focuses on the structural preservation and error control capabilities between the enhanced image and the reference image. Two widely used objective metrics are selected: Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR).

SSIM measures the consistency between the enhanced image and the reference image in terms of luminance, contrast, and structural information. Compared to traditional error-based pixel-wise comparison methods, SSIM emphasizes structural consistency as perceived by the human visual system. It is defined as:

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \tag{19}$$

Where $\mu_x$, $\mu_y$ are the means of images x and y ; $\sigma_x^2$,$\sigma_y^2$ are their variances; $\sigma_{xy}$ is the covariance between them; $C_1$,$C_2$ are constants to avoid division by zero. SSIM ranges from [0,1], with values closer to 1 indicating higher structural similarity between the two images.

PSNR is a distortion metric based on Mean Squared Error (MSE), primarily used to measure pixel-level reconstruction errors. It is calculated as follows:

$$\mathrm{MSE}=\frac{1}{N}\sum_{i=1}^{N}\left(I_i-\hat{I}_i\right)^2 \tag{20}$$

$$\mathrm{PSNR}=10\log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right) \tag{21}$$

Where $I_i$ and $\hat{I}_i$ denote the pixel values of the reference and enhanced images respectively, and N is the total number of pixels. MAX represents the maximum possible pixel value (typically 255).The unit of PSNR is decibels (dB); higher PSNR values indicate lower distortion and better image quality.

To quantitatively assess the performance of the GDEIM-SF framework in object detection tasks, the following four evaluation metrics are employed:Recall (r),Precision (p),Average Precision (AP),mean Average Precision (mAP).

Recall measures the proportion of true objects correctly detected by the model, while Precision reflects the proportion of correct detections among all detected results. These two metrics are generally complementary. To provide a more comprehensive evaluation of detection performance, AP and mAP are introduced as composite indicators.The corresponding mathematical definitions are as follows:

$$\mathrm{Recall}=\frac{TP}{TP+FN} \tag{22}$$

$$\mathrm{Precision}=\frac{TP}{TP+FP} \tag{23}$$

$$\mathrm{AP}=\int_0^1 p\,(r)\,dr \tag{24}$$

$$\mathrm{mAP}=\frac{1}{C}\sum_{c=1}^{C}\mathrm{AP}_c \tag{25}$$

Where TP, FP, and FN represent True Positives, False Positives, and False Negatives, respectively. C is the number of object classes, and $\mathrm{AP}_c$ denotes the average precision for class c.

In practical evaluation, the following two criteria are used for performance comparison:mAP@50: IoU ≥ 0.5 is considered a successful detection. This is a relatively lenient

 doi:10.20944/preprints202512.1468.v1



evaluation standard;mAP@50:95: The average is computed over IoU thresholds ranging from 0.5 to 0.95 in steps of 0.05. This is a stricter and more comprehensive metric, and serves as the primary evaluation index in this study.

### *3.1. Image Dehazing*

To further validate the adaptability and restoration capability of the Learning Hazing to Dehazing (LHD) method in real-world urban traffic surveillance scenarios, typical urban road and parking lot images were selected from the VisDrone dataset as test samples for dehazing experiments.Figure 3 presents a visual comparison of the images before and after the dehazing process. As shown, the original images are severely affected by haze, exhibiting a noticeable gray-white veiling effect that significantly reduces color contrast. The textures of building façades and vegetation areas become blurred, while critical structural boundaries such as vehicle contours and road markings are indistinct—greatly compromising the stability and accuracy of subsequent object detection tasks.

After processing with the LHD method, the brightness, color saturation, and sharpness of the images are significantly improved. Edge contours appear sharper, and fine structural details such as vehicles and crosswalks are effectively reconstructed. The texture layers in buildings and background regions are also enhanced. The visual appearance becomes more similar to that of haze-free reference images, indicating that LHD performs well in restoring realistic scene structure and detail.

**Figure 3.** Dehazing result comparison based on the LHD method.

As shown in Table 1, the LHD method significantly outperforms existing mainstream methods such as Dehamer and AOD-Net in two key objective metrics for dehazing tasks—PSNR and SSIM. Specifically, LHD achieves 25.30 dB PSNR and 0.9265 SSIM, which are markedly higher than those of the comparison models. This demonstrates its superior capability in structural restoration and texture detail reconstruction.

In contrast, Dehamer and AOD-Net lag behind LHD by 0.0030 and 0.0352 in SSIM, respectively, revealing their deficiencies in edge reconstruction and luminance consistency. Particularly in the context of real-world traffic surveillance images characterized by dense textures and complex background interference, both methods show insufficient robustness in structural fidelity and color preservation.

The superior performance of LHD in this task is attributed to its diffusion model-based structural prior guidance mechanism and statistical alignment operation (AlignOp), which effectively mitigate the degradation of edge sharpness and color contrast caused by haze. These mechanisms enhance the overall depth and spatial perception quality of the image.

Furthermore, LHD's combined advantages in structural restoration and visual naturalness greatly expand its practical application potential in urban traffic surveillance, intelligent driving, and edge vision perception systems. The high-quality dehazing output not only improves the readability of the images themselves but also provides clearer and more stable feature inputs for subsequent high-level vision tasks such as object detection and semantic segmentation. This significantly enhances system robustness and operational safety in complex environments such as haze and low visibility.

 doi:10.20944/preprints202512.1468.v1



**Table 1.** Comparison of dehazing algorithm evaluation metrics.

| Task | Algorithm | PSNR(dB) ↑ | SSIM ↑ |
|---|---|---|---|
| Dehaze | LHD | 25.3 | 0.9265 |
| | Dehanmer | 20.153 | 0.9235 |
| | Aod-net | 17.682 | 0.8913 |

*3.2. Image Illumination Enhancement*

To validate the performance of the proposed HVI-CIDNet algorithm in enhancing low-light urban scenes, representative nighttime road surveillance images were selected from the VisDrone dataset for low-illumination enhancement experiments.Figure 4 illustrates the visual comparison before and after enhancement. As observed, the original images exhibit significantly low overall brightness due to poor lighting conditions, with a prominent dark-gray tone. Key regions such as road surfaces, vehicles, and vegetation suffer from severe visibility degradation, texture blurring, and local overexposure or underexposure. Such image quality deterioration greatly constrains the performance of downstream modules for object detection and recognition.

After enhancement by the HVI-CIDNet algorithm, the overall brightness of the images is significantly improved. Especially in non-uniform lighting and dark regions, detail restoration is markedly enhanced. Key structures such as vehicle contours, lane markings, and building edges are effectively reconstructed. Brightness and color balance are noticeably improved, contributing to a more natural and realistic visual appearance.The channel-interaction modeling and high-variance compensation mechanisms of HVI-CIDNet effectively alleviate structural interference caused by non-uniform illumination, improving the perceptual readability of low-light images in dense urban traffic scenarios.

**Figure 4.** Visual comparison of illumination enhancement using HVI-CIDNet.

As shown in Table 2, HVI-CIDNet demonstrates superior performance across all metrics for low-light image enhancement. It achieves the best results in both PSNR and SSIM, reaching 27.35 dB and 0.9165, respectively—significantly outperforming state-of-the-art methods such as RetinexNet and Zero-DCE. HVI-CIDNet achieves breakthroughs in both structural fidelity and illumination restoration.In comparison, although RetinexNet shows certain advantages in improving overall brightness, it suffers from substantial detail loss during enhancement, resulting in an SSIM of only 0.743 and a PSNR of just 20.15 dB, which falls short of meeting high-precision demands for edge and texture preservation in object detection. While Zero-DCE shows a slightly higher PSNR (21.37 dB), its SSIM remains at 0.846, indicating inadequate performance in texture retention and contrast control.

HVI-CIDNet introduces a channel-aware mechanism and high-variance compensation strategy, effectively suppressing overexposure in bright regions and noise in dark areas. It expands the dynamic range while preserving edge sharpness. The enhanced output exhibits global luminance consistency and local detail integrity, significantly improving perceptual quality while maintaining naturalness.In conclusion, HVI-CIDNet provides a high-reliability, low-distortion image preprocessing capability for nighttime traffic surveillance, intelligent driving, and edge vision systems. It significantly boosts the robustness and accuracy of downstream tasks such as object

 doi:10.20944/preprints202512.1468.v1



detection, trajectory analysis, and behavior recognition under extreme illumination conditions, offering strong practical deployment value and scalability.

With its dual advantage in suppressing overexposure in bright zones and noise in dark zones, HVI-CIDNet effectively extends the dynamic range while preserving fine structures. It serves as a reliable and low-distortion preprocessing foundation for high-frequency visual scenarios such as nighttime traffic monitoring and autonomous driving, significantly improving the robustness and accuracy of downstream object recognition and behavior analysis systems in extreme illumination environments.

**Table 2.** Quantitative comparison of illumination enhancement algorithms.

| Task | Algorithm | PSNR(dB) ↑ | SSIM ↑ |
|---|---|---|---|
| Low-light Image Enhancement | HVI-CIDNet | 27.35 | 0.9165 |
| | Retinexnet | 20.15 | 0.743 |
| | Zero-dce | 21.37 | 0.846 |

*3.3 Training Details and Evaluation Protocol*

In the actual training process, the proposed GDEIM-SF model was trained using an NVIDIA RTX 4090 GPU environment. To balance convergence efficiency and generalization capability, a set of fine-tuned training hyperparameters was adopted, as detailed in Table 3.

The model input size was set to 640×640, and training was conducted over 160 epochs to fully leverage the feature representation capability at high resolutions. Taking into account the trade-off between memory constraints and throughput, the batch size was set to 8, while an equivalent batch size of 16 was achieved using 2× gradient accumulation, thereby benefiting from improved gradient stability associated with large batches.

The optimizer used was AdamW, with parameters $\beta_1=0.9$, $\beta_2=0.98$, and a weight decay factor of 0.05, in line with the normalization and attention-heavy architecture of deep networks. The initial learning rate was set to 0.0002, and scaled in proportion to the batch size based on the Linear Scaling Rule.For learning rate scheduling, a hybrid strategy was employed: Linear Warm-up for the first 3 epochs, followed by Cosine Annealing to enable smooth convergence in early stages and progressive fine-tuning later. To further enhance training stability and final accuracy, the Exponential Moving Average (EMA) technique was incorporated, with a decay coefficient of 0.9998. Additionally, Automatic Mixed Precision (AMP) training was enabled to improve memory efficiency and throughput.

Regarding data augmentation, a structure-preserving-first strategy was followed. Mosaic augmentation was moderately applied with a probability $p=0.5p = 0.5p=0.5$. Other techniques included RandomScale in the range （0.5–1.5）, multi-scale training via dynamic sampling from {896, 960, 1024, 1088, 1152}, and lightweight affine transformations (rotation ±5°, translation ≤0.1). For color space augmentation, HSV jitter and mild ColorJitter were used, along with MixUp = 0.1 to improve sample diversity.The above configuration enables a faster optimization process while maintaining high precision and robustness, particularly for small objects and boundary features, without excessive perturbation.

Figure 5 presents the object detection results of the GDEIM-SF model under typical urban traffic scenarios. The figure includes four test samples from the VisDrone dataset, covering various road structures and traffic density conditions. The model effectively identifies three key categories of traffic participants: Car, Motorcycle, and Pedestrian, using color-coded bounding boxes. The attached confidence scores reflect the reliability of the detections.

From the figure, it is evident that GDEIM-SF maintains stable detection performance under challenging conditions such as multi-scale targets, occlusion, and lighting variations. It accurately detects densely packed vehicles and interspersed motorcycles and pedestrians, demonstrating strong feature representation and scale adaptation capabilities.

In the four test images, the system detected a total of 24 cars, 11 motorcycles, and 1 pedestrian, corresponding to approximately 65%, 30%, and 5% of the total targets, respectively. This detection distribution is consistent with real-world urban traffic compositions, further validating the

 doi:10.20944/preprints202512.1468.v1



robustness and generalization capability of the proposed GDEIM-SF model under complex traffic environments.

**Table 3.** Training parameter configuration for the GDEIM-SF model.

| Category | Parameter | Value |
|---|---|---|
| Input | Image size | 640×640 |
| | Epochs | 160 |
| | Batch size | 8 |
| Optimizer | Optimizer | AdamW |
| | $\beta_1$, $\beta_2$ | $\beta_1$, $\beta_2$ |
| | Weight Decay | 0.05 |
| LR Strategy | Initial LR | $2\times10^{-4}$ |
| | Scheduler | Linear warm-up + cosine decay |
| EMA / AMP | EMA Decay | 0.9998 |
| | Mixed Precision | Enabled (AMP) |
| Data Aug. | Mosaic | p = 0.5 |
| | Random Scale | [0.5–1.5] |
| | Multi-scale Training | 896–1152 |
| | Rotation / Translation | ±5°, ≤0.1 |
| | HSV / Color Jitter | Light perturbation |
| | MixUp | 0.1 |

**Figure 5.** Detection results based on the GDEIM-SF model.

## 4. Discussion

As shown in Table 4, we conduct a systematic comparison between the proposed GDEIM-SF model and several state-of-the-art lightweight object detectors across key dimensions such as detection accuracy, computational complexity, and inference speed.The YOLO series has long been regarded as the benchmark for edge detection tasks due to its well-balanced trade-off between accuracy and real-time performance. However, a detailed examination of the table reveals that existing lightweight variants have not yet achieved an ideal Pareto-optimal solution in terms of model size, speed, and detection performance.

**Table 4.** Ablation study comparing the proposed model with various baseline detectors.

 doi:10.20944/preprints202512.1468.v1



| Models | Precision | Recall | mAP50-90 | map50 | flops | Params×105 | FPS |
|---|---|---|---|---|---|---|---|
| YOLOv8s | 43.6 | 32.9 | 0.173 | 0.307 | 28.5 | 11.13 | 82.1 |
| YOLOv10s | 45.2 | 34.8 | 0.179 | 0.323 | 21.4 | 7.22 | 93.5 |
| YOLOv11s | 44.8 | 35.1 | 0.176 | 0.313 | 21.3 | 9.42 | 90.1 |
| YOLOv12s | 45.7 | 34.9 | 0.176 | 0.312 | 21.2 | 9.23 | 91.3 |
| YOLOX-Tiny | 40.5 | 29.4 | 0.148 | 0.278 | 7.578 | 5.035 | 102.6 |
| RTDETR-R18 | 50.1 | 37.6 | 0.208 | 0.363 | 57 | 19.885 | 60.2 |
| Deim-d-fine-s | 51.3 | 39.2 | 0.219 | 0.394 | 24.8595 | 10.18 | 87.7 |
| Ours | 53.1 | 39.8 | 0.245 | 0.425 | 22.1646 | 10.15 | 88.7 |

Specifically:YOLOv8s improves mAP50–90 to 0.173, but incurs a cost of 28.5 GFLOPs and 1.113M parameters, which remains difficult to deploy on resource-constrained edge devices.The evolution from YOLOv10s to YOLOv12s gradually compresses computational cost to around 21 GFLOPs and increases FPS to the 90 range. However, their detection performance shows marginal returns, with mAP50–90 stagnating at 0.176–0.179.YOLOX-Tiny, with only 0.5035M parameters and 7.578 GFLOPs, achieves 102.6 FPS, but suffers a drop in mAP50–90 to 0.148, showing high miss rates in densely occluded and small-object scenarios.RT-DETR-R18 raises mAP50–90 to 0.208, but at the cost of 57 GFLOPs and nearly 2M parameters, reducing FPS to 60.2, thereby compromising real-time deployment potential.Deim-d-fine-s further improves accuracy to 0.219, but the parameter count exceeds 1M, and its efficiency is insufficient to meet the demands of large-scale deployment.

In contrast, the proposed GDEIM-SF model achieves a significant performance leap:With only 1.015M parameters and 22.16 GFLOPs, it boosts mAP50–90 to 0.245 and mAP50 to 0.425, reaching a level comparable to RT-DETR-R18 (mAP50–90 improved by +0.037), while maintaining a high inference speed of 88.7 FPS—striking an optimal balance between high precision and real-time inference in edge computing environments.

Comparative highlights: Versus YOLOv12s:+6.9 percentage points in mAP50–90, –108K fewer parameters, +3.4 FPS in inference speed Versus YOLOX-Tiny:+65.5% relative gain in mAP50–90, Computational cost increases only by 1.92×, FPS drops by 13.5%, yet still maintains high throughput。

Benefiting from the compact multi-scale attention mechanism and dynamic feature reuse strategy, the proposed model demonstrates outstanding scale robustness across three typical ITS scenarios:High-density vehicles at urban intersections, Occluded pedestrians in mixed pedestrian–vehicle traffic, Small targets under nighttime low illumination。

Notable performance gains include:Pedestrian AP improvements of 4.2–7.1 percentage points, AP gain of 5.8 pp for motorcycles under 20 px。This effectively alleviates the performance degradation often observed in lightweight detectors on extremely small targets.

Overall, the GDEIM-SF framework achieves approximately 39% improvement in detection accuracy at the cost of only ~9% additional computational overhead, showcasing excellent suitability for edge deployment. It can be widely applied in AI surveillance cameras, low-power GPUs, and other constrained devices, offering a high-accuracy, high-throughput, low-latency vision perception solution for next-generation intelligent transportation systems (ITS).

### *4.1. Ablation Study*

To evaluate the individual and combined contributions of the proposed architectural components, we conducted a comprehensive ablation study on the three core modules: Model A (GoldYOLO lightweight backbone), Model B (CAPR), and Model C (ASF). The experimental results are summarized in Table 5.

**Table 5.** Ablation study on key modules of the proposed architecture.

 doi:10.20944/preprints202512.1468.v1



| ModelA | ModelB | ModelC | mAP50-90 | map50 | aps | Params×105 |
|---|---|---|---|---|---|---|
| √ | | | 0.173 | 0.315 | 0.087 | 9.86 |
| √ | √ | | 0.223 | 0.322 | 0.123 | 9.95 |
| √ | | √ | 0.186 | 0.406 | 0.105 | 10.02 |
| √ | √ | √ | 0.245 | 0.425 | 0.128 | 10.15 |

to systematically assess the impact of each module on object detection performance, we use the lightweight backbone as the base model and progressively introduce the CAPR and ASF modules. All experiments were conducted on the VisDrone dataset, and results were analyzed primarily from the perspectives of:Accuracy: mAP@50–90, mAP@50, Small object detection: APS, Model complexity: parameter count。

Starting from the lightweight baseline (Model-A), the network achieves a basic performance of:mAP@50–90 = 0.173, mAP@50 = 0.315, APS = 0.087。with only $9.86 \times 10^5$ parameters, validating the effectiveness and deployment-friendliness of the compact backbone design.

Upon introducing the Channel-Aware Projection Refinement (CAPR) module (Model-B), detection performance improves significantly:mAP@50–90 increases to 0.223 (+28.9%), APS rises to 0.123 (+41.4%), mAP@50 shows a slight increase to 0.322。This result demonstrates that the CAPR module effectively guides shallow high-frequency information to compensate for edge and texture details in mid-to-low pyramid layers, enhancing the discriminability and detectability of small objects in low-texture and occluded regions.

Further incorporating the Scale-Adaptive Fusion (ASF) module (Model-C) leads to a notable boost in mAP@50, reaching 0.406, but causes some performance regression in:mAP@50–90 = 0.186, APS = 0.105。Based on visual analysis of experimental images, we attribute this drop to the frequency-domain prior enhancement mechanism of ASF. While it successfully enlarges the receptive field and strengthens cross-scale consistency, it may also introduce redundant spectral components or boundary alignment errors, which in turn lead to increased localization errors at higher IoU thresholds—especially under occlusion or in dense-object scenes.

Finally, in the complete model (Model-final) where both CAPR and ASF are jointly integrated, the network achieves the best overall performance:mAP@50 = 0.425, mAP@50–90 = 0.245, APS = 0.128, with a parameter count of only $10.15 \times 10^5$。Compared to the baseline, the three core indicators improve by:+34.9% (mAP@50), +41.6% (mAP@50–90), +47.1% (APS)

These results further validate the complementary synergy between CAPR and ASF:CAPR enhances local edge and high-frequency texture modeling, ASF introduces stable long-range dependencies and hierarchical consistency。Their joint integration effectively suppresses the independent error sources introduced by each module, ultimately achieving dual enhancement of detection accuracy and robustness in complex urban traffic environments.

### *4.2 Target Detection with and without Fog and Low Illumination*

To systematically evaluate the contribution of the front-end image enhancement module to the overall detection pipeline, we designed a rigorous paired comparison experiment based on the principle of single-variable control. Except for the inclusion of "dehazing + low-light enhancement" in the preprocessing step, all other experimental variables were held constant across setups. These include: the backbone detection network, decoder architecture, training data partition, optimizer type, learning rate scheduler, and random seed.

Two experimental groups were constructed under this setting:Raw-Input Group: original, unenhanced images as input；Enhanced-Input Group: input images processed via dehazing and illumination enhancement modules。Both groups were trained under identical data augmentation and iteration conditions, eliminating the influence of sample distribution and gradient path variation, ensuring fair comparability.

Figure 6 presents visual detection comparisons in typical nighttime urban road scenes:In the Raw-Input results (left), the overall brightness is extremely low, with distant areas nearly black. Object textures and edges are severely submerged in background noise, vehicle structures are fragmented and blurred, most bounding boxes are concentrated in the foreground with low

 doi:10.20944/preprints202512.1468.v1



confidence scores, and distant vehicles suffer large-scale missed detections；In contrast, the Enhanced-Input results (right) benefit from brightness and contrast enhancements that stretch the grayscale dynamic range. High-frequency features—such as vehicle contours, headlights, and windows—are effectively restored. Bounding boxes become more numerous, evenly distributed, and tightly aligned to actual targets, indicating improved spatial consistency and scale robustness.

**Figure 6.** Detection comparison before and after illumination enhancement.

Figure 7 shows detection performance under heavy haze:The Raw-Input (left) suffers from extensive aerosol scattering, forming low-contrast "milky" regions and low-frequency fog layers. These significantly compress radiometric contrast between objects and background, causing blurred boundaries and lost texture in distant small targets. This results in high miss rates, loose bounding boxes, low confidence, and unstable regression；The Enhanced-Input (right), post dehazing + global brightness restoration, shows clear improvements in contrast and scene depth. Building façades and tree textures are restored, and object boundaries are sharper—especially in medium to long-range views, where contour separation between vehicles and non-motorized objects improves. The detection boxes are tighter, more accurate, and better aligned, with reduced false positives and more uniform confidence scores.

**Figure 7.** Detection comparison before and after dehazing enhancement.

In summary, the front-end image enhancement module provides higher-quality input signals under challenging conditions such as low light and fog. It significantly improves image visibility and structural discernibility, reflected not only in increased brightness and contrast, but also in:Texture recovery, Edge clarity, Spatial hierarchy reconstruction。These enhancements provide a more robust foundation for feature extraction. Compared with the original input, the enhanced images show systemic improvements in three critical dimensions:1.Expanded Detection Coverage: Detection, once limited to the foreground, now extends to medium and long-range targets, enabling multi-scale perception across a wider field of view.2.Improved Box Tightness and Alignment: Bounding boxes are more accurately aligned with object contours, reducing false positives and localization errors.3.Optimized Confidence Distribution: Weak targets exhibit significantly increased

 doi:10.20944/preprints202512.1468.v1



confidence scores, enhancing discriminability for difficult samples while reducing background false detections.Moreover, the dual-stage preprocessing (dehazing + illumination enhancement) ensures that internal modules like CAPR and SAF receive clearer, less redundant features—effectively mitigating classification and regression uncertainties caused by degraded imaging (e.g., blurred edges, incomplete textures).

Overall, image enhancement not only improves perceptual image quality, but also enables front–back synergy within the detection pipeline. It significantly enhances robustness and stability for recognizing distant and low-visibility targets in UAV-based urban object detection—providing solid support for real-world deployment in intelligent traffic perception systems under complex environments.

To rigorously assess the generalization benefits of image enhancement under adverse weather conditions, an ablation study was conducted on the RADIATE dataset [33]. Table 6 presents detection performance under four preprocessing settings:No enhancement；Illumination enhancement；Dehazing enhancement；Joint (two-stage) enhancement。Four key metrics are used: Precision (P), Recall (R), mAP@50, and mAP@50–90.

1. No Enhancement (Baseline):P = 43.5%, R = 21.6%, mAP@50 = 0.373, mAP@50–90 = 0.155。Significant limitations in performance, especially in recall and mAP@50–90, indicating severe loss in perception and localization accuracy under fog/low-light degradation.
2. Illumination Enhancement:P = 47.4% (+3.9 pp), R = 32.9% (+11.3 pp), mAP@50 = 0.382, mAP@50–90 = 0.217 (+40%)。Highlights HVI-CIDNet's effectiveness in restoring dark details and dynamic range, particularly improving recall and structural preservation.
3. Dehazing Enhancement:P = 45.7%, R = 31.8%, mAP@50–90 = 0.201。Slightly lower than illumination enhancement, but better than baseline. Demonstrates LHD's ability to restore edge clarity and contrast lost to haze, especially for small distant targets.
4. Joint Enhancement (Dehazing + Illumination):P = 50.6%, R = 36.4%, mAP@50 = 0.409, mAP@50–90 = 0.233, Achieves the best results across all metrics. Compared to baseline:+7.1 pp (P), +14.8 pp (R), +9.6% (mAP@50), +50.3% (mAP@50–90)

This confirms a strong synergistic effect from the two-stage strategy:Illumination enhancement recovers brightness and fine detail；Dehazing enhancement reinforces structural clarity and edge separability。Together, they significantly improve spatial discriminability and semantic modeling for complex object detection tasks.

**Table 6.** Quantitative performance of different image preprocessing strategies.

| ModelA | Precision | Recall | map50 | mAP50-90 |
|---|---|---|---|---|
| No Enhancement (Baseline) | 43.5 | 21.6 | 0.373 | 0.155 |
| Illumination Enhancement | 47.4 | 32.9 | 0.382 | 0.217 |
| Dehazing Enhancement | 45.7 | 31.8 | 0.381 | 0.201 |
| Joint Enhancement (Dehazing + Illumination) | 50.6 | 36.4 | 0.409 | 0.233 |

## 5. Conclusions

This paper addresses the challenge of small object detection in UAV-based visual perception under adverse imaging conditions such as haze and low illumination. A multi-stage collaborative detection framework integrating image preprocessing and lightweight detection is proposed.

At the preprocessing stage, a cascaded module combining physical modeling and perceptual enhancement is designed to simultaneously improve image visibility and structural detail fidelity. At the detection stage, a lightweight and robust GDEIM-SF backbone is constructed, and the integration of the CAPR module significantly enhances the modeling of object edges and textures, particularly

 doi:10.20944/preprints202512.1468.v1



for small targets. Additionally, the Scale-Adaptive Fusion (SAF) module is introduced, which leverages frequency-domain priors to improve feature consistency and discriminability for small objects.

Experiments on multiple representative urban traffic scene datasets demonstrate that the proposed method achieves notable improvements in detection accuracy and structural consistency under complex imaging conditions, while maintaining inference efficiency and lightweight deployment adaptability. In particular, the proposed method shows significant gains in both mAP@50 and mAP@50–90 compared with other models, indicating strong robustness against multi-scale variations, dense occlusions, and low-quality images.Future work will explore end-to-end joint optimization mechanisms between the preprocessing and detection stages, as well as the generalization capability of the framework in broader low-altitude intelligent scenarios.

## References


1. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In CVPR 2016 (pp. 779–788). IEEE. https://doi.org/10.1109/CVPR.2016.91
2. Yin, D., Wan, M., Qian, W., Xu, Y., Kong, X., et al. (2025). SGA-YOLO: A Lightweight Real-Time Object Detection Network for UAV Infrared Images. IEEE Transactions on Instrumentation and Measurement. https://ieeexplore.ieee.org/document/11197501
3. Chen, J., Lai, D., Kang, K., Xu, K., Ma, X., et al. (2025). Enhancing UAV Object Detection with an Efficient Multi-Scale Feature Fusion Framework. PLOS ONE, 20(10), e0332408. https://doi.org/10.1371/journal.pone.0332408
4. Wang, J., Gao, M., Zhang, Z., Zhao, H., Li, W., et al. (2025). REA-YOLO for Small Object Detection in UAV Aerial Images. The Journal of Supercomputing. https://doi.org/10.1007/s11227-025-07836-0
5. Nikouei, M., Baroutian, B., Nabavi, S., Taraghi, F., & Aghaei, A. (2025). Small Object Detection: A Comprehensive Survey on Challenges, Techniques and Real-World Applications. ResearchGate Preprint. https://www.researchgate.net/publication/394126205
6. Li, J., Sun, L., Lin, C., Wang, X., Wang, Z., et al. (2025). SED-UAV: A Synergistic Framework of Lightweight Chaotic Encryption and Multi-Scale Feature Detection for Secure UAV Applications. IEEE Internet of Things Journal. https://ieeexplore.ieee.org/document/11153637
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. ECCV 2020, 213–229. https://arxiv.org/abs/2005.12872
8. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable DETR: Deformable Transformers for End-to-End Object Detection. ICLR 2021. https://arxiv.org/abs/2010.04159
9. Han, K., Sun, B., Zhang, X., et al. (2023). Small Object Detection from UAV Images Using Deformable Transformer Networks. IEEE Access, 11, 84521–84534.
10. Li, S. (2024). RT-DETR: Real-Time End-to-End Object Detection with Accelerated Training. arXiv preprint arXiv:2304.03766.
11. Chen, Q., et al. (2024). Freq-DETR: Frequency-Aware Transformer for High-Fidelity Object Detection. IEEE Transactions on Image Processing. [DOI pending publication]
12. Han, K., et al. (2025). CAEM-DETR: Contrastive Attention and Multi-Domain Fusion for Complex UAV Scene Detection. Scientific Reports, 15, Article 18881. https://www.nature.com/articles/s41598-025-18881-3
13. Wang, R., Zheng, Y., Zhang, Z., Li, C., Liu, S., Zhai, G., & Liu, X. (2025). Learning hazing to dehazing: Towards realistic haze generation for real-world image dehazing. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 23091-23100).
14. Yan, Q., Shi, K., Feng, Y., Hu, T., Wu, P., Pang, G., & Zhang, Y. (2025). HVI-CIDNet+: Beyond Extreme Darkness for Low-Light Image Enhancement. arXiv preprint arXiv:2507.06814.
15. Wang, C., He, W., Nie, Y., Guo, J., Liu, C., Wang, Y., & Han, K. (2023). Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. Advances in Neural Information Processing Systems, 36, 51094-51112.
16. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Ling, H. (2018). VisDrone-DET2018: The Vision Meets Drone Object Detection in Image Challenge Results. ECCV Workshops 2018.

17. He, K.; Sun, J.; Tang, X. (2011). Single Image Haze Removal Using Dark Channel Prior. IEEE TPAMI, 33(12), 2341–2353. https://doi.org/10.1109/TPAMI.2010.168
18. Zhu, Q.; Mai, J.; Shao, L. (2015). A Fast Single Image Haze Removal Algorithm Using Color Attenuation Prior. IEEE TIP, 24(11), 3522–3533. https://doi.org/10.1109/TIP.2015.2446191
19. Berman, D.; Avidan, S. (2016). Non-local Image Dehazing. CVPR 2016, 1674–1682. https://doi.org/10.1109/CVPR.2016.184
20. Cai, B.; Xu, X.; Jia, K.; Qing, C.; Tao, D. (2016). DehazeNet: An End-to-End System for Single Image Haze Removal. IEEE TIP, 25(11), 5187–5198. https://doi.org/10.1109/TIP.2016.2598681
21. Qu, Y.; Chen, Y.; Huang, J.; Xie, Y. (2019). Enhanced Pix2pix Dehazing Network. CVPR 2019, 8160–8168.
22. T. Gao, Y. Liu, P. Cheng, T. Chen and L. Liu, "Multi-Scale Density-Aware Network for Single Image Dehazing," in IEEE Signal Processing Letters, vol. 30, pp. 1117-1121, 2023, doi: 10.1109/LSP.2023.3304540.
23. D. Engin, A. Genc and H. K. Ekenel, "Cycle-Dehaze: Enhanced CycleGAN for Single Image Dehazing," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 2018, pp. 938-9388, doi: 10.1109/CVPRW.2018.00127.
24. J. Liu, H. Wu, Y. Xie, Y. Qu and L. Ma, "Trident Dehazing Network," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 1732-1741, doi: 10.1109/CVPRW50498.2020.00223.
25. Liu, W., Tang, H., Hu, X., Lin, Y., & Zhou, S. (2025). A Lightweight Multi-Stage Visual Detection Approach for Complex Traffic Scenes. Sensors, 25(16), 5014. https://doi.org/10.3390/s25165014
26. Wang, Y., Cao, Y., Zha, Z. J., Zhang, J., Xiong, Z., Zhang, W., & Wu, F. (2019, October). Progressive retinex: Mutually reinforced illumination-noise perception network for low-light image enhancement. In Proceedings of the 27th ACM international conference on multimedia (pp. 2015-2023).
27. K. Lu and L. Zhang, "TBEFN: A Two-Branch Exposure-Fusion Network for Low-Light Image Enhancement," in IEEE Transactions on Multimedia, vol. 23, pp. 4093-4105, 2021, doi: 10.1109/TMM.2020.3037526. keywords: {Lighting;Noise reduction;Image enhancement;Estimation;Image reconstruction;Image color
28. C. Guo et al., "Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 1777-1786, doi: 10.1109/CVPR42600.2020.00185.
29. Yang, X., Chen, J., & Yang, Z. (2025). Learning physics-informed color-aware transforms for low-light image enhancement. arXiv preprint arXiv:2504.11896.
30. Yi, X., Xu, H., Zhang, H., Tang, L., & Ma, J. (2023). Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 12302-12311).
31. Hou, J., Zhu, Z., Hou, J., Liu, H., Zeng, H., & Yuan, H. (2023). Global structure-aware diffusion process for low-light image enhancement. Advances in Neural Information Processing Systems, 36, 79734-79747.
32. Wang K, Zhao Y (2025) Improving object detection in challenging weather for autonomous driving via adversarial image translation. PLOS ONE 20(10): e0333928. https://doi.org/10.1371/journal.pone.0333928
33. Sheeny, M., De Pellegrin, E., Mukherjee, S., & Newman, P. (2021). RADIATE: A Radar Dataset for Automotive Perception in Bad Weather. IEEE Transactions on Robotics.