remote sensing

MDPI

Article

DEAF-Net: Dual-Domain Enhanced Adaptive Fusion Network for UAV Visible–Infrared Object Detection

Qian Weng [1,2,3,*], Yu Zhang [1], Xiansheng Huang [1], Liming Deng [4] and Jiawen Lin [1,2,3]

[1] College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China; 241027056@fzu.edu.cn (Y.Z.); 241020035@fzu.edu.cn (X.H.), ljw@fzu.edu.cn (J.L.)
[2] Engineering Research Center of Big Data Intelligence, Ministry of Education, Fuzhou 350116, China
[3] Fujian Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou University, Fuzhou 350116, China
[4] Graduate College for Engineers, Fuzhou University, Fuzhou 350108, China; denglm@fzu.edu.cn
* Correspondence: fzuwq@fzu.edu.cn

Highlights

What are the main findings?

- The DDPR jointly enhances features in frequency and spatial domains, effectively suppressing noise and improving texture details.
- The CDGF adaptively fuses features by modeling cross-modal consistency and discrepancy, guided by local spatial reliability and global entropy-based scene priors.

What are the implications of the main findings?

- Spatial-frequency collaborative enhancement prior to adaptive fusion establishes purified features for cross-modal integration, elevating the upper bound of multimodal fusion performance.
- DEAF-Net enables robust all-weather UAV object detection, reducing false positives and missed detections in challenging environments.

Abstract

In Unmanned Aerial Vehicle (UAV) object detection tasks, complex lighting conditions and variable weather render robust all-weather perception challenging when relying solely on the visible modality. Although infrared modalities can provide complementary information, the reliability of individual modalities is highly scene-dependent. Existing multimodal detection methods typically adopt static fusion strategies, which ignore spatial heterogeneity of modal reliability and under-explore spatial-frequency collaborative representation, thus limiting detection robustness in dynamic environments. To address these issues, this paper proposes a Dual-domain Enhanced Adaptive Fusion Network (DEAF-Net), with two core innovative modules to tackle the above challenges. First, the Dual Domain Progressive Refinement (DDPR) module mitigates feature degradation caused by poor imaging conditions via the joint design of frequency-domain learnable filtering and scale-aware contextual refinement in the spatial domain, effectively suppressing noise, enhancing textures, and yielding a purified feature basis for fusion. Second, the Consistency–Discrepancy Guided Fusion (CDGF) strategy leverages the selective scanning mechanism of VMamba to model consistent and differential patterns across modalities, dynamically generates local modal contribution maps for adaptive fusion, and integrates global scene prior via entropy weights for calibration. Extensive experiments on the DroneVehicle and VEDAI datasets show that DEAF-Net outperforms mainstream multimodal detection methods, achieving mAP@0.5 scores of 81.9% and 76.2%, respectively, while delivering improved robustness in low-light, dense fog, and sparse-category scenarios.

Check for updates

Academic Editors: Xian Li and Aleksandra Pizurica

Received: 7 May 2026
Revised: 18 June 2026
Accepted: 29 June 2026
Published: 7 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: UAV visible–infrared object detection; dual-domain enhancement; VMamba; adaptive cross-modal feature fusion

1. Introduction

With the continuous advances in low-altitude remote sensing technology and the growing maturity of UAV platforms, UAV object detection has become a significant research focus in intelligent perception. Visible images (RGB) are extensively used in various UAV vision tasks because of their abundant texture detail and structural information. However, in complex environments such as those with low illumination, strong shadows, target occlusion, or adverse weather, the imaging quality of visible light significantly deteriorates, reducing target discernibility and severely compromising the robustness and adaptability of detection models. In practical applications like traffic monitoring [1], power line inspections [2], personnel search and rescue [3], and military reconnaissance [4], tasks often demand models with stable perception capabilities across all times and diverse environments to enable efficient responses and accurate judgments during emergencies, particularly under extreme conditions such as nighttime, smog, rain, or snow. However, single-modal detection methods relying solely on visible light images have difficulty meeting practical demands. To enhance the overall performance and environmental adaptability of object detection algorithms for UAVs in complex scenes, there is an urgent need to introduce more robust multimodal perception methods.

To this end, the infrared (IR) modality has become an effective complementary modality. As shown in Figure 1a, under extremely low-light conditions, the visible modality severely degrades, making the IR modality the dominant and reliable information source. However, the lack of color information and fine-grained textures in the IR modality leads to semantic ambiguity, while in locally illuminated areas, the visible modality can provide richer semantic details than IR. Figure 1b further demonstrates that relying solely on the IR modality makes it difficult to distinguish between a truck (red box) and a metal shed (yellow box) with similar contours, whereas the semantically rich RGB modality can provide crucial complementary information. Moreover, as illustrated in Figure 1c, the IR modality may introduce artifacts due to its imaging mechanism. These observations reveal a fundamental challenge in visible–infrared object detection: in complex and ever-changing scenes, there is no globally optimal or consistently reliable single modality. Instead, the reliability and information value of each modality exhibit spatial heterogeneity, meaning they vary dynamically across different local regions and under different environmental conditions. Although significant progress has been made in the research of cross-modal feature interaction at present, most methods still mainly rely on static fusion strategies, which treat both modalities equally or with fixed fusion weights, fail to fully consider the spatial heterogeneity of modal reliability, and easily lead to limited expressive power and coarse granularity of fused features. When two modalities exhibit significantly different information richness within the same region, this naïve fusion strategy tends to cause negative fusion effects [5].

Furthermore, most of the existing methods focus on feature extraction and enhancement in the spatial domain [6], inadequately accounting for the decline in feature representation ability caused by imaging degradation in harsh environments, and overlooking the role of frequency-domain processing in suppressing background noise and enhancing critical textures [7–9], as illustrated in Figure 2. In fact, the effectiveness of feature fusion fundamentally depends on the quality of the input features themselves. By introducing

frequency-domain analysis, the quality of features can be effectively enhanced [10–12], thereby providing a cleaner and more discriminative foundation for subsequent fusion.

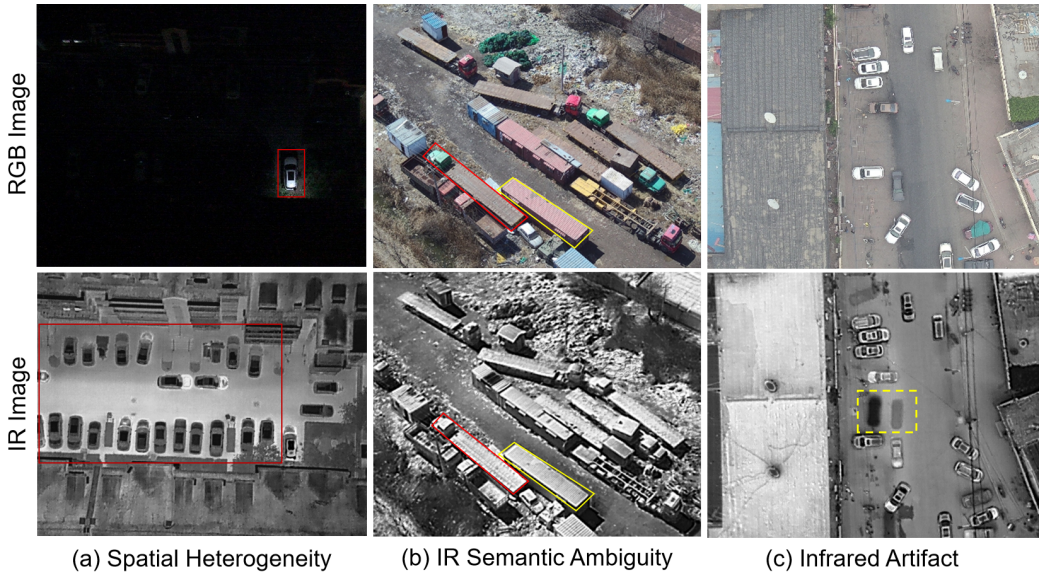


Figure 1. Complementary strengths and limitations of RGB and IR modalities. (**a**) In extremely dark environments, RGB is only effective in locally illuminated areas, while infrared can still maintain the overall visibility of the target. This indicates that there is no globally optimal pattern. (**b**) RGB provides rich edge textures and color information in well-lit environments that resolve semantic ambiguity caused by blurred infrared contours. (**c**) IR may contain artifacts due to its imaging mechanism, which can interfere with detection models.

However, current multimodal methods lack a systematic approach to dynamically perceive modality reliability and obtain modality-complementary fusion features under conditions where input features are degraded by noise or insufficient illumination. Based on the above analysis, we propose the following scientific hypothesis: explicitly modeling the spatial heterogeneity of modality reliability, together with joint frequency-spatial domain feature enhancement, leads to more robust and discriminative multimodal representations for UAV object detection under complex environments. Therefore, we propose a Dual-domain Enhanced and Adaptive Fusion Network (DEAF-Net), which synergistically integrates spatial-frequency domain-aware feature enhancement with a scene-aware and location-adaptive fusion mechanism. Specifically, we first design a Dual Domain Progressive Refinement (DDPR) module that adaptively enhances high-frequency components through learnable soft high-pass filters, thereby filtering out background interference in the visible modality and enhancing texture details in the IR modality. The enhanced features are then processed collaboratively via a dual-path architecture, which simultaneously incorporates frequency-domain channel attention and spatial depth-wise convolution to achieve feature co-optimization and information complementarity across domains, followed by multi-scale feature extraction to capture objects at different scales within UAV scenes.

Building upon these purified and enhanced features, we introduce an adaptive fusion strategy centered on the Consistency–Discrepancy Guided Fusion (CDGF) module. This module explicitly disentangles cross-modal interactions into two distinct signals: collaborative consistency features and complementary discrepancy features. The global dependencies between these two signals are captured via the state-space modeling mechanism of VMamba. Guided by this dual-signal representation, CDGF generates a modality contribution matrix, enabling dynamic spatial confidence weighting based on local reliability. To further improve robustness under extreme conditions, we incorporate an information entropy-based global prior to quantify inter-modal uncertainty, which globally calibrates

the data-driven local weighting mechanism. By unifying dual-domain feature enhancement with adaptive dynamic fusion, DEAF-Net establishes a robust and effective framework for visible–infrared object detection in UAV-based applications.

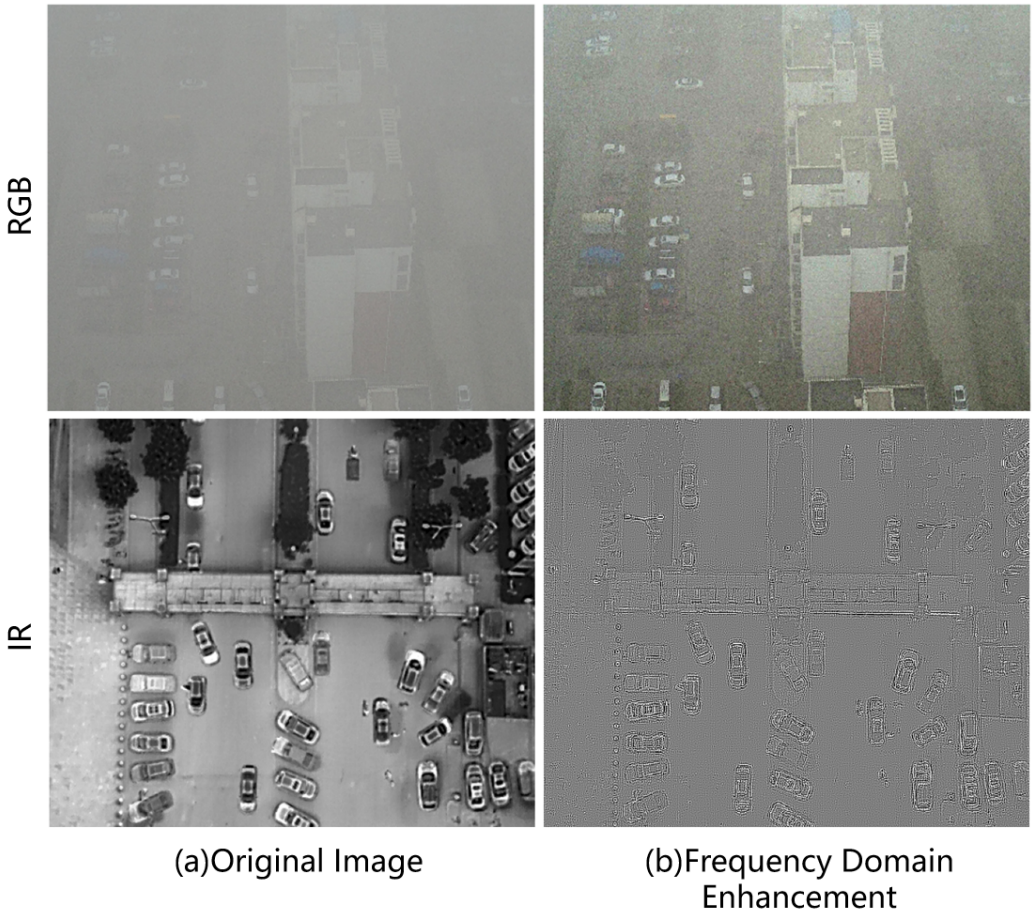


Figure 2. Visual comparison between original images (**a**) and enhanced results using the proposed soft high-pass frequency filter (**b**). For visible images (top row), it primarily suppresses background noise such as haze; for infrared images (bottom row), it enhances their lacking texture and edge details.

Our main contributions can be summarized as follows:

1. We propose the DDPR module that integrates frequency-spatial collaborative enhancement with scale-aware contextual refinement. It mitigates feature degradation under poor imaging conditions through dual-domain enhancement, suppresses background noise, enhances texture details, and captures contextual information to provide a purified and enriched feature basis for subsequent fusion.
2. We introduce the CDGF module, which dynamically fuses complementary multimodal information through locally generated spatial fusion weights guided by dual signals and globally calibrated by information entropy-based statistical priors, effectively addressing the spatial heterogeneity of reliability between visible and infrared modalities in complex scenes.
3. We conduct extensive experiments on DroneVehicle and VEDAI datasets, demonstrating that DEAF-Net outperforms other methods and exhibits effective adaptability to challenging UAV scenes such as low illumination and fog.

2. Related Work

2.1. Visible–Infrared Object Detection

Visible–infrared object detection leverages the complementary strengths of the RGB modality's rich semantic features and the IR modality's stable target perception capability, enabling robust all-weather perception for UAVs in complex scenes.

Early studies primarily adopt static fusion methods, which combine multimodal features through fixed operations such as channel concatenation or element-wise addition. Half Fusion [13] demonstrated that fusing features at intermediate layers can effectively exploit complementary modality information. Xiao et al. [14] further introduced difference enhancement, small-target enhancement, and large-target semantic expansion to improve multi-scale feature representation before adopting channel concatenation for fusion. Despite their simplicity and effectiveness, static fusion methods fail to account for dynamic variations in modality quality, often resulting in coarse fusion representations under complex illumination or adverse imaging conditions.

To overcome these limitations, adaptive weighted fusion methods have been proposed to dynamically adjust the contribution of each modality. UA-CMDet [15] introduced an uncertainty-aware module (UAM) that estimates target uncertainty via cross-modal IoU and illumination cues, thereby reweighting the regression loss to mitigate the negative impact of unreliable samples. GLFNet [16] jointly models local spatial saliency and global channel dependencies, enabling the network to adaptively emphasize RGB or IR features under varying lighting conditions. Although these methods introduce adaptive weighting mechanisms, their reliability modeling is often coarse or target-level, lacking fine-grained spatial awareness of modality reliability.

More recently, deep cross-modal interaction methods leverage Transformer or Mamba architectures to enable richer and more flexible information exchange between modalities. C^2Former [17] employs an Inter-modality Cross-Attention (ICA) module to align and complement RGB and IR features. DMM [18] proposes a Disparity-guided Cross-modal Fusion Mamba (DCFM) module, which models inter-modal differences using state space models to achieve efficient cross-modal feature fusion. However, most existing interaction-based methods remain confined to spatial-domain modeling, overlooking the potential of frequency-domain enhancement for low-quality features. Moreover, they typically employ global or channel-wise fusion weights, neglecting the spatial heterogeneity of modality reliability. In contrast, we propose a dual-domain enhancement strategy that jointly leverages frequency and spatial information, coupled with a spatially adaptive fusion mechanism guided by explicit consistency-discrepancy modeling and scene-level entropy calibration to achieve more robust and discriminative multimodal features.

2.2. Frequency-Domain Learning

The frequency domain provides a complementary perspective for visual representation in object detection tasks. Low-frequency components characterize the global structure of a scene, while high-frequency components encode edges and fine-grained textures, which are crucial for small-object perception and target-background discrimination. HS-FPN [19] adopted Discrete Cosine Transform (DCT) high-pass filtering to enhance small-object features. GLFDet [20] employs the Discrete Wavelet Transform (DWT) for multi-scale frequency decomposition to suppress noise. FMCAF [21] uses Fast Fourier Transform (FFT) feature suppression with cross-modal interactive attention to enhance RGB-IR complementary information for subsequent fusion.

Despite the promising progress of these frequency-domain methods, most focus only on frequency-domain enhancement and lack joint optimization with spatial multi-scale contextual modeling for multimodal UAV detection. Unlike existing frequency-domain

methods that employ hard frequency cutoffs, our DDPR module introduces a learnable soft high-pass filter for modality-adaptive spectral modulation, coupled with spatial-domain contextual refinement to achieve synergistic dual-domain enhancement.

2.3. Mamba Model

Mamba [22] is an efficient sequence modeling framework built upon structured State Space Models (SSM), which enables long-range dependency modeling with linear computational complexity. To extend Mamba to visual perception, VMamba [23] introduces a 2D selective scanning (SS2D) mechanism, which performs quad-directional scanning on feature maps to capture global contextual information efficiently, as shown in Figure 3. Its fundamental building block, the Visual State Space Block (VSSBlock), combines local convolutional modeling with global dependency learning via state space models. RSMamba [24] incorporates a multi-path activation mechanism to achieve non-causal global dependency modeling on 2D remote sensing images, thereby enhancing the representational capacity of spatial relationships. SMILE [25] designs a dual-path attention network based on Mamba's multi-dimensional scanning mechanism to achieve sparse high-frequency feature enhancement for infrared small targets.

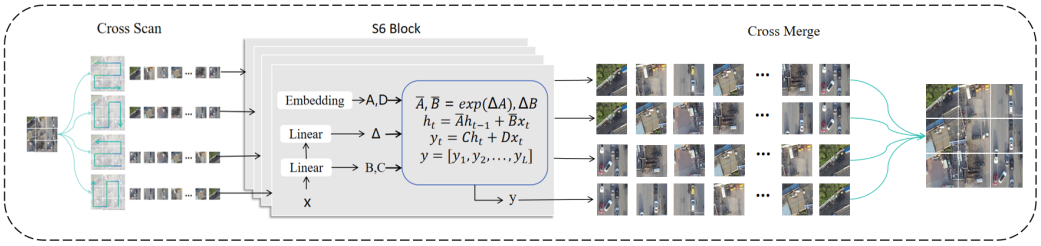


Figure 3. Illustration of the 2D selective scanning (SS2D) mechanism in VMamba.

These studies collectively verify the effectiveness of Mamba-style state space modeling as a general-purpose mechanism for visual feature representation and global context modeling. However, existing explorations are still predominantly limited to single-modal settings or general feature enhancement. The potential of Mamba-based models for explicitly modeling the fine-grained consistency and discrepancy signals in visible–infrared fusion scenarios remains insufficiently investigated.

3. Proposed Method

The overall architecture of the proposed Dual-domain Enhanced Adaptive Fusion Network (DEAF-Net) is shown in Figure 4. First, a dual-stream VMamba [23] is adopted as the backbone, which employs four VSSBlocks to extract multi-level features from visible and infrared images, respectively (stages 1 to 4). To achieve effective cross-modal information fusion, we introduce DEAF blocks at stages 2, 3, and 4 of the backbone, progressively refining and fusing multi-level features. Each DEAF block comprises two collaboratively designed components: the Dual Domain Progressive Refinement (DDPR) module, which enhances modality-specific features via spatial-frequency collaboration, and the Consistency-Discrepancy Guided Fusion (CDGF) module. The CDGF module adaptively fuses visible and infrared features by combining locally adaptive spatial weighting with information entropy-based global prior calibration. Finally, the fused features are fed into the detection head to predict category labels and oriented bounding boxes.

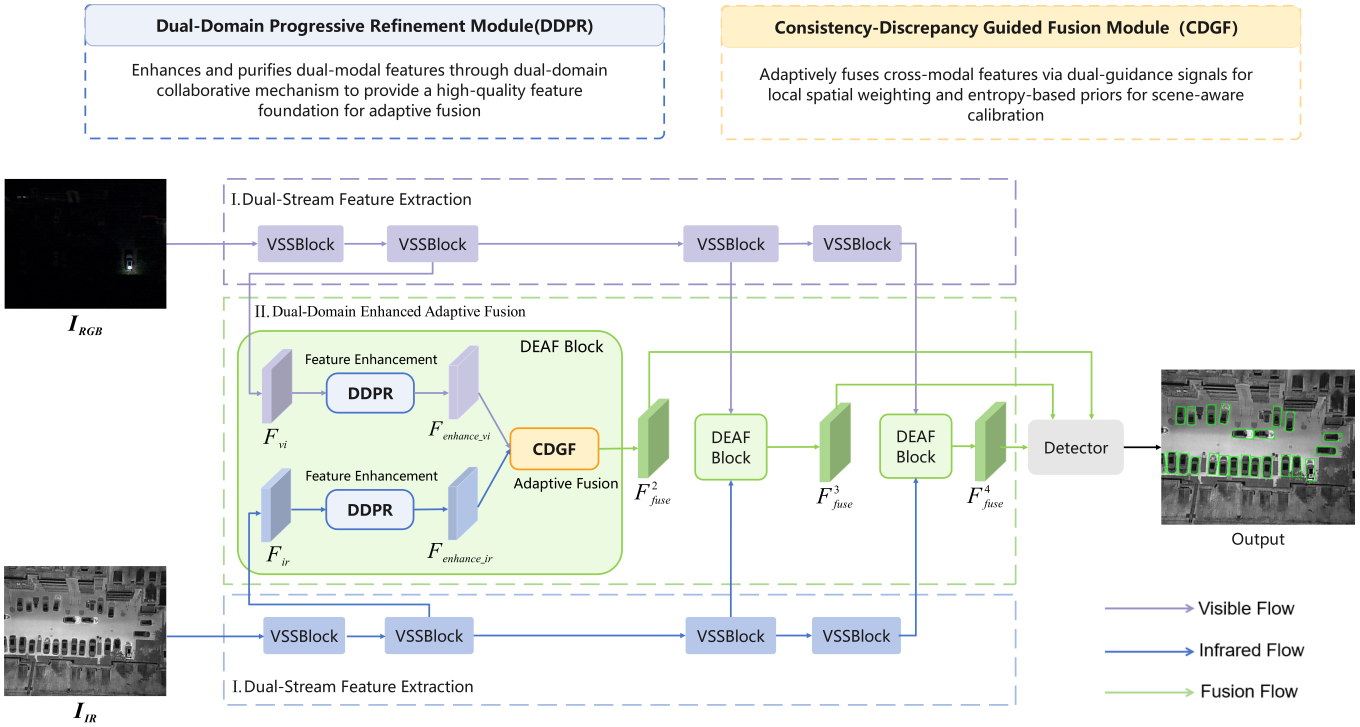


Figure 4. Overall architecture of DEAF-Net. (I) Dual-Stream Feature Extraction, which separately encodes RGB and IR inputs via a VMamba backbone to obtain multi-level preliminary features, while computing global information entropy priors to quantify modality-wise informativeness; (II) Dual-domain Enhanced Adaptive Fusion, which progressively refines features through the DDPR and CDGF modules for frequency-spatial enhancement and adaptive cross-modal fusion.

3.1. Problem Definition

Given the aligned visible image I_{RGB} and infrared image I_{IR}, UAV visible–infrared object detection tries to predict the category labels c and rotated bounding boxes $\mathbf{b} = (x, y, w, h, \theta)$ for all objects, where (x, y) denote the center coordinates of the bounding box, (w, h) denote width and height, and θ is the rotation angle. Two independent VMamba [23] backbones are adopted to encode visible image I_{RGB} and infrared image I_{IR}, respectively. The extracted feature maps at t-layer are expressed as $\mathbf{F}_{vi}^{t}, \mathbf{F}_{ir}^{t} \in \mathbb{R}^{C_t \times H_t \times W_t}$, where the spatial resolution (H, W) progressively decreases along the network depth, and $C \in \{96, 192, 384, 768\}$. Subsequently, the feature maps $\{\mathbf{F}_{vi}^{t}, \mathbf{F}_{ir}^{t}\}_{t=1}^{4}$ at each layer are directly added and fed into the Single-Shot Alignment detector [26]. The overall loss function can be computed as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cls}}(c, \hat{c}) + \mathcal{L}_{\text{reg}}(\mathbf{b}, \hat{\mathbf{b}}) \tag{1}$$

where $\mathcal{L}_{\text{cls}}(c, \hat{c})$ denotes the classification loss between the ground-truth category c and the predicted category $\hat{c}$, which is computed using Focal Loss [27]. $\mathcal{L}_{\text{reg}}(\mathbf{b}, \hat{\mathbf{b}})$ denotes the regression loss between the ground-truth bounding boxes $\mathbf{b} = (x, y, w, h, \theta)$ and the predicted bounding boxes $\hat{\mathbf{b}} = (\hat{x}, \hat{y}, \hat{w}, \hat{h}, \hat{\theta})$. Smooth L1 Loss [28] is used to constrain the bounding-box regression.

However, both visible and infrared images are prone to degradation under adverse illumination and atmospheric conditions, such as low light or foggy scenes. Visible images suffer from noise and low contrast, while infrared images lack fine-grained textures. These degradations may interfere with the original feature maps $\{\mathbf{F}_{vi}^{t}, \mathbf{F}_{ir}^{t}\}_{t=1}^{4}$, ultimately limiting the object detection accuracy in dynamic UAV environments. Furthermore, the visible and infrared modalities vary dynamically depending on the imaging environment. For example, in nighttime scenes, the infrared modality usually dominates. If local light

sources are present, such as streetlights or vehicle lights, the visible modality may provide richer details in those areas. Similarly, in daytime smoke-obscured regions, the quality of visible images degrades, while the infrared modality may become reliable. Relying on naïve strategies such as concatenation, addition, or global weighting not only fails to fully exploit complementary information but may also introduce noise interference or cause feature dilution, thereby weakening the model's discriminative capability [29]. It is essential to design an adaptive fusion mechanism that can perceive global scene content and dynamically evaluate the reliability of each modality in local regions. To address these challenges, both feature enhancement module DDPR and the adaptive fusion module CDGF are proposed, which together form the core DEAF block of our method to learn a discriminative and robust multimodal representation.

3.2. Dual-Domain Progressive Refinement Module

In UAV visible–infrared object detection, fine-grained structures and contour details are crucial for distinguishing small or similar-shaped objects. Previous work has demonstrated that frequency-domain analysis can effectively suppress background noise in visible images while enhancing the scarce texture details in IR images [30,31]. Although the VMamba [23] backbones are good at capturing long-range dependencies, they tend to amplify low-frequency features such as background information and also suppress high-frequency features such as edges and textures [32]. To improve feature representation quality, a Dual Domain Progressive Refinement (DDPR) module is proposed to jointly perform frequency-domain enhancement and spatial-domain structural refinement, as illustrated in Figure 5.

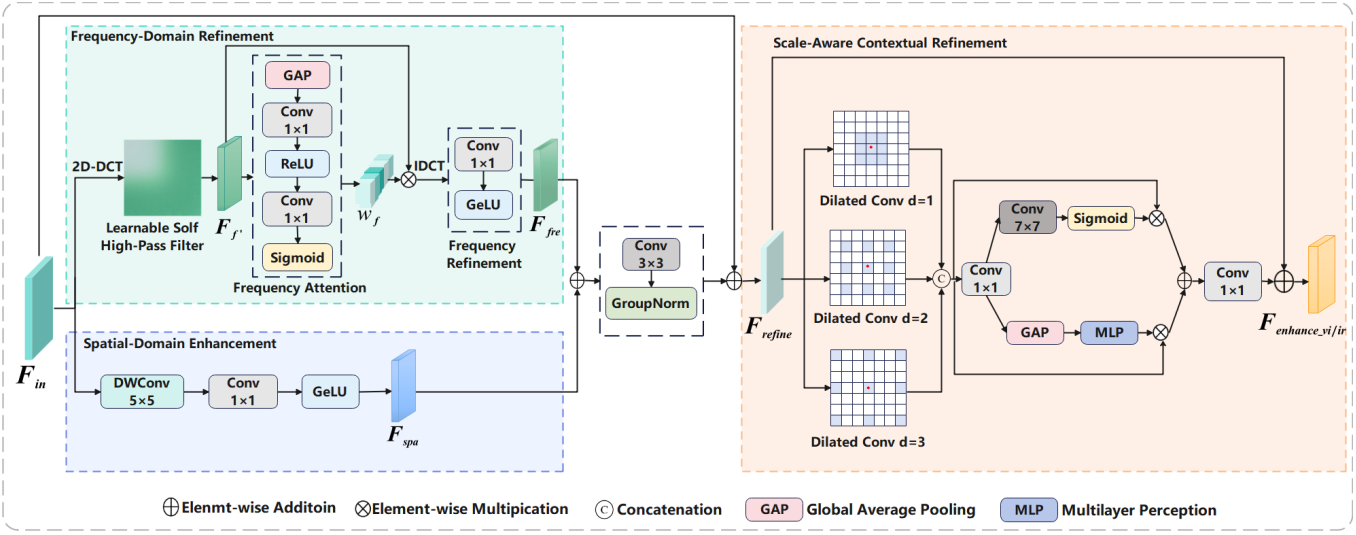


Figure 5. The structure of our proposed DDPR module.

Specifically, in the frequency domain refinement, a learnable soft high-pass filter is designed to selectively enhance discriminative high-frequency responses while suppressing low-frequency interference. In parallel, the spatial-domain refinement branch is employed to preserve structural continuity and semantic consistency of feature maps, thereby mitigating potential local distortions caused by frequency transformations. Building upon these purified dual-domain representations, a scale-aware contextual refinement strategy is further employed to aggregate multi-scale contextual information via dilated perception, thereby enhancing the representation of objects with significant scale variations in UAV scenes.

3.2.1. Frequency-Domain Refinement

Let $\mathbf{F}_{in}^{t} \in \mathbb{R}^{C_t \times H_t \times W_t}$ denote the intermediate feature from either modality $\mathbf{F}_{vi}^{t}$ or $\mathbf{F}_{ir}^{t}$ produced by the VMamba backbone, and where $t \in \{1, 2, 3, 4\}$, where H_t and W_t represent the spatial resolution at the t-th stage. We first transform $\mathbf{F}_{in}^{t}$ into the frequency domain by applying a channel-wise 2D Discrete Cosine Transform (2D DCT):

$$\mathbf{F}_{f}^{t} = \mathrm{DCT2}(\mathbf{F}_{in}^{t}) = \mathbf{D}_{H_t}\, \mathbf{F}_{in}^{t}\, \mathbf{D}_{W_t}^{\top} \tag{2}$$

where $\mathrm{DCT2}(\cdot)$ denotes 2D DCT operation, $\mathbf{D}_{H_t} \in \mathbb{R}^{H_t \times H_t}$ and $\mathbf{D}_{W_t}^{\top} \in \mathbb{R}^{W_t \times W_t}$ are the DCT transformation matrices.

In the frequency domain, high-frequency components generally correspond to edges and fine-grained textures, while low-frequency components mainly capture smooth background information. A common strategy for frequency enhancement is to apply a hard cutoff threshold to separate these components. However, due to the distinct spectral characteristics of visible and infrared modalities, their frequency energy distributions differ significantly. As a result, a fixed high-frequency threshold cannot simultaneously achieve optimal enhancement for both modalities.

To address this issue, we introduce a learnable soft high-pass filter that adaptively emphasizes discriminative high-frequency responses while suppressing low-frequency interference. Specifically, a soft high-pass mask $\mathbf{M} \in \mathbb{R}^{H_t \times W_t}$ is defined as:

$$\mathbf{M}_{h,w} = 1 - \sigma\left(\beta\left[\alpha_h - \frac{h-1}{H_t}\right]\right) \times \sigma\left(\beta\left[\alpha_w - \frac{w-1}{W_t}\right]\right) \tag{3}$$

where $h \in \{1, 2, \ldots, H_t\}$, $w \in \{1, 2, \ldots, W_t\}$, $\alpha_h, \alpha_w \in (0, 1)$ are learnable frequency parameters that control the starting position of the high-pass filter in the height and width directions, respectively. $\beta > 0$ is a learnable slope parameter that governs the steepness of the transition band from low to high frequencies. $\sigma(\cdot)$ denotes the sigmoid function. $\mathbf{M}$ is then applied to suppress degradation-related low-frequency components, yielding the masked frequency representation $\mathbf{F}_{f'}^{t} = \mathbf{F}_{f}^{t} \otimes \mathbf{M}$. We initialize α_h and α_w to 0.5. This places the cutoff at the mid-frequency position, allowing the network to balance low-frequency and high-frequency components from the start. β is initialized to 12, which yields a soft transition edge, enabling a soft cutoff between low and high frequencies.

The masked frequency features are further processed by a Frequency Attention mechanism, which generates channel-wise importance weights to adaptively emphasize informative frequency responses.

$$\mathbf{w}_f = \sigma\Big(\mathrm{MLP}\Big(\mathrm{GAP}(\mathbf{F}_{f'}^{t})\Big)\Big) \tag{4}$$

where the MLP is implemented using two successive 1×1 convolution layers with a ReLU activation in between, $\mathrm{GAP}(\cdot)$ denotes global average pooling,

The frequency-aware features are then reweighted and transformed back to the spatial domain via the Inverse DCT, followed by a lightweight Frequency Refinement operation:

$$\mathbf{F}_{fre}^{t} = \mathrm{GeLU}\Big(\mathrm{Conv}_{1\times1}\Big(\mathrm{IDCT}\Big(\mathbf{F}_{f'}^{t} \otimes \mathbf{w}_f\Big)\Big)\Big) \tag{5}$$

where $\otimes$ denotes element-wise multiplication with broadcasting of $\mathbf{w}_f \in \mathbb{R}^{C}$ across spatial dimensions, $\mathrm{Conv}_{1\times1}(\cdot)$ represents point-wise convolutions for cross-channel interaction, $\mathrm{GeLU}(\cdot)$ is the Gaussian Error Linear Unit activation function, and $\mathrm{IDCT}(\cdot)$ denotes the inverse DCT applied independently to each channel to convert frequency-domain features back to the spatial domain.

3.2.2. Spatial-Domain Enhancement

Although the frequency-domain branch effectively enhances high-frequency details via spectral modulation, it tends to attenuate spatial structures that rely on low-frequency components. To preserve essential local spatial dependencies, particularly object boundaries and structural continuity, we introduce a parallel spatial-domain enhancement path that aggregates neighborhood information, complementing the frequency domain refinement.

$$\mathbf{F}_{spa}^{t} = \mathrm{GeLU}\big(\mathrm{Conv}_{1\times1}\big(\mathrm{DWConv}_{5\times5}(\mathbf{F}_{in}^{t})\big)\big) \tag{6}$$

where $\mathrm{DWConv}_{5\times5}(\cdot)$ denotes a depth-wise convolution with a 5×5 kernel that captures local spatial structures independently for each channel.

The resulting feature $\mathbf{F}_{spa}^{t}$ represents the spatially enhanced feature map at stage t that preserves structural continuity and is subsequently fused with the frequency-enhanced feature $\mathbf{F}_{fre}^{t}$ to construct a comprehensive dual-domain representation.

3.2.3. Dual Domain Feature Integration

To synergistically combine the complementary information from the two domains, we fuse the frequency-enhanced feature $\mathbf{F}_{fre}^{t}$ and the spatially enhanced feature $\mathbf{F}_{spa}^{t}$. Specifically, we employ a residual learning strategy to ensure the preservation of original structural information while integrating the enhanced cues:

$$\mathbf{F}_{refine}^{t} = \mathrm{GroupNorm}\Big(\mathrm{Conv}_{3\times3}\big(\mathbf{F}_{spa}^{t} + \mathbf{F}_{fre}^{t}\big)\Big) + \mathbf{F}_{in}^{t} \tag{7}$$

where the element-wise summation $\mathbf{F}_{spa}^{t} + \mathbf{F}_{fre}^{t}$ initially merges the dual-domain representations. The subsequent 3×3 convolution integrates spatial-frequency cues into a unified representation. Finally, the residual connection adds the input $\mathbf{F}_{in}^{t}$ back to the output, facilitating gradient flow and stabilizing the optimization process via Group Normalization.

3.2.4. Scale-Aware Contextual Refinement

Following frequency and spatial-domain refinement, we further introduce a scale-aware contextual refinement operation to progressively enhance contextual awareness and robustness to scale variations in UAV scenes. This operation acts upon the dual-domain refined features and aims to selectively amplify scale-relevant and complementary responses.

Specifically, multi-scale contextual features are extracted by applying dilated convolutions with increasing dilation rates, where different dilation rates capture spatial context across multiple receptive field sizes. These features are then concatenated and projected into a unified representation:

$$\begin{aligned} \mathbf{F}_{i}^{t} &= \mathrm{Conv}_{3\times3,d=i}(\mathbf{F}_{refine}^{t}), \quad i = 1, 2, 3 \\ \mathbf{F}_{mix}^{t} &= \mathrm{Conv}_{1\times1}(\mathrm{Concat}(\mathbf{F}_{1}^{t}, \mathbf{F}_{2}^{t}, \mathbf{F}_{3}^{t})) \end{aligned} \tag{8}$$

where $\mathrm{Conv}_{3\times3,d=i}(\cdot)$ denotes a 3×3 dilated convolution with dilation rate i.

Within the scale-aware contextual refinement stage, a dual-path attention operation is employed to adaptively emphasize informative channels and spatially salient regions, yielding the enhanced features $\mathbf{F}_{enhance}^{t}$:

$$\begin{aligned} \mathbf{F}_{\mathrm{att}}^{t_c} &= \mathbf{F}_{\mathrm{mix}}^{t} \otimes \mathrm{MLP}(\mathrm{GAP}(\mathbf{F}_{\mathrm{mix}}^{t})) \\ \mathbf{F}_{\mathrm{att}}^{t_s} &= \mathbf{F}_{\mathrm{mix}}^{t} \otimes \sigma(\mathrm{Conv}_{7\times7}(\mathbf{F}_{\mathrm{mix}}^{t})) \\ \mathbf{F}_{enhance}^{t} &= \mathrm{Conv}_{1\times1}(\mathbf{F}_{\mathrm{att}}^{t_s} + \mathbf{F}_{\mathrm{att}}^{t_c}) + \mathbf{F}_{refine}^{t} \end{aligned} \tag{9}$$

This scale-aware contextual refinement stage effectively integrates globally modulated frequency cues with locally refined spatial features. Building upon the purified and enhanced features from both domains, it further captures scale-sensitive contextual information, thereby progressively constructing a more discriminative and robust feature representation.

3.3. Consistency-Discrepancy Guided Fusion Module

Visible and infrared modalities exhibit strong scene-dependent reliability: the visible spectrum provides fine-grained textures under good illumination but degrades significantly in darkness, fog, or occlusion, whereas infrared sensing retains thermal cues yet generally lacks structural detail. Most existing fusion methods assume spatially uniform modality reliability and therefore adopt global or channel-wise weighting strategies. However, UAV scenes often involve spatially non-uniform degradation. For instance, visible images captured at night may include both well-lit regions that are information-rich and shadowed areas that suffer from information loss. This spatial heterogeneity renders conventional fusion approaches fundamentally inadequate.

To address this challenge, we propose the Consistency-Discrepancy Guided Fusion (CDGF) module, as illustrated in Figure 6, which establishes an uncertainty-guided, spatially adaptive fusion paradigm. Following DDPR, CDGF completes the Dual-domain Enhanced Adaptive Fusion (DEAF) by performing cross-modal interaction guided by both local uncertainty cues and global reliability priors. The core idea is to decouple cross-modal consensus and complementarity through two guidance signals, leverage SS2D to model their long-range semantics and generate a local spatial modality weight matrix, and calibrate fusion responses using an entropy-based reliability prior.

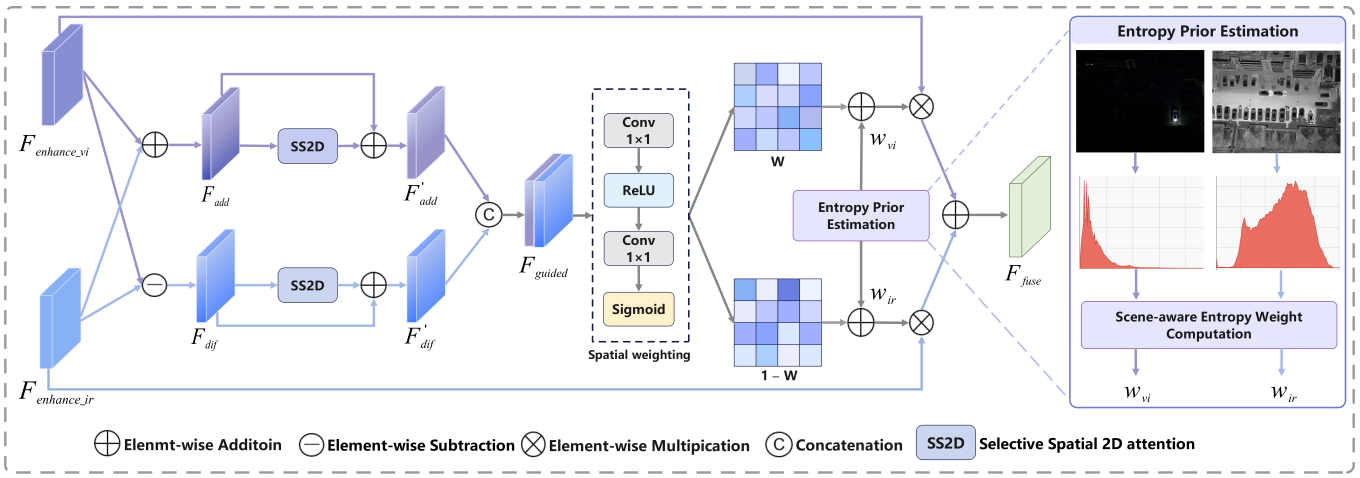


Figure 6. The structure of our proposed CDGF module.

3.3.1. Dual-Signal Guided Context Modeling

Given the DDPR enhanced features $\mathbf{F}^t_{enhance_vi}$ and $\mathbf{F}^t_{enhance_ir}$ from visible and infrared streams, CDGF first computes two guidance signals. To suppress local noise and enhance semantic coherence, both signals undergo long-range dependency modeling using two independent VMamba-based SS2D blocks:

$$\begin{aligned} \mathbf{F}^t_{add} &= \mathbf{F}^t_{enhance_vi} + \mathbf{F}^t_{enhance_ir} \\ \mathbf{F}^t_{dif} &= |\mathbf{F}^t_{enhance_vi} - \mathbf{F}^t_{enhance_ir}| \\ \mathbf{F}^{t'}_{add} &= SS2D(\mathbf{F}^t_{add}) + \mathbf{F}^t_{add} \\ \mathbf{F}^{t'}_{dif} &= SS2D(\mathbf{F}^t_{dif}) + \mathbf{F}^t_{dif} \end{aligned} \tag{10}$$

where $\mathbf{F}_{add}^{t'}$ captures shared or mutually reinforced information between the two modalities, and $\mathbf{F}_{dif}^{t'}$ highlights their complementary discrepancies arising from modality-specific characteristics or degradation. Both SS2D modules adopt the quad-directional scanning mechanism of VMamba, use default hyperparameters, and employ the NCHW layout.

3.3.2. Spatially Adaptive Fusion

Motivated by the observation that cross-modal reliability varies not only across different scenes but also within different regions of the same image, we design a two-level fusion mechanism that integrates scene-aware global priors with spatially-modulated local weights. This dual-level design ensures robustness against both large-scale environmental changes and fine-grained local degradations.

Based on $\mathbf{F}_{add}^{t'}$ and $\mathbf{F}_{dif}^{t'}$ compute the spatial modality weight matrix $\mathbf{W} \in [0,1]^{H \times W}$:

$$\begin{aligned} \mathbf{F}_{guided}^{t} &= \mathrm{Concat}\left[\mathbf{F}_{add}^{t'}, \mathbf{F}_{dif}^{t'}\right] \\ \mathbf{W} &= \sigma\left(\mathrm{Conv}_{1\times1}(\mathrm{ReLU}\left(\mathrm{Conv}_{1\times1}(\mathbf{F}_{guided}^{t})\right))\right) \end{aligned} \tag{11}$$

By jointly exploiting consistency and discrepancy cues, the mask assigns higher weights to the visible modality in texture-rich and well-illuminated regions, while emphasizing infrared features in low-illumination or thermally salient areas.

3.3.3. Scene-Aware Prior Calibration

To further enhance the stability and robustness of fused features across different scenes, we introduce a global prior weight based on the information entropy of input images for statistical calibration of locally adaptive weights. Information entropy, serving as a measure of image information complexity [33,34], effectively reflects the differences in information richness and reliability of each modality within the overall scene. This global prior weight, independent of network training and derived directly from the statistical properties of input data, provides scene-level guidance for local fusion decisions, suppressing fusion deviations caused by local feature noise or ambiguity, thereby enhancing the model's adaptive capability in complex and varying environments.

Specifically, given an input image I, we convert it to a grayscale image and compute its histogram. The information entropy of I is then calculated as:

$$\begin{aligned} p_i &= \frac{h_i}{\sum_{j=0}^{255} h_j + \epsilon}, \quad i = 0, \ldots, 255 \\ H(I) &= -\sum_{i=0}^{255} p_i \log(p_i + \epsilon) \end{aligned} \tag{12}$$

where h_i refers to the frequency of the i-th grayscale value, and ϵ is a smoothing constant.

Let $H_{vi} = H(I_{RGB})$ and $H_{IR} = H(I_{IR})$ denote the entropy of the visible and infrared images, respectively. The relative information entropy weights w_{ir} and w_{vi} for the two modalities are then defined as:

$$w_{vi} = \frac{H_{vi}}{H_{vi} + H_{ir}}, w_{ir} = \frac{H_{ir}}{H_{vi} + H_{ir}} \tag{13}$$

The final fusion output integrates the spatially-local weighting with scene-aware global guidance, yielding the multi-level fused features $\mathbf{F}_{\mathrm{fuse}}^{t}$:

$$\mathbf{F}_{\mathrm{fuse}}^{t} = (w_{vi} + \mathbf{W}) \otimes \mathbf{F}_{\mathrm{enhance_vi}}^{t} + (w_{ir} + (1 - \mathbf{W})) \otimes \mathbf{F}_{\mathrm{enhance_ir}}^{t} \tag{14}$$

where scalar global weights $w_{vi}, w_{ir} \in [0,1]$ are broadcast across the spatial dimensions of the local weight map $\mathbf{W} \in [0,1]^{H\times W}$, and the sum of local and global weights for each modality is constrained to $[0,2]^{H\times W}$, which ensures that the total scaling factor remains spatially constant, avoiding disruptive feature distribution fluctuations.

By integrating spatially adaptive weighting with scene-aware statistical calibration, CDGF effectively addresses spatial heterogeneity and enables robust cross-modal fusion under complex UAV imaging conditions.

Finally, the multi-level fused features $\{\mathbf{F}^t_{\text{fuse}}\}^4_{t=2}$ are fed into the S^2A detector to produce the detection results.

4. Experiment

This section presents comprehensive experiments to evaluate the effectiveness of the proposed DEAF-Net. We first introduce the datasets in Section 4.1, followed by implementation details and evaluation metrics in Section 4.2. Quantitative and qualitative comparisons with state-of-the-art methods are provided in Section 4.3, and ablation studies validating each component are presented in Section 4.4.

4.1. Datasets

We evaluate our method on two UAV RGB-infrared object detection benchmarks, DroneVehicle [15] and VEDAI [35]. Figure 7 illustrates the class distribution of the two datasets.

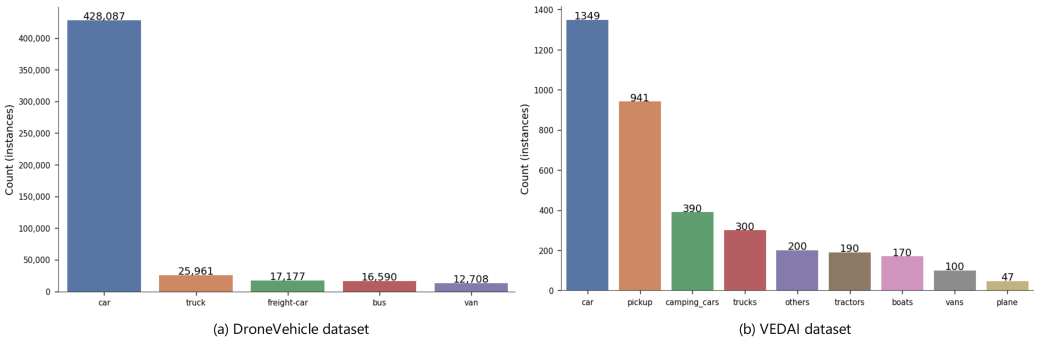


Figure 7. Class distribution in DroneVehicle and VEDAI Datasets.

4.1.1. DroneVehicle

DroneVehicle is a large-scale aerial vehicle detection dataset captured under diverse illumination and flight conditions (altitudes: 80–120 m; viewing angles: 15–45°). It contains 28,439 aligned visible–infrared image pairs with 953,087 oriented bounding box annotations across five vehicle classes (car, truck, bus, van, freight car). The dataset covers urban streets, highways, parking lots, and residential areas, featuring small-scale, densely occluded, and visually similar targets. We follow the official split: 17,990 for training, 1469 for validation, and 8980 for testing. Before training, we crop out the 100-pixel white borders present in all original images.

4.1.2. VEDAI

This dataset comprises 1246 visible and infrared image pairs, featuring over 3700 annotated object instances. The images are provided at resolutions of 1024×1024 and 512×512 pixels and include 9 object categories covering a variety of urban and rural scenes. In our experiments, the 512×512 version is adopted, and ten-fold cross-validation is employed for model training and testing.

4.2. Experimental Settings

4.2.1. Training Details

All experiments are conducted using the MMRotate and PyTorch frameworks on an NVIDIA A40 GPU with CUDA 12.2 and PyTorch 2.5.1. The input resolution is 512 × 640, with a batch size of 8. The AdamW optimizer is employed with an initial learning rate of 0.001 and a weight decay coefficient of 0.05. Due to the different data scales, this model is trained for 15 epochs on the DroneVehicle dataset and 100 epochs on the VEDAI dataset. To maintain fairness and consistency in comparison with previous approaches, the IR modality annotations are used as ground-truth labels.

4.2.2. Evaluation Metrics

In this experiment, mAP@0.5 is adopted as the primary evaluation metric, which represents the mean Average Precision (AP) at an IoU threshold of 0.5, where AP for each class is computed as the area under the precision–recall curve. To ensure fair comparison with other methods, all AP values are computed following the PASCAL VOC standard with the 11-point interpolation method [36].

4.3. Comparison Experiments

This paper compares the proposed method with the following methods: the classical single-modality object detection algorithms, including ROITrans [37], Oriented R-CNN [38], and S^2ANet [26]; as well as several state-of-the-art visible–infrared object detection methods, namely UA-CMDet [15], TSFADet [39], CLGNet [40], C^2Former [17], and DMM [18].

4.3.1. Comparisons on the DroneVehicle Dataset

A comparison of experimental results is detailed in Table 1. Overall, visible–infrared methods demonstrate significant advantages over the single-modality approaches. The proposed method achieves improvements of 21.1% and 12.0% over visible and infrared single-modality methods, respectively, confirming that multimodal feature fusion can leverage the complementary advantages of both modalities to generate more discriminative feature representations under complex lighting conditions, thereby enhancing overall detection performance. Among the compared multimodal methods, our approach achieves the highest mAP@0.5 of 81.9%, surpassing all other methods across all five vehicle categories. Notably, it outperforms the benchmark model UA-CMDet (72.1%) by 9.8 points and the DMM method (79.4%), which also employs VMamba as the backbone, by 2.5 points. Furthermore, the proposed method demonstrates stronger detection capabilities for sparse sample categories, notably trucks, vans, and freight cars, indicating improved robustness performance on categories with limited samples.

Table 2 presents a comprehensive comparison of computational costs between our DEAF-Net and mainstream approaches, all evaluated with an input size of 512 × 640. A thorough analysis reveals that our proposed DEAF-Net strikes a better balance among detection accuracy, model complexity, and inference speed. Compared to the benchmark method UA-CMDet, DEAF-Net achieves substantial improvements in accuracy while maintaining competitive model complexity and inference speed. It reduces the number of parameters by 16.2 M, increases FPS by 3.8 FPS, and enhances detection accuracy by 9.8%. When compared with C^2Former, DEAF-Net achieves a 7.7% gain in detection accuracy while maintaining a comparable number of model parameters, albeit with a slight decrease in inference speed. This design trade-off takes into account the stringent detection accuracy requirements in practical UAV applications, such as fault detection and search-and-rescue missions, where high detection accuracy is far more important than minor differences in latency.

Table 1. Comparative performance (mAP@0.5, in %) of different methods on the DroneVehicle dataset, all class-wise values are AP@0.5 (%). All detectors use OBB heads. The best results are emphasized in bold.

Method	Reference	Modality	Car	Truck	Bus	Van	Freight Car	mAP@0.5 (%)
ROITrans [37]	CVPR 19	RGB	78.4	52.5	79.0	41.9	39.8	58.3
Oriented R-CNN [38]	ICCV 21	RGB	80.1	53.8	85.4	43.3	41.6	60.8
S^2ANet [26]	TGRS 21	RGB	77.8	47.9	77.7	35.2	31.8	54.1
ROITrans [37]	CVPR 19	IR	90.2	67.0	89.1	47.9	55.2	69.9
Oriented R-CNN [38]	ICCV 21	IR	89.8	57.4	89.3	45.4	53.1	67.0
S^2ANet [26]	TGRS 21	IR	90.3	68.4	89.9	47.9	53.8	69.9
UA-CMDet [15]	TCSVT 22	RGB + IR	88.6	73.0	88.4	54.1	56.5	72.1
TSFADet [39]	ECCV 22	RGB + IR	89.9	67.9	89.8	54.0	63.7	73.1
CLGNet [40]	J-STARS 23	RGB + IR	90.4	72.6	89.9	59.5	64.6	75.4
C^2Former [17]	TGRS 24	RGB + IR	90.2	68.3	89.8	58.5	64.4	74.2
GLFDet [20]	CJA 25	RGB + IR	89.5	71.5	89.4	62.2	64.2	75.4
DMM [18]	TGRS 25	RGB + IR	90.4	79.8	89.9	68.6	68.2	79.4
Ours	—	RGB + IR	**90.5**	**83.2**	**90.0**	**73.0**	**73.0**	**81.9**

Table 2. Computational efficiency comparison of visible–infrared detection methods on the DroneVehicle dataset. The best results are emphasized in bold.

Method	mAP@0.5 (%)	Params (M)	FPS
UA-CMDet [15]	72.1	138.70	6.1
C^2Former [17]	74.2	**120.80**	**11.0**
Ours	**81.9**	122.52	9.9

Figure 8 provides a visual comparison of detection results on three typical scenarios from the DroneVehicle dataset: daytime, nighttime, and foggy conditions. In the daytime case, the IR-only modality lacks sufficient color and texture cues, leading to the misidentification of a metal-roofed structure as a truck (blue dashed box). The RGB-only modality, despite its rich color information, fails to detect vehicles whose colors blend into the background (yellow dashed box). While existing multi-modal fusion methods attempt to integrate the two modalities, they still produce a certain degree of false positives and missed detections in complex environments. In contrast, our DEAF-Net first performs denoising and feature enhancement through the DDPR module and then conducts adaptive fusion through the CDGF module. The collaboration of these two components enables strong robustness and environmental adaptability in challenging environments such as those with low light and heavily foggy conditions, and demonstrates strong effectiveness in various real-world drone imaging scenarios.

4.3.2. Comparisons on the VEDAI Dataset

The experimental results are presented in Table 3. Under single-modality conditions, visible-based methods consistently outperform infrared-based ones, primarily because the dataset was collected exclusively during daytime. Under sufficient illumination, the RGB modality provides rich color, texture, and spatial structural information, whereas the structural advantages of the IR modality become less pronounced. Among multimodal approaches, our proposed method achieves 76.2% mAP@0.5, outperforming the best RGB-only and IR-only approaches by 9.8% and 15.3%, respectively. It also surpasses the second-best multimodal method DMM by 1.2%, with notable gains on sparse categories, which include improvements of 7.0% for trucks and 4.2% for planes. Notably, another multi-modality method, C^2Former and GLFDet, achieves only 55.6% and 58.2 mAP@0.5, which is

even lower than the best single RGB modality result of 66.4%. This observation indicates that under well-lit conditions, visible features already possess strong discriminative power. With an inappropriate fusion strategy, IR features can be introduced as interference, which ultimately leads to negative fusion. In contrast, our adaptive fusion mechanism introduces global statistical constraints through scene-aware calibration during spatial modulation. This enables our approach to amplify complementary strengths while avoiding interference, thereby achieving superior overall performance.



Figure 8. Qualitative comparison of detection results on the DroneVehicle dataset under diverse scenes.

Table 3. Comparative performance (mAP@0.5, in %) of different methods on the VEDAI dataset; all class-wise values are AP@0.5 (%). All detectors use OBB heads. The best results are emphasized in bold.

Method	Modality	Car	Truck	Tractor	Camping Car	Van	Other	Pickup	Boat	Plane	mAP@0.5 (%)
ROITrans [37]	RGB	77.3	56.1	64.7	73.6	60.2	42.8	71.5	56.7	85.7	65.4
Oriented R-CNN [38]	RGB	77.6	59.7	62.8	76.6	60.9	43.6	72.3	60.1	84.0	66.4
S^2ANet [26]	RGB	74.5	47.3	55.6	61.7	32.5	39.8	65.1	16.7	7.1	44.5
ROITrans [37]	IR	76.1	51.7	51.9	71.2	64.3	28.3	70.7	46.9	83.3	60.5
Oriented R-CNN [38]	IR	77.0	55.0	47.5	73.6	63.2	30.5	72.2	49.4	79.6	60.9
S^2ANet [26]	IR	73.0	39.2	41.9	59.2	32.3	23.1	65.6	13.9	12.0	40.0
C^2Former [17]	RGB + IR	76.7	52.0	59.8	63.2	48.0	41.9	68.7	43.3	47.0	55.6
GLFDet [20]	RGB + IR	72.6	48.9	58.4	68.0	53.9	42.0	64.6	43.7	71.6	58.2
CMAFF [41]	RGB + IR	81.7	58.8	68.7	78.4	68.5	51.5	76.3	66.0	72.7	69.2
DMM [18]	RGB + IR	**84.2**	65.7	72.3	**79.0**	**72.5**	56.3	78.8	72.3	93.6	75.0
Ours	RGB + IR	83.9	**72.7**	**75.0**	78.2	68.5	**58.0**	**79.2**	**72.3**	**97.8**	**76.2**

As illustrated in Figure 9, DEAF-Net exhibits strong robustness across rural, urban, airport, and mountainous scenes. It successfully detects small vehicles that other methods often miss, and suppresses false positives on visually similar structures such as sheds or roofs. These results show that DEAF-Net maintains high sensitivity to fine-grained instances and strong discrimination against background clutter, validating the effectiveness of our proposed DEAF-Net in real-world scenarios.

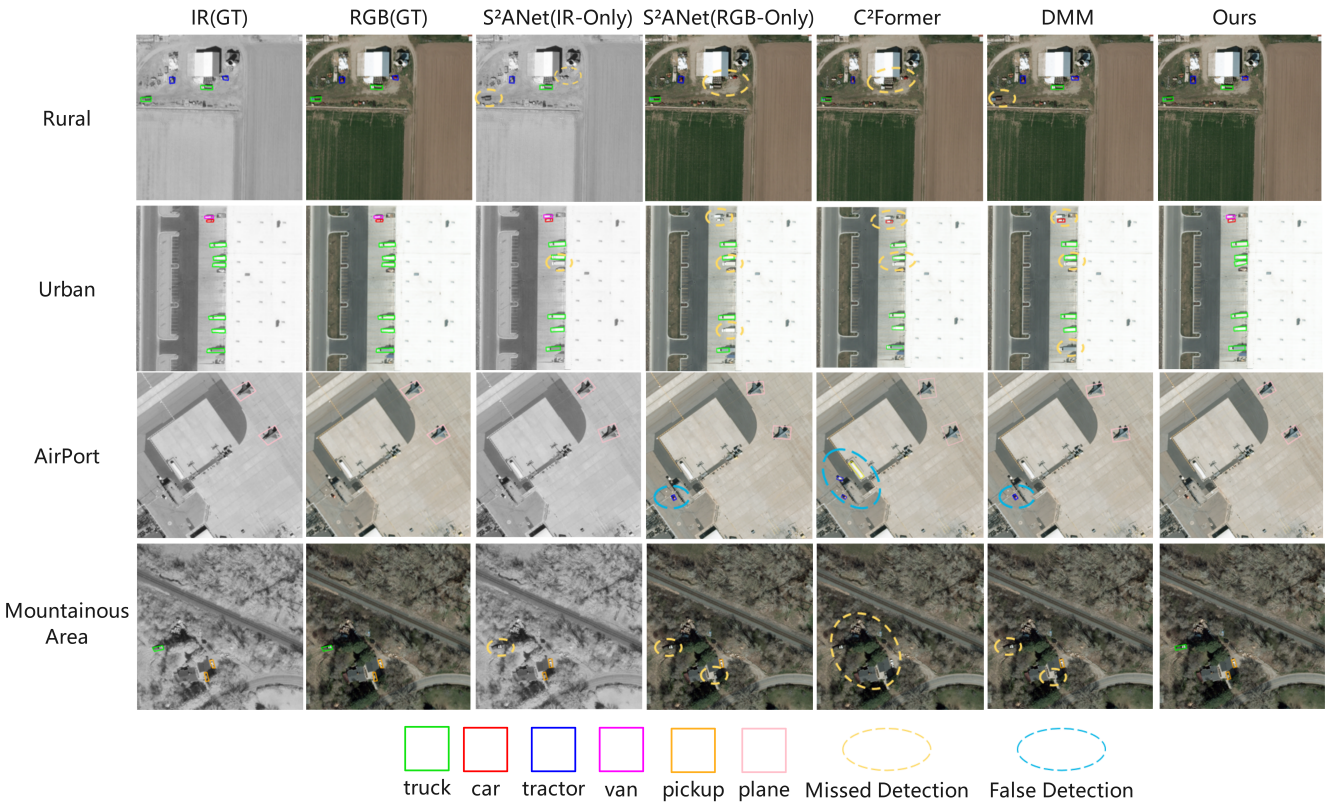


Figure 9. Visualization of detection results by various methods on the VEDAI dataset under diverse scenes.

4.4. Ablation Study

Since the VEDAI dataset is limited to daytime scenes, we conduct the ablation study on the more comprehensive DroneVehicle dataset, which includes diverse and challenging illumination conditions such as daytime, nighttime, and foggy environments. This enables us to comprehensively evaluate the robustness and adaptability of our multimodal fusion components under complex real-world scenes.

To evaluate the contribution of each key component in DEAF-Net, we conduct a systematic ablation study on the DroneVehicle dataset. Specifically, we analyze the impact of the two proposed core modules: DDPR and CDGF. The experiments follow an incremental design: starting from a baseline that only contains the VMamba backbone, FPN, and S^2A detection head, where multi-modal features are integrated via element-wise addition, we successively integrate DDPR and CDGF to isolate their individual and combined effects on detection performance.

As shown in Table 4, each proposed module independently improves performance, and the integration of the two modules leads to significant synergistic gains. DDPR alone brings a significant gain of 2.1%, demonstrating its effectiveness in feature purification. In contrast, CDGF alone yields only a marginal gain of 0.2%. This is because CDGF generates spatially adaptive fusion weights by modeling cross-modal consistency and discrepancy signals. When applied to raw backbone features, these signals can be corrupted by noise, weak textures, and unreliable modality responses under poor imaging conditions, leading to suboptimal fusion weights. However, when CDGF is combined with DDPR, it achieves an additional 1.1% gain over DDPR alone, and the full model outperforms the baseline by 3.2%. This synergy occurs because DDPR first suppresses low-frequency noise and enhances high-frequency details via frequency-domain filtering, followed by spatial-domain contextual refinement. The resulting clean and discriminative features enable

CDGF to accurately estimate local modality reliability and generate effective adaptive fusion weights. In essence, DDPR provides a purified feature foundation, while CDGF performs reliable fusion decisions upon it, their collaboration is essential for the potential of adaptive multimodal fusion.

Table 4. Ablation study of DEAF-Net components on the DroneVehicle dataset (mAP@0.5, in %). "✓" denotes that the corresponding module is integrated into the baseline, and empty means the module is not used. The best results are emphasized in bold.

Method	DDPR	CDGF	mAP@0.5 (%)	FLOPs (G)
Baseline			78.7	90.92
Baseline + DDPR	✓		80.8	120.77
Baseline + CDGF		✓	78.9	91.28
Ours	✓	✓	**81.9**	121.13

As shown in Table 5, both sub-modules of DDPR positively contribute to detection performance. Dual-domain refine enhances feature representation from two domains, while Scale-aware refine focuses on addressing the issue of drastic scale variations of objects from the UAV perspective. Their combination achieves improved performance, verifying their complementarity. Table 6 presents the ablation study of CDGF internal components. Using DDPR alone brings the major performance gain, albeit with substantial computational overhead, primarily due to frequency-domain transformation operations and parallel computation of multi-scale dilated convolutions. Adding spatial local fusion further improves mAP to 81.3% with only a 0.36 G FLOPs increase. Introducing the entropy-based Scene-aware prior further raises mAP to 81.9% with no FLOPs increase, as this method only involves entropy computation and scalar operations. The spatial local fusion handles fine-grained modality selection, while the entropy prior provides coarse-grained scene calibration, and the two complement each other. This further validates that combining frequency-spatial joint enhancement with explicit modeling of spatial heterogeneity of modality reliability can effectively improve object perception capability in complex environments.

Table 5. Ablation study of DDPR internal sub-modules on the DroneVehicle dataset (mAP@0.5, in %). "✓" denotes that the corresponding module is integrated. The best results are emphasized in bold.

Method	DDPR		CDGF	mAP@0.5 (%)	FLOPs (G)
Baseline	Dual-Domain Refine	Scale-Aware Refine			
✓				78.7	90.92
✓	✓			79.2	102.46
✓		✓		79.5	109.24
✓		✓	✓	79.6	109.59
✓	✓		✓	80.1	102.81
✓	✓	✓	✓	**81.9**	121.13

Table 7 presents the ablation results of applying the DEAF block at different stages of the VMamba backbone. The results show that integrating DEAF into the middle and late stages, namely stages 2 through 4, yields the most notable performance improvement. In contrast, applying DEAF only to the deepest stage (stage 4) leads to a limited gain. This suggests that refining features exclusively at high-level stages is insufficient for building robust representations under complex scenes. Based on these observations, DEAF is finally deployed at stages 2 to 4 of the backbone, enabling joint refinement of structural and semantic features and providing a more discriminative feature foundation for subsequent cross-modal adaptive fusion.

Table 6. Ablation study of CDGF internal sub-modules on the DroneVehicle dataset (mAP@0.5, in %). "✓" denotes that the corresponding module is integrated. The best results are emphasized in bold.

Method		CDGF		mAP@0.5 (%)	FLOPs (G)
Baseline	DDPR	Spatial Local Fusion	Scene-Aware Prior Calibration		
✓				78.7	90.92
✓	✓			80.8	120.77
✓	✓	✓		81.3	121.13
✓	✓	✓	✓	**81.9**	121.13

Table 7. Performance of the model after using the DEAF block at different stages of the backbone. "✓" denotes the DEAF block is applied to the corresponding stage.

Method	Stage1	Stage2	Stage3	Stage4	mAP@0.5 (%)
Baseline					78.7
Ours	✓	✓	✓	✓	81.2
Ours		✓	✓	✓	81.9
Ours			✓	✓	80.6
Ours				✓	79.4

To intuitively verify the advantages of the proposed fusion mechanism, we visualize fused feature response heatmaps in Figure 10. Compared with the baseline, DEAF-Net exhibits stronger activation on hidden targets in low-light or occluded regions (red boxes) and effectively suppresses false responses in areas with contour-similar interference (yellow boxes). This confirms that CDGF adaptively adjusts fusion weights based on local modality reliability under diverse imaging conditions, enabling precise attention to salient regions while suppressing irrelevant or misleading cues. Overall, DEAF-Net produces fused features with stronger target response and higher semantic consistency, providing more discriminative and robust representations for downstream detection tasks.

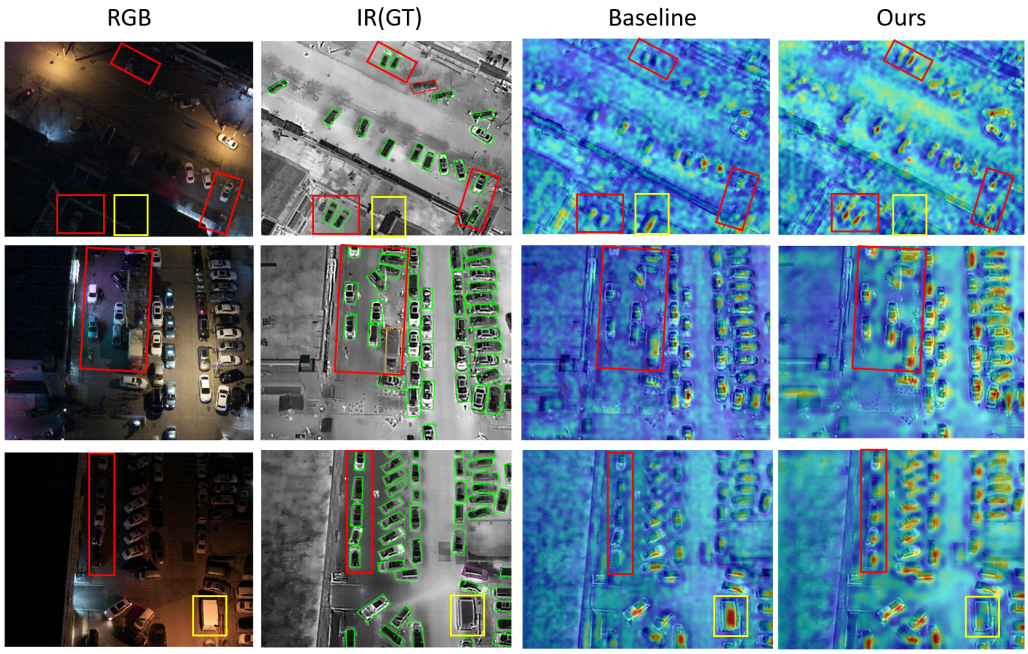


Figure 10. Visualization of fused feature activation maps on the DroneVehicle dataset. From (**left**) to (**right**): visible image, infrared image, fused features from the baseline, and fused features from DEAF-Net. Red boxes indicate low-illumination or occluded targets, while yellow boxes denote contour-similar interference regions in the infrared modality.

5. Discussion

Experimental results on DroneVehicle and VEDAI demonstrate that DEAF-Net achieves optimal detection performance under complex scenes such as low-light and foggy conditions, confirming the effectiveness of dual-domain enhancement and adaptive fusion for UAV-based visible–infrared object detection.

The core contribution lies in two synergistic modules. The DDPR module applies a learnable soft high-pass filter in the frequency domain to suppress background interference and enhance edge texture details, complemented by spatial-domain refinement and scale-aware contextual aggregation. The CDGF module decouples cross-modal interactions into consistency and discrepancy signals, generating spatially adaptive fusion weights calibrated by entropy-based global priors. Together, DDPR supplies purified feature foundations, and CDGF performs context-aware adaptive fusion.

Despite these advancements, this study has several limitations. First, the computational overhead is relatively high, which may pose challenges for deployment on resource-constrained edge devices. Second, our method assumes well-aligned visible–infrared image pairs; systematic evaluations of misalignment and controlled degradations have not been conducted. To tackle these challenges more effectively, future work could leverage current trends in computer vision. Specifically, the cross-attention mechanism in DETR could implicitly learn spatial correspondences between modalities, potentially easing the alignment handling, while foundation models such as CLIP and SAM could serve as powerful feature extractors to minimize domain gaps and reduce annotation dependence. Additionally, we will investigate model behavior under varying levels of noise, fog, illumination changes, and occlusions to provide a more comprehensive assessment of robustness.

In summary, despite these limitations, DEAF-Net enhances all-weather UAV object perception in complex environments by integrating spatial-frequency collaborative enhancement with consistency-discrepancy guided fusion.

6. Conclusions

In this work, we propose DEAF-Net, a dual-domain enhanced adaptive fusion framework for visible–infrared object detection under all-weather conditions in UAV applications. The framework integrates a DDPR module for frequency-spatial feature enhancement and a CDGF strategy for spatially adaptive cross-modal fusion. Extensive experiments on DroneVehicle and VEDAI demonstrate that DEAF-Net shows promising performance on challenging UAV scenarios. Future work will explore structured pruning for lightweight design and robust fusion strategies to handle potential misalignment between visible and infrared images caused by mechanical jitter or asynchronous capture in real-world UAV deployment.

Author Contributions: Conceptualization, Q.W., Y.Z. and L.D.; methodology, Q.W., Y.Z. and J.L.; software, Y.Z.; validation, Y.Z.; formal analysis, Q.W., L.D. and J.L.; investigation, Y.Z. and X.H.; resources, Q.W. and J.L.; data curation, Q.W. and Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, Q.W., X.H. and J.L.; visualization, X.H.; project administration, Q.W. and J.L.; funding acquisition, Q.W. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Natural Science Foundation Project of Fujian Province 2023J01432 and 2025J01538, in part by the Industry-Academy Cooperation Project of Fujian Province under Grant 2024H6006, in part by the Collaborative Innovation Platform Project of Fuzhou City under Grant 2023-P-002, and in part by the 2025 Special Fund Project for Promoting High-Quality Development of Marine and Fishery Industries in Fujian Province under Grant FJHYF-L-2025-07-004.

Data Availability Statement: The datasets used in this study include the DroneVehicle dataset and the VEDAI dataset. Both datasets are publicly available and can be accessed at https://github.com/VisDrone/DroneVehicle (accessed on 20 November 2025) and https://downloads.greyc.fr/vedai/ (accessed on 22 November 2025).

Acknowledgments: During the preparation of this manuscript, the authors used DeepSeek-Chat for the purposes of language polishing and fluency refinement. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xiao, Y.; Wang, J.; Zhao, Z.; Jiang, B.; Li, C.; Tang, J. UAV video vehicle detection: Benchmark and baseline. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5609814. [CrossRef]
2. Senthilnath, J.; Kumar, A.; Jain, A.; Harikumar, K.; Thapa, M.; Suresh, S.; Anand, G.; Benediktsson, J.A. BS-McL: Bilevel segmentation framework with metacognitive learning for detection of the power lines in UAV imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5604712.
3. Xu, J.; Fan, X.; Jian, H.; Xu, C.; Bei, W.; Ge, Q.; Zhao, T. Yoloow: A spatial scale adaptive real-time object detection neural network for open water search and rescue from uav aerial imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5623115.
4. Santos, N.P.; Rodrigues, V.B.; Pinto, A.B.; Damas, B. Automatic detection of civilian and military personnel in reconnaissance missions using a UAV. In *2023 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*; IEEE: New York, NY, USA, 2023; pp. 157–162.
5. Fang, A.; Li, Y. Dynamic and static fusion mechanisms of infrared and visible images. *Pattern Recognit.* **2024**, *155*, 110689. [CrossRef]
6. Bao, M.; Lyu, S.; Xu, Z.; Zhou, H.; Ren, J.; Xiang, S.; Li, X.; Cheng, G. Vision Mamba in Remote Sensing: A Comprehensive Survey of Techniques, Applications and Outlook. *arXiv* **2025**, arXiv:2505.00630.
7. Li, K.; Wang, D.; Hu, Z.; Li, S.; Ni, W.; Zhao, L.; Wang, Q. Fd2-net: Frequency-driven feature decomposition network for infrared-visible object detection. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 4797–4805.
8. Li, H.; Hu, X.; Wang, H. Interpretable Unsupervised Joint Denoising and Enhancement for Real-World low-light Scenarios. In *The Thirteenth International Conference on Learning Representations*; International Conference on Learning Representations (ICLR): Appleton, WI, USA, 2025.
9. Hu, K.; Zhang, Q.; Yuan, M.; Zhang, Y. SFDFusion: An efficient spatial-frequency domain fusion network for infrared and visible image fusion. *arXiv* **2024**, arXiv:2410.22837.
10. Luo, Z.; Wang, Y.; Chen, L.; Yang, W. Frequency spectrum features modeling for real-time tiny object detection in remote sensing image. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6011205. [CrossRef]
11. Liu, Z.; Deng, X.; Jiang, P.; Lv, C.; Min, G.; Wang, X. Edge perception camouflaged object detection under frequency domain reconstruction. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 10194–10207. [CrossRef]
12. Gao, X.; Zhao, D.; Yuan, B.; Li, W. Dual-Stream Spectral Decoupling Distillation for Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5639711. [CrossRef]
13. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral deep neural networks for pedestrian detection. *arXiv* **2016**, arXiv:1611.02644.
14. Xiao, X.; Wang, B.; Miao, L.; Li, L.; Zhou, Z.; Ma, J.; Dong, D. Infrared and visible image object detection via focused feature enhancement and cascaded semantic extension. *Remote Sens.* **2021**, *13*, 2538. [CrossRef]
15. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713.
16. Kang, X.; Yin, H.; Duan, P. Global–local feature fusion network for visible–infrared vehicle detection. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6005805.
17. Yuan, M.; Wei, X. C^2former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5403712.
18. Zhou, M.; Li, T.; Qiao, C.; Xie, D.; Wang, G.; Ruan, N.; Mei, L.; Yang, Y.; Shen, H.T. Dmm: Disparity-guided multispectral mamba for oriented object detection in remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5404913. [CrossRef]
19. Shi, Z.; Hu, J.; Ren, J.; Ye, H.; Yuan, X.; Ouyang, Y.; He, J.; Ji, B.; Guo, J. HS-FPN: High frequency and spatial perception FPN for tiny object detection. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 6896–6904. [CrossRef]
20. Chen, Z.; Ji, H.; Zhang, Y. Global-local feature optimization based RGB-IR fusion object detection on drone view. *Chin. J. Aeronaut.* **2025**, *39*, 103781.
21. Berjawi, J.; Dupas, Y.; Cérin, C. Towards a Generalizable Fusion Architecture for Multimodal Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2025; pp. 2192–2200.

22. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In Proceedings of the First Conference on Language Modeling, Philadelphia, PA, USA, 7–9 October 2024.
23. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063. [CrossRef]
24. Chen, K.; Chen, B.; Liu, C.; Li, W.; Zou, Z.; Shi, Z. Rsmamba: Remote sensing image classification with state space model. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 8002605. [CrossRef]
25. Li, Y.; Wang, L.; Chen, S. Smile: Spatial-spectral mamba interactive learning for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5005214.
26. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5602511. [CrossRef]
27. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017; pp. 2980–2988.
28. Girshick, R. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2015; pp. 1440–1448.
29. Cao, B.; Sun, Y.; Zhu, P.; Hu, Q. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2023; pp. 23555–23564.
30. Huang, Y.; Zhi, X.; Hu, J.; Yu, L.; Han, Q.; Chen, W.; Zhang, W. FDDBA-NET: Frequency domain decoupling bidirectional interactive attention network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5004416. [CrossRef]
31. Duan, W.; Ji, L.; Chen, S.; Zhu, S.; Ye, M. Triple-domain feature learning with frequency-aware memory enhancement for moving infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5006014.
32. Ma, X.; Ni, Z.; Chen, X. Tinyvim: Frequency decoupling for tiny hybrid vision mamba. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2025; pp. 23519–23529.
33. Wei, H.; Liu, J.; Chen, B.; Lin, L.; Chen, W.; Zhao, T. Geometry-based spherical JND modeling for 360° display. *arXiv* **2023**, arXiv:2303.03703.
34. Yin, D.; Zhou, S.; Wang, P.; Lin, M.; Song, H.; Ke, F.; Luo, K. A feature points extraction algorithm based on adaptive information entropy. *IEEE Access* **2020**, *8*, 127134–127141. [CrossRef]
35. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]
36. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338.
37. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2019; pp. 2849–2858.
38. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2021; pp. 3520–3529.
39. Yuan, M.; Wang, Y.; Wei, X. Translation, scale and rotation: Cross-modal alignment meets RGB-infrared vehicle detection. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 509–525.
40. Xie, J.; Nie, J.; Ding, B.; Yu, M.; Cao, J. Cross-modal local calibration and global context modeling network for RGB–infrared remote-sensing object detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 8933–8942.
41. Fang, Q.; Wang, Z. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognit.* **2022**, *130*, 108786. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.