# RHG-DETR: Riemannian Hyper-Graph Transformerwith Dynamic Receptive Fields for Detecting Special Targets in Degraded UAV Imagery

Kaipeng Wang , Guanglin He * , Wenhao Kong , Yuzhe Fu , Zongze Li

Posted Date: 24 April 2026

doi: 10.20944/preprints202604.1731.v1

Keywords: UAV remote sensing; special target detection; composite image degradation; dynamic receptive field; hyper-graph attention; adaptive sparse encoding

 doi:10.20944/preprints202604.1731.v1



*Article*

# RHG-DETR: Riemannian Hyper-Graph Transformer with Dynamic Receptive Fields for Detecting Special Targets in Degraded UAV Imagery

**Kaipeng Wang, Guanglin He *, Wenhao Kong, Yuzhe Fu and Zongze Li**

National Key Laboratory of Proximity Detection and Control, Beijing Institute of Technology, Beijing 100081, China
* Correspondence: heguanglin@bit.edu.cn; Tel.: +86-150-1096-5661

**Abstract**

Accurate detection of special targets in unmanned aerial vehicle (UAV) remote sensing imagery under complex degradation conditions remains a critical challenge for intelligent surveillance systems. Existing detectors exhibit significant performance degradation when confronted with composite degradation factors such as blur, rain, snow, fog, low illumination, strong light, and electromagnetic interference. To address this limitation, we propose RHG-DETR (Riemannian Hyper-Graph Detection Transformer), a novel detection framework for robust special target detection under multi-type degradation in UAV remote sensing imagery. Using RT-DETR as the baseline, three synergistic innovations are introduced at the backbone, neck, and encoder levels. The Dynamic Receptive-field Hyper-graph Attention Network (DRHANet) replaces the conventional ResNet backbone, employing anisotropic dynamic depthwise separable convolution and a Riemannian Hyper-graph Fusion (RHGF) mechanism to model high-order semantic topology dependencies among target components. The Bi-directional Weighted Adaptive Fusion Network (BWAFN) constructs a two-stage bidirectional feature pyramid with learnable scale contribution weights and a lightweight spatial compensation upsampler to maintain cross-scale semantic consistency under atmospheric degradation. The Adaptive Sparse Multi-scale Encoder with Dynamic normalization (ASMED) reconstructs the AIFI encoder module by introducing sparse window self-attention to suppress background interference, a spatial-gated feedforward fusion to preserve geometric topology constraints of target sub-components, and coordinated dynamic normalization modules to stabilize encoding under extreme illumination and electromagnetic interference. On a self-constructed special target dataset comprising tanks, multiple launch rocket systems, and soldiers under seven degradation types, RHG-DETR achieves an $mAP_{50}$ of 78.5%, surpassing the RT-DETR baseline by 3.7%, while reducing GFLOPs and parameter count by 34.4% and 28.8%, respectively, at an inference speed of 84.2 FPS. Consistent improvements on VisDrone2019 and BDD100K further validate the cross-domain generalization capability of the proposed framework.

**Keywords:** UAV remote sensing; special target detection; composite image degradation; dynamic receptive field; hyper-graph attention; adaptive sparse encoding

## 1. Introduction

UAV-based remote sensing technology has been widely adopted in target detection owing to its broad area coverage, flexible deployment capability, and the ability to acquire high-resolution imagery at low altitudes [1,2]. Rapid and accurate identification of special targets in complex environments constitutes a core challenge for intelligent surveillance and reconnaissance systems. In real-world scenarios, composite degradation factors are pervasive, including atmospheric degradation (fog, rain, and snow), adverse illumination conditions (low light and overexposure), motion blur, and electromagnetic interference, all of which substantially degrade image quality and fundamentally constrain the reliability of detection systems [3,4]. Throughout this paper, special vehicles refer to various military

 doi:10.20944/preprints202604.1731.v1



and civilian purpose-built vehicles, including but not limited to tanks, armored vehicles, multiple launch rocket systems (MLRS), radar systems, fire trucks, ambulances, and engineering vehicles. Although our study concentrates on special targets, the proposed technical framework—particularly its robustness to degradation factors such as rain, snow, and fog—holds considerable potential for civilian applications, including all-weather autonomous driving, disaster response operations, and infrastructure inspection [5].

Deep learning-based object detection has made remarkable strides in recent years, yet maintaining high-performance robust detection across diverse degradation conditions remains an open research problem demanding urgent attention. Deep learning object detection has progressed through an evolutionary path from two-stage to single-stage and subsequently to end-to-end paradigms. Faster R-CNN [6] established the foundational paradigm for two-stage detection through its learnable region proposal mechanism. Single-stage detectors represented by the YOLO series [7–9] unify localization and classification within a single forward pass, enabling near-real-time inference. End-to-end transformer detectors exemplified by the DETR series [10,11] model long-range dependencies through global self-attention, demonstrating strong performance on general benchmarks. Nevertheless, directly applying these architectures to UAV-based special target detection under degraded imaging conditions exposes fundamental limitations at three levels: backbone feature extraction, multi-scale neck fusion, and encoder semantic modeling.

At the feature extraction level, backbone networks built upon fixed-receptive-field residual stacking suffer from systematic multi-scale feature representation degradation under blurred and low-contrast backgrounds. Fixed convolutional kernels cannot dynamically adjust their spatial receptive range according to the directional distribution of target features, resulting in severely inadequate capacity for simultaneous modeling of fine-grained local textures and cross-layer semantic correlations. This limitation is especially pronounced in occlusion scenarios and component-level localization tasks—for instance, separately detecting the turret and hull of a tank, or identifying the launch pod of a rocket system—where discriminative structural features are spatially concentrated and strongly anisotropic [12,13]. At the multi-scale fusion level, Feature Pyramid Network (FPN) and its variants typically adopt fixed unidirectional fusion paths and equal-weight summation strategies, lacking adaptive capacity to dynamically balance scale contributions. Under composite atmospheric degradation, alignment errors between shallow spatial detail features and deep semantic representations accumulate progressively across successive fusion stages, easily causing feature dominance imbalance in scenes where targets of multiple spatial granularities coexist, such as complete vehicles and sub-components [14,15]. At the encoder modeling level, transformer encoder modules designed for natural image understanding fail to adequately account for the structural characteristics of special targets in aerial UAV imagery. Standard global self-attention performs indiscriminate aggregation over semantically sparse background regions, continuously diluting the feature representation of target regions; spatial geometry-dependent sub-component semantics—such as turret-hull spatial constraints or launch pod positional relationships—cannot be effectively preserved during the encoding process [14,16].

To systematically address the foregoing challenges at all three levels, we propose RHG-DETR (Riemannian Hyper-Graph Detection Transformer), a novel framework for detecting special targets in UAV remote sensing imagery under multi-type degradation conditions, introducing three synergistic innovations at the backbone, neck, and encoder stages respectively. The main contributions are summarized as follows:

(1) We propose DRHANet (Dynamic Receptive-field Hyper-graph Attention Network) as a replacement for the conventional ResNet backbone. DRHANet introduces the Channel-split Dynamic Stream Fusion (CDSF) module, replacing fixed residual units with Dual-path Adaptive Mixing Blocks (DAMB) and employing Anisotropic Dynamic Depthwise Convolution (ADWC) to achieve multi-scale local feature extraction, enabling the receptive field orientation to adaptively match the directional distribution of special target structures. At key fusion nodes of the feature pyramid, the Riemannian

 doi:10.20944/preprints202604.1731.v1



Hyper-graph Fusion (RHGF) mechanism models high-order semantic topology dependencies among target components on the feature manifold, offering significant advantages over standard convolutional operators in cross-region semantic binding between the turret–hull and launch pod–cradle associations.

(2) We propose BWAFN (Bi-directional Weighted Adaptive Fusion Network) as a replacement for the conventional FPN neck architecture. BWAFN constructs a two-stage bidirectional feature pyramid, with learnable scale contribution weights normalized by a fast normalization attention mechanism at each node, enabling dynamic adaptive balancing of fusion ratios across scales. The Lightweight Spatial Compensation Upsampler (LSCU), composed of nearest-neighbor upsampling combined with a Shift-based Channel Aggregation with Group-equivariance (SCAG) module, eliminates spatial activation inconsistencies during resolution restoration. Cross-stage Multi-scale Receptive-field Transform (CMRFT) nodes maintain cross-scale semantic consistency under atmospheric degradation through iterative multi-scale receptive field expansion.

(3) We propose ASMED (Adaptive Sparse Multi-scale Encoder with Dynamic normalization) to systematically reconstruct the AIFI encoder module. ASMED introduces Sparse Window-based Self-Attention (SWSA), which restricts feature interactions to a high-saliency spatial subset within local windows, effectively suppressing background interference in semantically sparse scenes. The Spatial-Gated Feedforward Fusion (SGFF) module introduces a parallel spatial branch conditioned on the pre-attention feature map, explicitly injecting geometric topology constraints into the feedforward transformation stage to preserve spatial relational semantics among target sub-components. The Dynamic Channel-wise Nonlinear Modulator (DCNM) and Multi-scale Gated Coupling Adaptor (MGCA) further cooperate to stabilize the encoding process, effectively addressing the high dynamic range and non-stationary activation distributions induced by extreme illumination and electromagnetic interference.

Comprehensive experimental results demonstrate that the proposed method achieves superior detection performance over existing approaches under multi-type degradation conditions. The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 presents the proposed method in detail; Section 4 reports experimental results and ablation studies; and Section 5 concludes the paper with directions for future research.

## 2. Related Work

### *2.1. UAV Remote Sensing Object Detection under Multi-Type Degradation*

Object detection in UAV remote sensing imagery has long been a research hotspot in computer vision, characterized by diverse viewpoints, significant target scale variation, and complex backgrounds. Yu et al. [17] proposed FSTD-Net, a lightweight full-scale target detection network that addresses multi-scale and complex background challenges in UAV imagery through multi-scale contextual information extraction, deformable convolution for shape-adaptive feature extraction, and low-level feature enhancement. Li et al. [18] proposed a scale-aware multi-domain DETR framework for small target detection in UAV remote sensing imagery, tackling densely distributed small targets through a scale-aware attention backbone and multi-domain feature modeling.

However, most existing works have been conducted under standard imaging conditions without adequately considering the pervasive image degradation encountered in practical deployment. Factors such as fog, rain, snow, and low illumination substantially reduce image contrast, introduce noise, and blur target boundaries, all of which seriously impair detector performance. Pei et al. [19] systematically evaluated the impact of degradation phenomena such as blur and noise on convolutional neural network (CNN) image classification performance; Zhao et al. [20] proposed a generative strategy based on Retinex decomposition to construct a unified deep framework for low-light image enhancement; Ma et al. [21] proposed an end-to-end domain-adaptive framework combining a dehazing module with YOLOX to improve detection performance in foggy conditions; Kang et al. [22] proposed a deep learning recognition framework targeting three weather types—fog, rain, and snow; and Gupta et al. [23] investigated various training strategies to enhance detector robustness under adverse

 doi:10.20944/preprints202604.1731.v1



weather conditions. Nonetheless, these methods were each designed for a single or limited number of degradation types and struggle to handle scenarios involving composite overlapping degradations. There remains a conspicuous gap in research addressing the detection of special targets and their sub-components under composite degradation conditions.

### *2.2. RT-DETR: Architecture Overview and Baseline Analysis*

The Real-Time Detection Transformer (RT-DETR) was proposed by Zhao et al. [11] as the first end-to-end object detection framework that simultaneously achieves high accuracy and real-time inference. Unlike conventional detectors that rely on non-maximum suppression (NMS) post-processing, RT-DETR directly predicts sets of objects through a bipartite matching mechanism, eliminating the inference latency and hyperparameter sensitivity introduced by NMS. RT-DETR adopts ResNet [24] as its backbone, with the AIFI and CCFM modules within an efficient hybrid encoder performing global context modeling and cross-scale feature aggregation, while the decoder employs learnable object queries and an IoU-based query selection strategy to dynamically select high-confidence anchors, ultimately outputting class probabilities and bounding box coordinates through multi-layer cross-attention.

Although RT-DETR demonstrates outstanding accuracy–speed trade-offs on general detection benchmarks, direct application to UAV remote sensing special target recognition under multiple complex degradation conditions exposes inherent limitations: rain-snow noise and motion blur disrupt target texture features and cause disordered attention weights; night and strong-light conditions introduce domain shift, degrading backbone feature stability; the quadratic complexity of transformer self-attention limits spatial resolution in small-target scenarios; and furthermore, standard RT-DETR treats each target as an independent instance, incapable of leveraging hierarchical semantic constraints between components such as tanks and turrets to assist discrimination. These limitations collectively constitute the core motivation for the improvements proposed in this paper.

### *2.3. Special Target Detection Datasets*

High-quality annotated datasets provide an indispensable foundation for advancing remote sensing object detection research. VisDrone [25] is one of the largest UAV visual benchmarks, covering 10 target categories with over 400,000 annotated instances; however, its images were collected under standard conditions without covering complex degradation scenes. DOTA [26] and DIOR [27] focus on multi-class geographic feature detection in optical remote sensing images and similarly do not consider degraded imaging conditions, nor do they address component-level target recognition.

In this paper, we construct a UAV-perspective special target dataset covering three target categories: tank (with turret sub-component), MLRS (with launch pod sub-component), and soldier, designed to simultaneously encompass a broad range of scales, pose angles, and diverse complex scenes. All images were collected from publicly available Internet sources and underwent rigorous expert review and annotation. The complete dataset comprises 8,546 images partitioned into training (5,982), validation (854), and test (1,710) subsets in a 7:1:2 ratio, with stratified sampling employed to ensure consistent distribution across subsets.

The degradation simulation dataset consists of single-blur and composite-degradation images. Single-blur degradation is divided into mild, moderate, and severe grades at a 4:4:2 ratio. Composite degradation layers six types—rain, fog, snow, strong light, low illumination, and electromagnetic interference—onto the blur base to simulate complex realistic multi-interference imaging conditions; 800 images are selected for each composite degradation type and processed at mild, moderate, and severe intensity levels.

All degradation effects are synthesized using physics- or statistics-based methods: motion blur and defocus blur are realized through directional kernel convolution and Gaussian defocus kernels [28,29]; rain streaks are synthesized based on physics-driven rendering [30]; snowfall is simulated via random snowflake particle distribution [31]; fog is modeled using the Koschmieder atmospheric scattering model [32,33]; electromagnetic interference is implemented by superimposing periodic

 doi:10.20944/preprints202604.1731.v1



stripe patterns, salt-and-pepper noise, and block artifacts [34,35]; low illumination is jointly simulated through gamma correction and Poisson noise [36–38]; and strong light is realized through brightness saturation processing [35,39]. These strategies are consistent with established degraded image detection benchmark specifications [40,41], aiming to systematically reproduce the typical degradation scenarios encountered in actual deployment of target recognition systems and substantially improving the validity of the dataset for assessing algorithmic robustness. To illustrate the different degradation effects more clearly, we applied single-degradation processing of varying intensities to the same image, as shown in Figure 1. The first column shows the clean image, the second column shows the mildly degraded image, the third column shows the moderately degraded image, and the fourth column shows the severely degraded image.

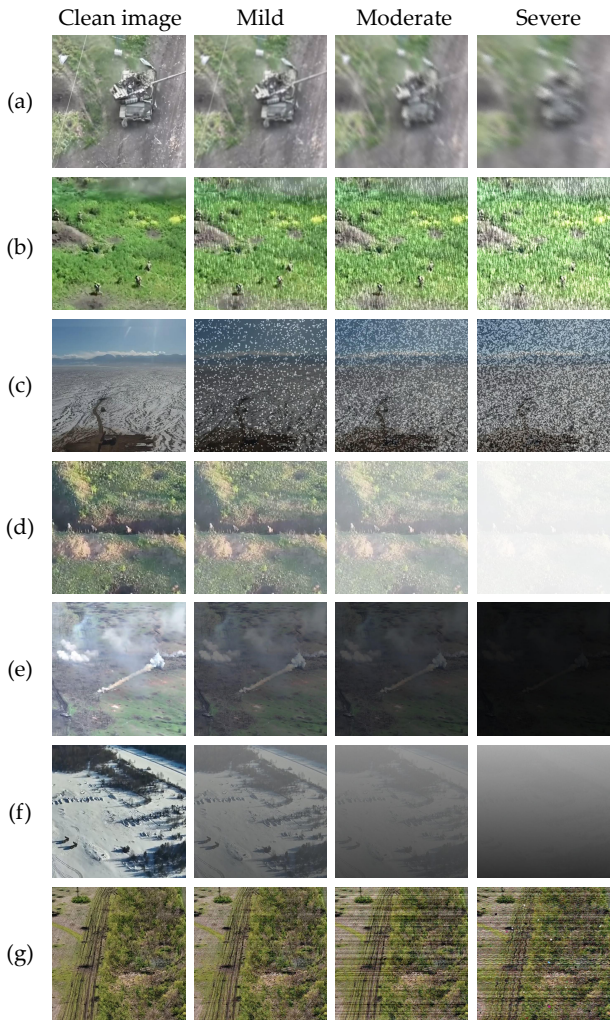


**Figure 1.** (a) Clean image and blur degradation at three intensity levels. (b) Clean image and rain degradation at three intensity levels. (c) Clean image and snow degradation at three intensity levels. (d) Clean image and strong-light degradation at three intensity levels. (e) Clean image and low-illumination degradation at three intensity levels. (f) Clean image and fog degradation at three intensity levels. (g) Clean image and electromagnetic interference degradation at three intensity levels.

Representative samples from the dataset are shown in Figure 2. Soldiers are marked in red; MLRS are marked in dark blue with their launch pods in light blue; tanks are marked in yellow with their turrets in orange. Due to data sensitivity, the dataset used in this study is not publicly archived. It is available only upon reasonable request for academic purposes; all inquiries should be submitted to the corresponding author.

 doi:10.20944/preprints202604.1731.v1



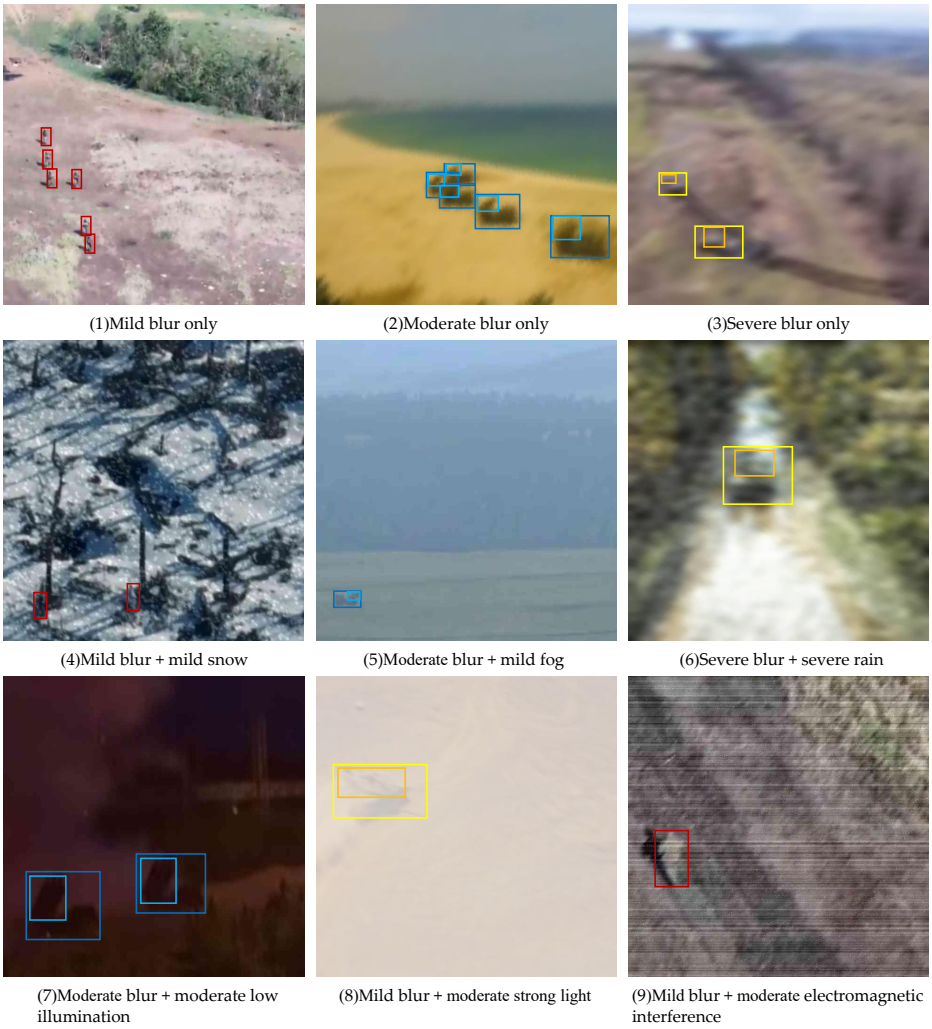


**Figure 2.** Representative samples from the self-constructed special target detection dataset, covering single-blur degradation and multi-source composite degradation scenarios.

## 3. Method

As illustrated in Figure 3, the overall RHG-DETR framework takes RT-DETR as its baseline and introduces three synergistic innovation modules at the backbone, neck, and encoder levels in a cooperative manner. First, DRHANet replaces the conventional ResNet backbone, realizing adaptive multi-scale local feature extraction through anisotropic dynamic depthwise separable convolution and introducing the RHGF mechanism at key feature pyramid fusion nodes to model high-order semantic topology dependencies among target components on the feature manifold. Subsequently, BWAFN replaces the conventional FPN neck architecture, constructing a two-stage bidirectional feature pyramid in which each fusion node dynamically balances the contribution weights of each scale through the fast normalization attention mechanism, maintaining cross-scale semantic consistency in scenes where multi-granularity targets coexist under composite degradation. Finally, ASMED systematically reconstructs the AIFI encoder module, suppressing background interference through sparse window self-attention and stabilizing the encoding process via the DCNM and MGCA. The three modules jointly constitute an end-to-end special target detection framework for composite degradation environments. The following subsections describe the structural design and operating principles of each component module in detail.

 doi:10.20944/preprints202604.1731.v1



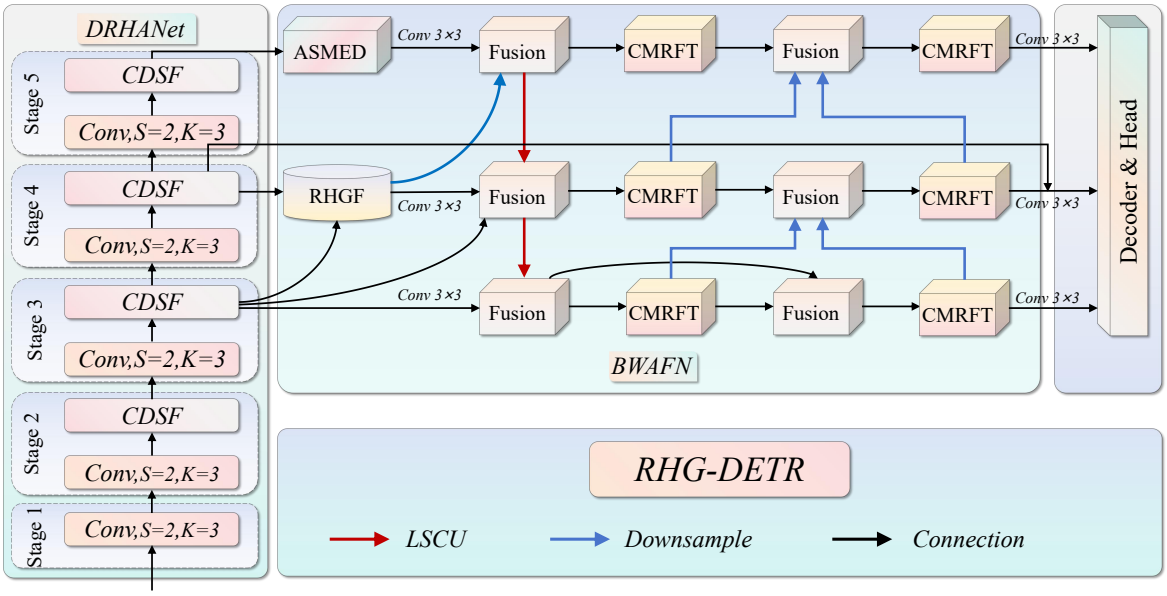


**Figure 3.** Overall architecture diagram of RHG-DETR. The framework introduces DRHANet, BWAFN, and ASMED as three innovation modules at the backbone, neck, and encoder levels respectively, cooperatively forming an end-to-end special target detection framework for composite degradation environments.

### *3.1. Dynamic Receptive-field Hyper-graph Attention Network*

The ResNet backbone adopted by conventional RT-DETR relies on fixed-receptive-field residual stacking, exhibiting fundamental limitations under multi-type degradation: fixed kernel sizes cannot dynamically adjust the spatial receptive range according to the directional distribution of input features, resulting in severely insufficient capacity for simultaneous modeling of multi-scale local texture details and cross-layer semantic correlations. This manifests as systematic feature representation degradation in occlusion and component-level localization tasks. To this end, we propose DRHANet (Dynamic Receptive-field Hyper-graph Attention Network), which replaces conventional residual units with dynamic receptive field mixing blocks to strengthen local multi-scale feature extraction and introduces a hyper-graph attention aggregation mechanism at key feature pyramid fusion nodes to model high-order semantic topology dependencies among target components—synergistically enhancing the detection capability for special targets and their specific sub-components under multi-type degradation conditions from both feature representation and semantic association dimensions, as shown in Figure 4.

CDSF splits the input features along the channel dimension and feeds them sequentially into $n$ DAMB modules for step-by-step transformation; all intermediate representations are accumulated along the feature flow direction and projected to produce the output. Let the hidden channel dimension be $c = \lfloor \frac{C}{2} \cdot r \rfloor$, and let the output of the $i$-th branch step be $z_i$; the feature accumulation process forms the following recursive functional:

$$z_{i+1} = \mathcal{F}^{i}_{\text{DIMB}}\left( z_i,\ Z = W_2 \cdot \int_{\mathcal{C}} \sum_{\delta(c-c_{z_j}) \mid z_j \in \{0,\ldots,i+1\}} dz \right) \tag{1}$$

where the integral form describes the measure accumulation of the discrete feature flow along channel space $\mathcal{C}$, $\delta(\cdot)$ is the Dirac kernel, and $W_2 \in \mathbb{R}^{C^2 \times (i+2)C}$ is the aggregation projection matrix. This structure ensures that shallow spatial details and deep semantic abstractions are simultaneously preserved at the output end, alleviating the progressive ablation of fine-grained features in deep sequential networks.

 doi:10.20944/preprints202604.1731.v1



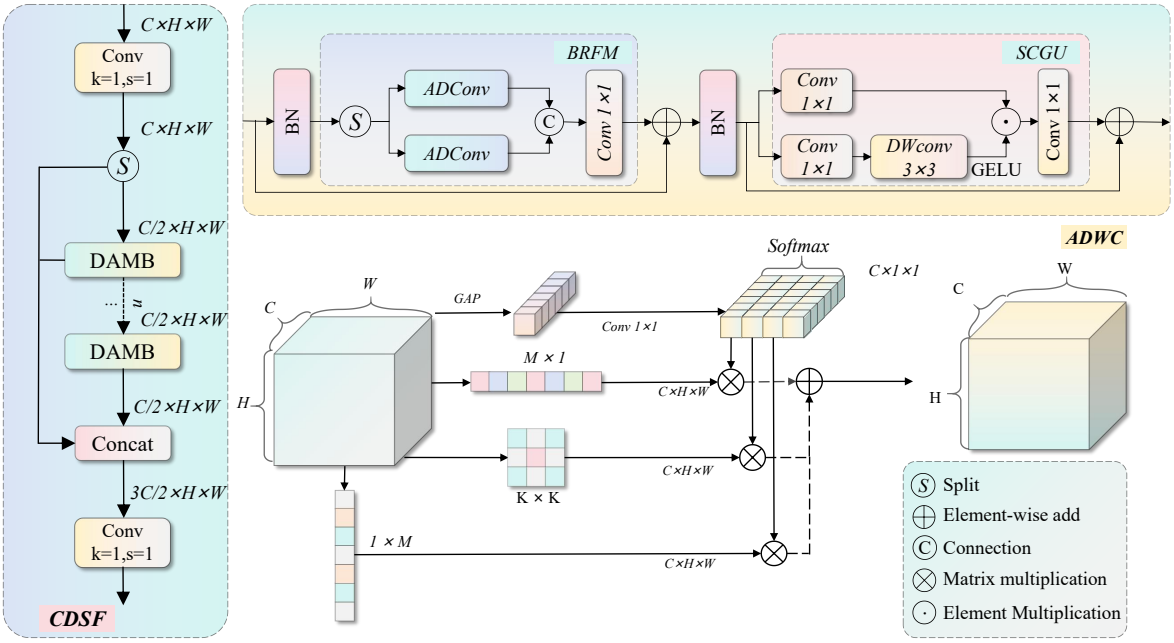


**Figure 4.** Internal structural diagram of the CDSF module and ADWC module within DRHANet.

Each DAMB (Dual-path Adaptive Mixing Block) sequentially applies a spatial mixing transform $\mathcal{T}_1$ and a gated feedforward transform $\mathcal{T}_2$ through dual residual paths, introducing per-channel learnable scale parameters $\lambda = \{\lambda_1, \lambda_2\} \in \mathbb{R}^C$ to independently modulate the residual contribution of each path, enabling adaptive allocation of the feature refinement strength of different transformation paths across the channel manifold. The two-stage residual update forms the following functional superposition:

$$x'' = x + \sum_{i=1}^{2} \int_{\mathcal{C}} \lambda_i(c) \cdot \mathcal{T}_i\left(x^{(i-1)}\right) dc \tag{2}$$

where $\lambda_i(c)$ is the continuous scale density function of the $i$-th path at channel $c$; the contributions of the spatial mixing and gated feedforward transforms accumulate independently on the channel manifold, achieving decoupled control of heterogeneous transformation paths.

BRFM (Bi-group Receptive-field Mixer) uniformly splits the input along the channel dimension into two groups, feeding them in parallel into ADWC branches with kernel scales $k_1$ and $k_2$ respectively. The two branches form local responses with different receptive field radii at the same spatial position, and the cross-group responses are fused through spatially weighted integration before projection output:

$$Z_{\text{mix}} = W \cdot \int_{\mathcal{X}} w(x) \cdot \left[\mathcal{F}_{k_1}(x_1; x) \oplus \mathcal{F}_{k_2}(x_2; x)\right] dx \tag{3}$$

where $\mathcal{X}$ is the spatial domain, $w(x)$ is an adaptive weight jointly determined by the response confidence of the two paths at position $x$, and the integral form characterizes the point-wise cooperation of local responses under different receptive field radii across the spatial domain, enabling complementary multi-scale local structural perception of special targets at the same position under complex blurred backgrounds. Within each ADWC (Anisotropic Dynamic Depthwise Convolution) branch, the dynamic weighted fusion of three anisotropic transforms is defined as:

$$y_x = \sum_{i=1}^{3} \frac{\exp(a_i(x))}{\sum_{j=1}^{3} \exp(a_j(x))} \cdot \mathcal{T}_i(x_i) \tag{4}$$

where $\mathcal{T}_i$ is the $i$-th anisotropic transform operator; the normalized weights adaptively distribute the contribution ratio of each directional receptive field based on global context statistics, exhibit-

 doi:10.20944/preprints202604.1731.v1



ing a prominent perceptual advantage for directionally salient structural features of special targets. Figure 5(a) shows the RHGF module and Figure 5(b) shows the DWSB module.

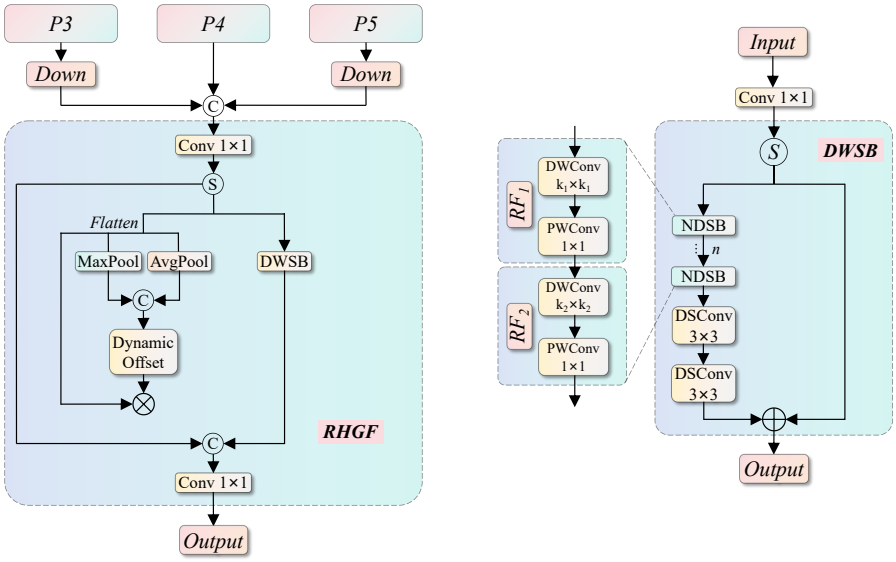


**Figure 5.** Internal structural diagrams of the RHGF module (a) and the DWSB module (b).

RHGF (Riemannian Hyper-graph Fusion) takes adjacent three-scale features $\{x^{(l-1)}, x^{(l)}, x^{(l+1)}\}$ as input and, after scale alignment, performs cross-scale measure fusion in the Riemannian manifold sense:

$$\bar{x} = W_{\text{fuse}} \cdot \exp_p\left(\sum_{l\in\{l-1,l,l+1\}} \log_p\left(\mathcal{P}_l\left(x^{(l)}\right)\right)\right) \tag{5}$$

where $\exp_p(\cdot)$ and $\log_p(\cdot)$ are the Riemannian exponential map and logarithmic map based at point $p$, elevating the aligned fusion to geodesic interpolation on the feature manifold. The fused representation is split into three paths and fed into a local refinement chain and dual hyper-graph attention branches; the joint output is:

$$Z = W_{\text{out}} \cdot \left[\sum_{k=1}^{n} \mathcal{B}^{(k)}(y) + \int_0^1 \frac{d}{d\tau}\mathcal{H}(y_1;\,\tau)\,d\tau\right] \tag{6}$$

where $\mathcal{H}(y_1;\tau)$ is the path-integral form of the hyper-graph attention branch with respect to the continuous interpolation parameter $\tau$, whose physical meaning is the continuous accumulation of target component semantic dependencies along the hyper-graph propagation path. This mechanism holds a decisive modeling advantage over standard convolutional operators in cross-region semantic binding between the tank turret–hull and the rocket system launch pod–cradle.

DWSB (Dilated depthwise-separable Split Block) adopts a dual-path split structure, where the nonlinear transformation path is implemented through a function composition chain of $n$ NDSB (Nested Dilated Separable Bottleneck) units:

$$Z_{\text{DWSB}} = W_3 \cdot \left[\left(\circ_{k=1}^{n}\mathcal{B}^{(k)}\right)(W_1 \cdot x) + W_2 \cdot x\right] \tag{7}$$

Each NDSB expands the effective receptive field through a series connection of two levels of dilated depthwise separable transforms; the receptive field expansion process satisfies the following nested integral relation:

$$y = \int_{\mathcal{X}} G_{k_2,d_2}(x',x)\left[\int_{\mathcal{X}} G_{k_1,d_1}(x'',x')\,x''\,dx''\right]dx' + x \cdot \mathbb{1}_{[c_1=c_2]} \tag{8}$$

 doi:10.20944/preprints202604.1731.v1



where $G_{k,d}(\cdot,\cdot)$ is the Green's function kernel determined by kernel size $k$ and dilation rate $d$; the two-level nested integral characterizes the progressive spatial expansion mechanism of the receptive field, with both components jointly constituting the lightweight local structure refinement subnetwork inside RHGF, providing structurally salient local priors for the hyper-graph attention branch.

DRHANet, through the organic fusion of dynamic anisotropic receptive field learning, layer-scaled residual modulation, and hyper-graph high-order semantic topology modeling, systematically strengthens the backbone network's feature extraction capability for special targets against complex degraded backgrounds from three dimensions: feature extraction accuracy, occlusion robustness, and component semantic decoupling. This provides multi-scale image features of stronger discriminability and higher structural integrity for subsequent neck multi-scale fusion.

### *3.2. Bi-Directional Weighted Adaptive Fusion Network*

When performing multi-scale detection of special targets—such as tanks with their turrets, rocket systems with their launch pods, and soldiers—under multi-type degradation conditions, the intrinsic challenges of maintaining cross-scale semantic consistency, dynamically balancing scale contributions, and preserving fine boundary structures collectively constitute the core problems demanding solution in neck network design. Unidirectional fusion paths permit features to undergo only one cross-scale propagation, leaving alignment errors between shallow details and deep semantics uncorrectable in subsequent layers; under composite degradation interference, this deficiency is particularly damaging to semantic consistency. Fixed equal-weight summation renders the network incapable of adaptive response to changes in target scale distribution, and in scenes where multi-granularity targets coexist—such as complete tank bodies alongside turret regions, or complete rocket systems alongside launch modules—it readily induces feature dominance imbalance. The spatial activation inconsistencies and parameter redundancy introduced by conventional transposed convolution upsampling during spatial resolution restoration further weaken boundary localization accuracy for fine local structures. To address these issues, we propose BWAFN (Bi-directional Weighted Adaptive Fusion Network), which systematically resolves the feature fusion problems of the neck network under composite degradation conditions through the cooperative design of a bidirectional adaptive weighted fusion path (Fusion module), multi-scale node feature transforms, and a lightweight spatial compensation upsampling mechanism.

BWAFN constructs a two-stage bidirectional pyramid on the three-scale features output by the backbone network. The Fusion module adaptively integrates multi-level features from different input branches through five feature fusion strategies—weight, adaptive, concat, BiFPN, and SDI—employing learnable weights or attention mechanisms to balance the contribution of each feature map, as illustrated in Figure 6. LSCU propagates along the top-down direction to generate intermediate fusion nodes carrying deep semantics, then performs a second enhancement of detail semantics along the bottom-up direction and outputs the final multi-scale features, as shown in Figure 7.

 doi:10.20944/preprints202604.1731.v1



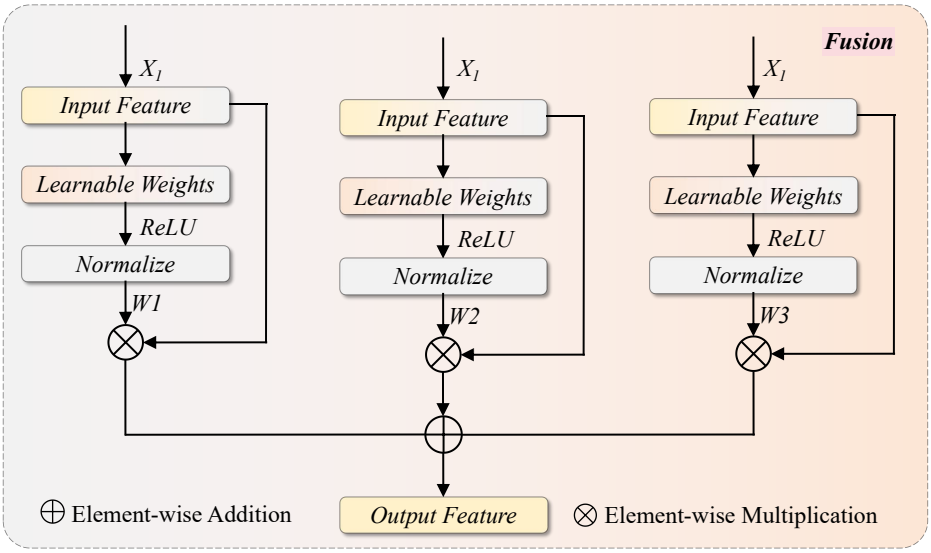


**Figure 6.** Structural diagram of the Fusion module.

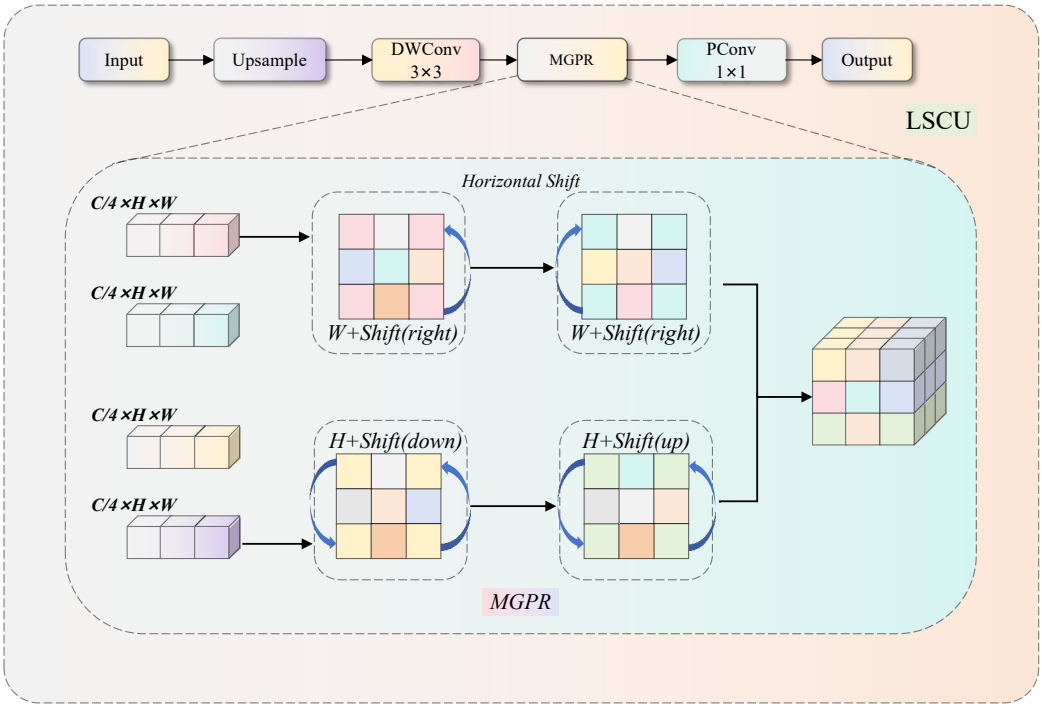


**Figure 7.** Structural diagram of the LSCU module.

Each fusion node is equipped with learnable scale contribution weights; the adaptive weight allocation and fusion output at layer $l$ are expressed as:

$$\alpha_{l,s} = \frac{w_{l,s}}{\epsilon + \sum_{j \in \mathcal{S}_l} w_{l,j}}, \qquad F_l^{\text{out}} = \mathcal{N}_l\left(\sum_{s \in \mathcal{S}_l} \alpha_{l,s} \cdot \phi_s\left(F_s^l\right)\right) \tag{9}$$

where $\alpha_{l,s}$ is the dynamic contribution weight of scale $s$ at layer $l$ after normalization, $w_{l,s} > 0$ is the learnable weight that updates adaptively during training, $\mathcal{S}_l$ is the set of input scales at this layer's fusion node, $\mathcal{N}_l$ is the node-level feature transform based on CMRFT (Cross-stage Multi-scale Receptive-field Transform), and $\epsilon$ is a numerical stability term. Within the overall architecture, CMRFT handles the feature transform at each fusion node, Multi-scale Grouped Permutation Recomposer (MGPR) is responsible for multi-receptive-field expansion inside each node, LSCU and SCAG cooperatively

 doi:10.20944/preprints202604.1731.v1



accomplish lightweight upsampling and spatial context compensation along the top-down path, and all four components fulfill their respective roles with clear hierarchical structure. The CMRFT module is shown in Figure 8.

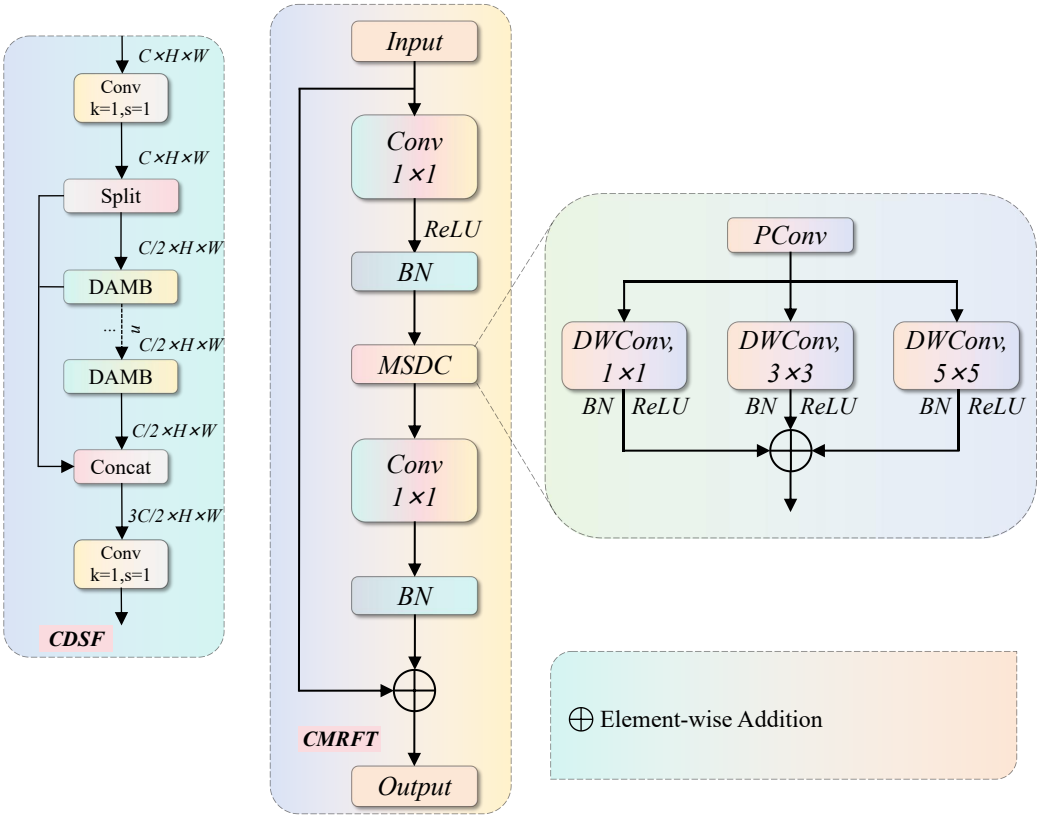


**Figure 8.** Structural diagram of the CMRFT module.

Single-receptive-field node transforms struggle to accommodate the feature response requirements at different spatial granularities when confronted with scenes containing multi-granularity target structures such as turrets alongside complete vehicles, or individual soldiers alongside groups. CMRFT takes cross-stage local structures as its framework, with the direct-connection branch $Y_l^0$ preserving the low-frequency semantic base of the input, while the multi-scale branch applies $n$ MSRB (Multi-Scale Receptive-field Block) units to $Y_l^1$ for progressive receptive field expansion. The two branch outputs, together with all intermediate features at each level, are uniformly mapped through projection; the cumulative iterative fusion process is expressed as:

$$Z_l = \mathcal{N}_l\left(Y_l^0 + \sum_{i=1}^{n} \mathcal{F}_i \circ \mathcal{F}_{i-1} \circ \cdots \circ \mathcal{F}_1\left(Y_l^1\right)\right) \tag{10}$$

where $\mathcal{F}_i$ is the $i$-th MSRB transform, $\circ$ is the functional composition operator, and the cumulative summation term stacks the outputs of iterative composite mappings at each depth level, allowing multi-scale responses at different depths to all participate in the final fusion. The presence of the direct-connection branch preserves low-frequency semantics after multi-layer transformation, forming a complementary rather than substitutive relationship with the multi-scale branch, thereby effectively suppressing progressive target feature degradation under low-contrast backgrounds.

Effective aggregation of multi-scale responses relies on applying $K$ grouped perceptual operators of different receptive field scales in parallel to the expanded feature $X_l$, with each branch output integrated across scales through element-wise summation while keeping the channel dimension unchanged:

 doi:10.20944/preprints202604.1731.v1



$$D_l = \sum_{k=1}^{K} G_k(X_l) \tag{11}$$

where $G_k$ is the grouped perceptual operator at scale $k$. The aggregated feature responses, with grouping granularity $g_l = \gcd(C_{\text{in}}, C_{\text{out}})$, undergo a linear transform under the permutation group through block permutation matrix $P_{g_l}$, reallocating the feature responses of each scale to different channel positions by group:

$$D'_l = P_{g_l} D_l P_{g_l}^{\top} \tag{12}$$

Since $P_{g_l}$ satisfies $P_{g_l} P_{g_l}^{\top} = I$, this transform is an isometric linear map that fully cross-recombines the feature responses of each receptive field branch into a new channel arrangement without introducing any additional parameters.

Spatial activation inconsistencies and parameter redundancy during spatial resolution restoration constitute intrinsic limitations of transposed convolution upsampling in fine local structure localization. LSCU replaces transposed convolution with a sequential combination of nearest-neighbor upsampling and grouped convolution, with SCAG embedded after the grouped convolution to perform spatial context aggregation on the upsampled features. SCAG uniformly partitions features into four subsets along the channel dimension and applies corresponding directional linear representation operators $\rho(g)$ to each channel subset under the discrete translation group $G_s = \{+\Delta h, -\Delta h, +\Delta w, -\Delta w\}$, completing four-direction spatial context aggregation:

$$\mathcal{M}_s(X_l) = \sum_{g \in G_s} \rho(g) \cdot X_l^g \tag{13}$$

where $G_s$ is the discrete translation group generated by stride $s$, $\rho(g)$ is the linear representation operator of group element $g$ on the feature space, and $X_l^g$ is the channel subset corresponding to direction $g$. By Burnside's lemma from group representation theory, the complete summation of linear representations over group $G_s$ renders the output feature spatially equivariant under $G_s$ action, and the symmetrically complete four-direction aggregation achieves effective integration of local spatial context during upsampling without introducing any learnable parameters.

The design of each sub-module in BWAFN progresses layer by layer according to task requirements: the bidirectional weighted fusion path establishes a bidirectional correction mechanism for cross-scale semantics at the path structure level; CMRFT expands receptive field diversity at each node and suppresses feature degradation through the complementary structure of direct connection and iterative composition; MGPR further promotes cross-channel recombination of multi-receptive-field responses through permutation group linear transforms; and LSCU with SCAG compensate for activation inconsistencies during resolution restoration through spatial context aggregation under a group representation framework in the upsampling path. The four-level progressive design jointly acts on the feature expression quality of the multi-type degraded neck network, providing structurally complete multi-scale inputs for the subsequent transformer decoder.

### *3.3. Adaptive Sparse Multi-scale Encoder with Dynamic Normalization*

Standard transformer encoders exhibit several inherent common limitations in feature interaction modeling. Global self-attention performs indiscriminate dense interaction across all position pairs; in semantically sparse scenes, redundant activations in background regions continuously interfere with target region feature responses through the attention path, causing the target semantic signals in the encoder output feature representations to be overwhelmed by background noise. Standard feedforward networks are entirely defined in the linear mapping space of the channel dimension, and the spatial topological structure of feature activations is irreversibly lost during the feedforward stage, making fine semantics that depend on spatial geometric distributions—such as turret contour

 doi:10.20944/preprints202604.1731.v1



constraints and launch module relative position relationships—impossible to effectively preserve during encoding.

The linear normalization assumption of LayerNorm produces statistical mismatch under high dynamic range activation scenarios in degraded remote sensing images, causing the activation separability between target and background regions in deep feature maps to continuously decline. These limitations are significantly amplified under composite degradation conditions by structural factors such as strong target sparsity, fine sub-component semantics, and non-stationary activation distributions, motivating us to systematically reconstruct the AIFI module and propose the ASMED (Adaptive Sparse Multi-scale Encoder with Dynamic normalization) encoder, establishing an adaptive encoding framework for structural feature distributions under multi-type degradation conditions, as shown in Figure 9.

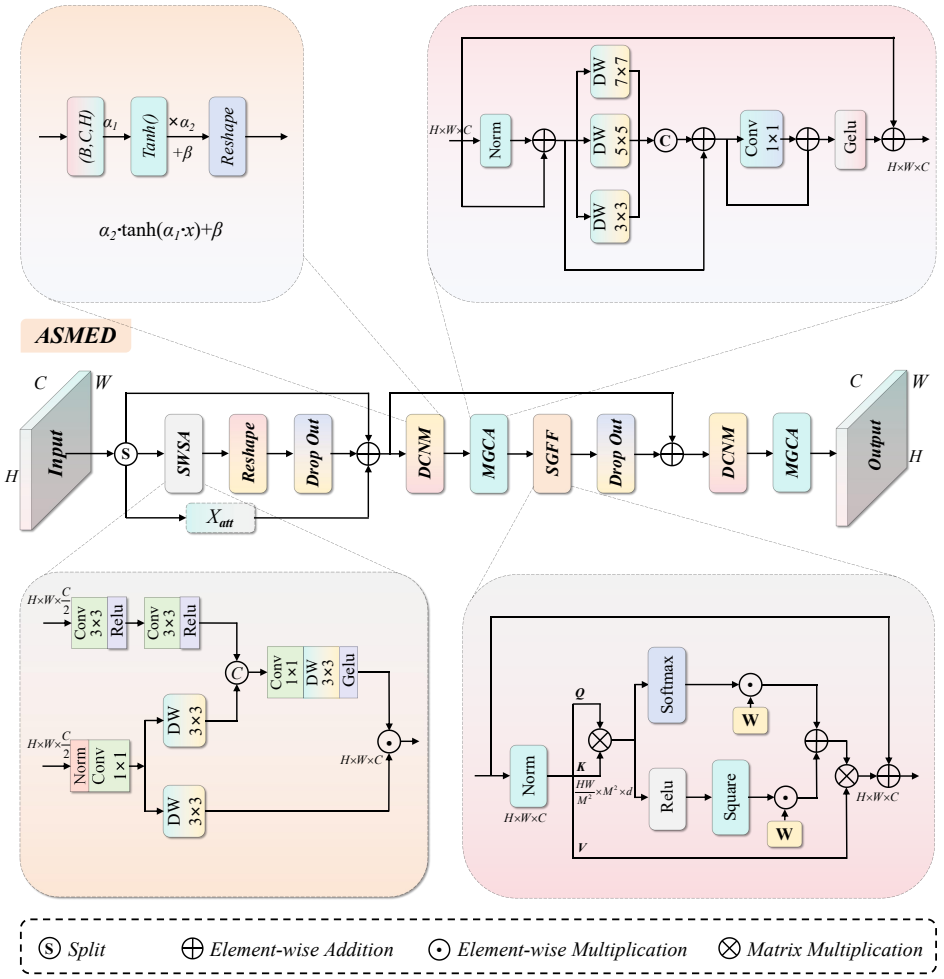


**Figure 9.** Overall structural diagram of the ASMED encoder, illustrating the two-stage residual pipeline organization of the SWSA sparse attention operator, SGFF spatial-gated feedforward fusion module, MGCA multi-scale gated coupling adaptor, and DCNM dynamic channel-wise nonlinear modulator.

The proposed module takes input feature map $X \in \mathbb{R}^{B\times C\times H\times W}$ as its starting point and completes feature encoding through a two-stage residual pipeline. Denoting $\mathcal{A}(\cdot)$, $\mathcal{F}(\cdot|\cdot)$, $\mathcal{M}(\cdot)$, and $\mathcal{N}(\cdot)$ as the SWSA sparse attention operator, the SGFF feedforward operator conditioned on the original spatial feature, the MGCA geometric adaptor, and the DCNM normalization operator respectively, the overall forward computation is:

$$Z_{\text{out}} = \mathcal{M}_2(\mathcal{N}_2(\mathcal{M}_1(\mathcal{N}_1(X + \mathcal{A}(X))) + \mathcal{F}(\mathcal{M}_1(\mathcal{N}_1(X + \mathcal{A}(X))),\, X))) \tag{14}$$

 doi:10.20944/preprints202604.1731.v1



where $\mathcal{M}_1$ and $\mathcal{M}_2$ are two structurally independent MGCA operator instances with non-shared parameters, embedded after the two normalization components respectively, and independently applying geometric structure modulation to intermediate feature representations through residual paths during both the attention encoding stage and the feedforward transformation stage.

The semantic energy of special targets is highly concentrated in local regions of the feature space; indiscriminate aggregation at low-response positions by global self-attention continuously interferes with the feature representation of target semantics. SWSA introduces a sparse attention constraint within a window partitioning framework, restricting the effective support domain of attention weights to the subset $\mathcal{S}_i \subseteq \Omega_i$ of positions satisfying the saliency condition within each local window, with a cyclic shift mechanism ensuring feature connectivity between adjacent windows. Denoting the sparse response weight distribution of the $h$-th attention head in the $i$-th window over spatial domain $\Omega_i$ as $a_{q,k}^{i,h}$, the global attention output is defined as:

$$\mathcal{A}(x) = \sum_{i=1}^{N_w} \sum_{h=1}^{N_{\text{head}}} W_h \int_{\Omega_i} a_{q,k}^{i,h} \cdot V_k^{i,h} \, dk \tag{15}$$

where the sparse response weights $a_{q,k}^{i,h}$ are given by the saliency-constrained normalized similarity:

$$a_{q,k}^{i,h} = \frac{\exp\Big(\langle \nabla_q q^{i,h}, \nabla_k k^{i,h} \rangle + b_{qk}\Big)}{\sum_{k' \in \mathcal{S}_i} \exp\Big(\langle \nabla_q q^{i,h}, \nabla_{k'} k^{i,h} \rangle + b_{qk'}\Big)} \tag{16}$$

where $\mathcal{S}_i$ is the set of valid positions filtered by saliency constraints, $\langle \cdot, \cdot \rangle$ denotes the inner product operation, $\nabla_q q^{i,h}$ and $\nabla_k k^{i,h}$ are gradient-aware representations of query and key vectors at positions $q$ and $k$, and $b_{qk}$ is a learnable relative position bias. Contracting the integration aggregation domain from $\Omega_i$ to $\mathcal{S}_i$ makes feature interactions occur strictly within locally high semantically responsive regions, thereby suppressing the interference of background noise on target feature representations.

The feature transformation process of a standard feedforward network is strictly confined to the channel-linear mapping space, where geometric correlations between spatial positions are invisible, preventing sub-component semantics of targets that depend on spatial structure distributions from being effectively encoded during the feedforward stage. SGFF introduces a parallel spatial branch, generating multi-scale spatial structural semantic representations from the pre-attention feature $X_0$ through $L$ stacked convolutional layers, and performing adaptive gated fusion with channel activations. The cumulative response of the spatial branch along the scale dimension is recorded as $Y_s = \sum_{i=1}^{L} \mathcal{T}_i(X_0)$, and the gated fusion output is defined via a functional integral of continuous geometric responses over the spatial domain:

$$\mathcal{F}(X, X_0) = \int_{\mathcal{X}} \left[ \sigma\left( \frac{\delta \mathcal{L}_s}{\delta X}(x) \right) \cdot Y_s(x) \cdot X(x) \right] dx \tag{17}$$

where $\sigma(\cdot)$ is the GELU nonlinear activation, $\frac{\delta \mathcal{L}_s}{\delta X}(x)$ is the variational derivative of the feature response at spatial position $x$ with respect to the spatial structure functional $\mathcal{L}_s$, quantifying the local sensitivity of the current channel activation to spatial geometric structure semantics, and $Y_s(x)$ provides the structural modulation signal from the multi-scale spatial branch. This functional integral introduces an explicit constraint on spatial geometric distributions in addition to the channel mapping during the feedforward transformation, allowing spatially discriminative cues of sub-components such as turret contours and launch modules to be preserved during the feedforward encoding stage.

The shared challenge of non-stationary activation distributions unifies the design motivations of DCNM and MGCA under the functional positioning of stabilizing the encoding process. DCNM intervenes at the normalization level, applying input-dependent nonlinear compression to each channel's activations through learnable parameter $\alpha$; its normalization output is given by:

 doi:10.20944/preprints202604.1731.v1



$$\mathcal{N}(x) = \sum_{c=1}^{C} \alpha_c \cdot \tanh\left( \int_0^{\alpha_c} \frac{d}{da} \tanh(a \cdot x_c) da + \beta_c \right) + \beta_c \tag{18}$$

where the integration path from zero to the current learnable value characterizes the cumulative effect of compression strength as it dynamically evolves through training, and $\alpha_c$ and $\beta_c$ are per-channel affine parameters. The bounded monotonicity of tanh guarantees effective compression of the normalization output for arbitrary-magnitude activations, exhibiting more stable feature calibration properties under high dynamic range activation scenarios compared to linear normalization. MGCA intervenes at the feature flow level, completing multi-receptive-field feature aggregation through grouped convolution $\mathcal{G}$, and controlling the dynamic coupling ratio of the main feature flow and the normalization branch through dual learnable gating scale parameters:

$$\mathcal{M}(X) = X + \sum_{l=1}^{n} \int_0^1 \frac{\partial}{\partial \tau} \left[ \gamma_l^f \cdot X_l + (1 - \gamma_l^n) \cdot \mathcal{N}(X_l) \right] d\tau \tag{19}$$

where $\tau \in [0, 1]$ is the interpolation path parameter, and the integral form characterizes the dynamic coupling process between the main feature flow $\gamma_l^f \cdot X_l$ and the normalization modulation branch $\gamma_l^n \cdot \mathcal{N}(X_l)$ along the continuous interpolation path. The dual gating scale parameters $\gamma_l^f$ and $\gamma_l^n$ independently control the contribution weights of the main feature flow and the normalization modulation branch to the final output; their complementary cooperation at the normalization level and the feature flow level jointly suppresses the impact of non-stationary activation distributions on encoding stability.

The overall design of the proposed encoder is organized around two core detection difficulties for special targets under multi-type degradation: missed detections caused by camouflaged targets and low-contrast background boundaries, and false detections caused by semantic confusion of target sub-components under composite degradation and local occlusion. The sparse attention constraint contracts feature interactions to high semantic response regions, preventing the weak activation signals of low-contrast targets from being diluted by background noise, suppressing missed detections from the attention path. The synergistic action of DCNM and MGCA jointly maintains the effectiveness of the above mechanism under non-stationary activation distributions at both the normalization level and the feature flow level, ensuring that the activation separability between targets and backgrounds does not significantly decay under extreme conditions such as dramatic illumination changes and atmospheric scattering.

## 4. Experiments

### *4.1. Public Datasets*

VisDrone2019 [25] is a large-scale UAV visual benchmark dataset constructed by the AISKYEYE team of the Machine Learning and Data Mining Laboratory at Tianjin University, designed for multiple computer vision tasks in UAV image and video analysis. For the object detection subtask (VisDrone2019-DET), the dataset is partitioned into four subsets: training set (6,471 images), validation set (548 images), test-dev set (1,610 images), and test-challenge set (1,580 images), with annotated categories covering 10 classes including pedestrians, vehicles, bicycles, and tricycles. The dataset ensures data diversity through varying weather conditions, illumination conditions, and flight altitudes, and is broadly representative in geographical distribution, environmental background, and capture viewpoint, making it one of the most representative public benchmarks in UAV remote sensing object detection.

BDD100K [42] was collected by the University of California, Berkeley through crowdsourcing, comprising 100,000 annotated keyframe images partitioned at a 7:1:2 ratio into training (70,000 images), validation (10,000 images), and test sets (20,000 images), with annotated categories covering 10 semantic classes including vehicles, pedestrians, riders, traffic signs, and traffic lights. The dataset exhibits significant diversity across three dimensions—weather conditions, scene types, and illumination

 doi:10.20944/preprints202604.1731.v1



states—with approximately half of the targets exhibiting occlusion and approximately 7% exhibiting truncation, making it an ideal benchmark for evaluating model cross-domain robustness.

### *4.2. Implementation Details and Training Configuration*

All experiments were conducted on a system running Ubuntu 22.04, with Python 3.10.16 and PyTorch 2.2.2. The hardware configuration consisted of one NVIDIA GeForce RTX 3090 GPU with CUDA version 12.1. During training, the AdamW optimizer was used with a learning rate of 0.0001 and a batch size of 8. The model was trained for a total of 300 epochs. The weight decay coefficient was set to 0.0001. All experiments used the same random seed to ensure reproducibility. Other hyperparameters adopted the default settings of RT-DETR. Frames per second (FPS) evaluation was performed on an RTX 3090 GPU with an input resolution of 640×640 pixels and a batch size of 1. All inference time measurements excluded data loading and post-processing time to ensure fair comparison across different methods.

### *4.3. Evaluation Metrics*

To comprehensively evaluate the detection performance of the proposed RHG-DETR framework, a set of standard quantitative metrics widely used in the object detection field was adopted. Precision (P) measures the proportion of true positive predictions among all predicted bounding boxes, reflecting the detector's accuracy in avoiding false detections. Recall (R) quantifies the proportion of correctly detected targets relative to all ground-truth annotated instances, characterizing the model's capacity to identify targets without omission. The mean average precision at an intersection over union (IoU) threshold of 0.5 ($mAP_{50}$, also denoted mAP@0.5) serves as the primary performance metric, providing a balanced comprehensive assessment of localization and classification accuracy under relatively lenient overlap criteria. Additionally, the mean average precision computed at IoU thresholds ranging from 0.5 to 0.95 in steps of 0.05 ($mAP_{50\text{-}95}$, also denoted mAP@0.5:0.95) is adopted for a stricter evaluation of bounding box localization quality.

Beyond detection accuracy metrics, computational efficiency is evaluated through floating-point operations (GFLOPs) and total model parameter count (Params). These metrics reflect the computational complexity and memory footprint of the model, which are critical considerations for deploying models in resource-constrained remote sensing application scenarios. Employing all of the above evaluation metrics enables a comprehensive and multi-dimensional assessment of the effectiveness and practicality of RHG-DETR in special target detection from remote sensing images under multi-type degradation conditions.

### *4.4. Ablation Study*

#### 4.4.1. Convolutional Kernel Ablation for Internal Components of DRHANet

To investigate the impact of convolutional kernel configuration within DRHANet on special target detection performance, we designed systematic ablation experiments around square kernels, asymmetric strip kernels, and multi-kernel hybrid strategies to validate the effectiveness of the multi-kernel dynamic fusion mechanism in ADWC. As shown in Table 1, six configurations were evaluated, spanning single square kernels, strip kernel extensions, and multi-kernel hybrid configurations.

**Table 1.** Ablation results for different convolutional kernel configurations of DRHANet internal components.

| Config. | Square Kernel | H-Band | V-Band | GFLOPs | Param/M | $mAP_{50}$ | $mAP_{50\text{-}95}$ |
|---|---|---|---|---|---|---|---|
| 1 | 3×3 | | | 45.2 | 13.11 | 74.3 | 41.5 |
| 2 | 3×3 | 1×11 | | 45.8 | 13.44 | 74.6 | 42.2 |
| 3 | 5×5 | | | 45.5 | 13.29 | 75.0 | 40.9 |
| 4 | 5×5 | 1×17 | 17×1 | 46.1 | 14.37 | 74.7 | 40.4 |
| 5 | 3×3 | 1×17 | 17×1 | 45.3 | 13.98 | 75.2 | 41.6 |
| 6 (ours) | 3×3 | 1×11 | 11×1 | 46.8 | 14.52 | 75.8 | 42.3 |

 doi:10.20944/preprints202604.1731.v1



From the data in Table 1, the single square kernel (Configuration 1) as the baseline maintains stable precision but exhibits relatively limited recall, with $mAP_{50}$ and $mAP_{50\text{-}95}$ of 74.3% and 41.5%, respectively. Introducing asymmetric strip kernels (Configuration 2) yields a 0.3% gain in $mAP_{50}$ over the baseline, indicating that directional features make a positive contribution to special target detection. Adopting a 5×5 square kernel (Configuration 3) further improves recall, with $mAP_{50}$ increasing by 0.7% over the baseline. The multi-kernel hybrid strategy (Configuration 5), without incorporating strip kernels, already achieves a 0.9% $mAP_{50}$ improvement, reflecting the complementary perceptual advantages of multi-scale square kernels. Our proposed complete multi-kernel hybrid configuration (Configuration 6), which simultaneously fuses square kernels and two sets of asymmetric strip convolutional kernels, achieves the best performance on all evaluation metrics, with $mAP_{50}$ and $mAP_{50\text{-}95}$ improving by 1.5% and 0.8% over the baseline, respectively. This result fully validates that ADWC, through the dynamic fusion of convolutional kernels of multiple scales and orientations, can more comprehensively perceive multi-directional spatial textures and local geometric structures of special targets, achieving coordinated optimization of detection accuracy and recall while introducing only 1.41M additional parameters.

#### 4.4.2. Ablation Study of ASMED

The ASMED encoder consists of four functionally distinct sub-modules: SWSA, SGFF, MGCA, and DCNM; the independent contribution and synergistic gain of each module to the final detection performance merit thorough investigation. To this end, we conducted an ablation analysis by progressively stacking modules one by one, as shown in Table 2, designing seven different component combination schemes in addition to the baseline configuration.

**Table 2.** Ablation results for sub-modules of the ASMED encoder.

| Model | SWSA | SGFF | MGCA | DCNM | P | R | $mAP_{50}$ | $mAP_{50\text{-}95}$ |
|---|---|---|---|---|---|---|---|---|
| 1. base | | | | | 85.3 | 70.2 | 74.8 | 41.5 |
| 2 | ✓ | | | | 86.1 | 72.7 | 76.0 | 42.4 |
| 3 | | ✓ | | | 86.8 | 73.5 | 76.4 | 43.2 |
| 4 | | | ✓ | | 85.6 | 72.4 | 75.8 | 42.1 |
| 5 | | | | ✓ | 86.4 | 73.0 | 75.5 | 42.8 |
| 6 | ✓ | ✓ | | | 87.1 | 74.1 | 76.9 | 43.4 |
| 7 | ✓ | ✓ | ✓ | | 88.0 | 73.7 | 77.7 | 44.1 |
| 8 (ours) | ✓ | ✓ | ✓ | ✓ | 87.6 | 74.4 | 78.5 | 44.6 |

Analyzing the experimental data of each configuration reveals that all components produce varying degrees of positive improvement over the baseline detection performance. When SGFF is introduced alone, it yields the most significant $mAP_{50}$ improvement of 1.6%, demonstrating the outstanding contribution of spatial-gated feedforward fusion to preserving the geometric structure semantics of target components; when SWSA is introduced alone, $mAP_{50}$ improves by 1.2%, reflecting the effectiveness of sparse attention constraints in suppressing complex background noise interference; and the introduction of MGCA and DCNM provides strong support for encoding stability in terms of geometric adaptive modulation and dynamic normalization, respectively. As components are progressively stacked, SWSA combined with SGFF improves $mAP_{50}$ by 2.1% over the baseline, and further adding MGCA yields a 2.9% improvement, fully demonstrating the complementary cooperative effect among the modules. The optimal configuration with all four sub-modules achieves the highest scores on precision, recall, and mAP metrics, with $mAP_{50}$ and $mAP_{50\text{-}95}$ improving by 3.7% and 3.1% over the baseline, respectively. These results fully demonstrate that the cooperative design of sparse attention mechanism, spatial-gated feedforward fusion, geometric adaptive modulation, and dynamic channel-wise nonlinear normalization in ASMED can effectively decouple target regions from background noise under complex blurred military backgrounds and enhance the encoder's capacity to represent structural semantics of target sub-components.

 doi:10.20944/preprints202604.1731.v1



4.4.3. Overall Ablation Study

DRHANet, BWAFN, and ASMED each operate at the feature extraction, multi-scale fusion, and semantic encoding levels respectively; their individual independent gains and mutual synergistic effects are key evidence for evaluating the rationality of the overall architecture design. We designed systematic overall ablation experiments with single-module and multi-module joint introduction, as shown in Table 3 and Figure 10, using RT-DETR-R18 as the baseline for a comprehensive comparison of seven module combination configurations from both detection accuracy and model lightweight dimensions.

**Table 3.** Overall ablation results for the three core modules DRHANet, BWAFN, and ASMED.

| Model | DRHANet | BWAFN | ASMED | GFLOPs | Param/M | P | R | $mAP_{50}$ | $mAP_{50-95}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1. base | | | | 57.3 | 19.88 | 85.3 | 70.2 | 74.8 | 41.5 |
| 2 | ✓ | | | 46.8 | 14.52 | 86.7 | 72.0 | 75.8 | 42.3 |
| 3 | | ✓ | | 48.6 | 17.11 | 85.9 | 71.5 | 76.5 | 42.0 |
| 4 | | | ✓ | 58.7 | 21.31 | 86.4 | 72.1 | 76.3 | 43.1 |
| 5 | ✓ | ✓ | | 36.2 | 11.89 | 87.3 | 73.0 | 77.2 | 42.8 |
| 6 | ✓ | | ✓ | 48.2 | 16.79 | 86.5 | 73.8 | 77.8 | 43.9 |
| 7 | | ✓ | ✓ | 50.0 | 19.38 | 87.0 | 74.7 | 78.1 | 44.2 |
| 8 (ours) | ✓ | ✓ | ✓ | 37.6 | 14.16 | 87.6 | 74.4 | 78.5 | 44.6 |

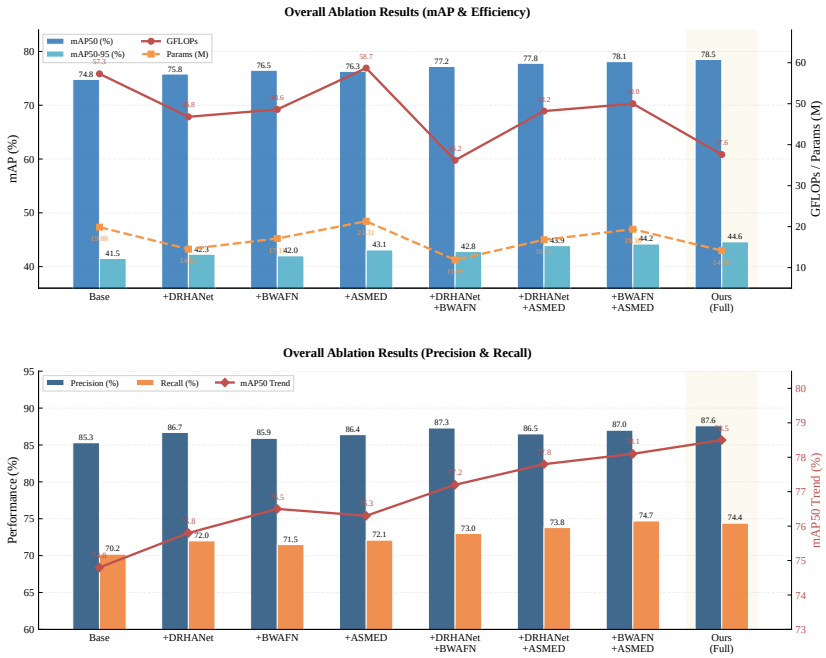


**Figure 10.** Overall ablation results for the three core modules DRHANet, BWAFN, and ASMED.

Taking all metrics together, the three modules each possess lightweight advantages while enhancing detection performance, and the synergistic gains among modules are significant. When DRHANet is introduced alone, $mAP_{50}$ improves by 1.0% over the baseline while GFLOPs and parameter count are reduced by 18.3% and 26.9%, respectively, simultaneously optimizing both detection accuracy and computational efficiency; when BWAFN is introduced alone, $mAP_{50}$ improves by 1.7%, reflecting the enhancement of bidirectional weighted fusion paths on multi-scale feature semantic consistency; and when ASMED is introduced alone, $mAP_{50}$ improves by 1.5%, validating the effectiveness of the adaptive sparse encoding mechanism at the encoder level. Notably, when DRHANet and BWAFN are jointly introduced, GFLOPs are further compressed to 36.2 and parameter count drops to 11.89M, demonstrating a remarkable lightweighting effect. The optimal configuration integrating all three mod-

 doi:10.20944/preprints202604.1731.v1



ules achieves the best results in both accuracy and efficiency dimensions, with $mAP_{50}$ and $mAP_{50-95}$ improving by 3.7% and 3.1% over the baseline respectively, while GFLOPs and parameter count are reduced by 34.4% and 28.8%. This result fully validates the synergistic complementary design philosophy of the three core modules, enabling RHG-DETR to achieve substantial performance gains while significantly reducing computational overhead, ultimately realizing a high degree of unification of accuracy and lightweighting.

To validate the effectiveness of the proposed RHG-DETR framework in special target detection tasks under multi-type degradation and reveal the advantages of its attention mechanism, we designed and conducted attention heat map visualization comparison experiments. Several representative complex degradation detection scenes were selected; the relevant visualization results are shown in Figure 11. The first column represents clean images and the second column represents degraded images. RHG-DETR demonstrates significant advantages over RT-DETR-R18 (Base) in attention distribution accuracy and target focusing capability, and as DRHANet (+A), BWAFN (+B), and ASMED (+C) are progressively introduced, the target focusing quality of the attention heat maps exhibits a systematic progressive improvement. Degradation scenarios include: camouflaged soldiers in complex woodland under occlusion and moderate blur; tanks in the wilderness under smoke occlusion; camouflaged tanks in the field under moderate rain; camouflaged soldiers in fallow land under mild strong light; rocket systems on floodplains under moderate electromagnetic interference and moderate blur; and rocket systems on desert terrain under night vision and mild blur, among other typical challenging application scenarios.

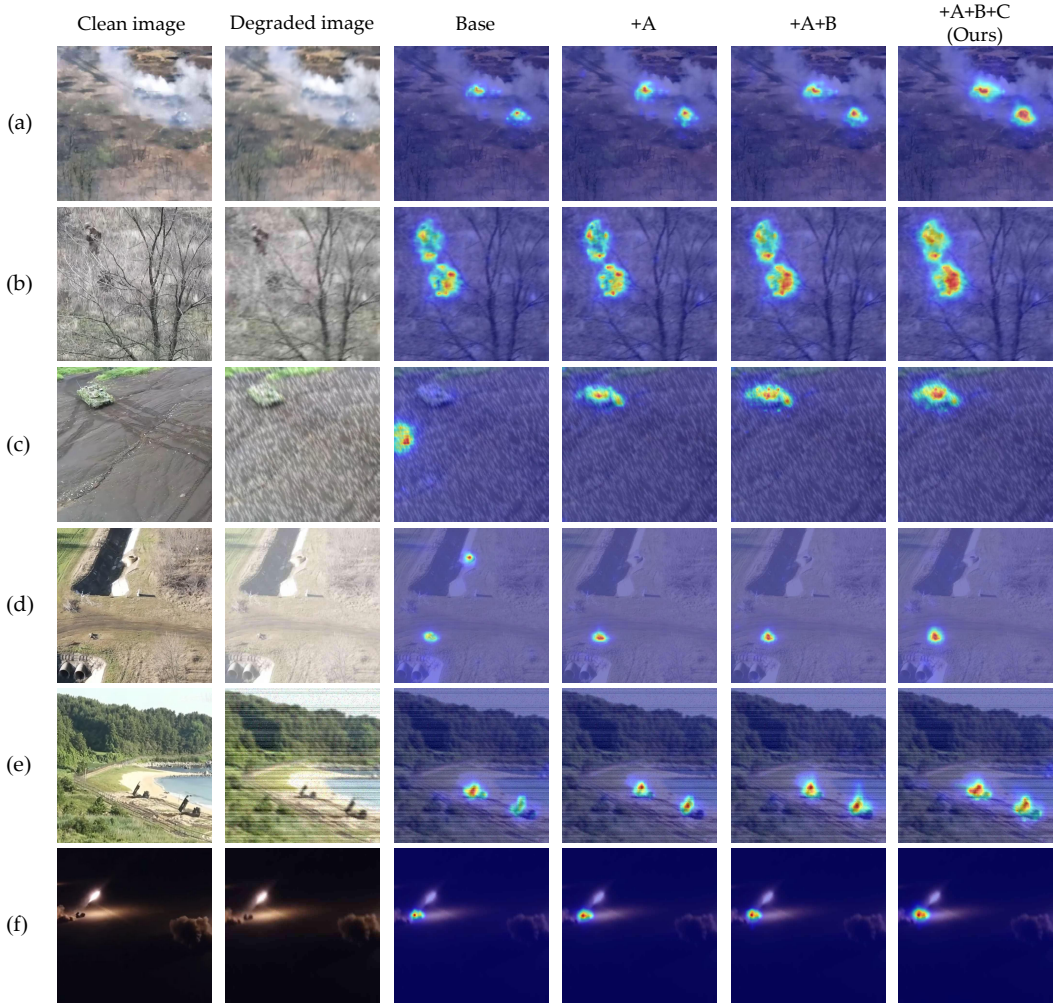


**Figure 11.** Attention heat map visualization comparison of RHG-DETR versus the RT-DETR baseline across six typical degradation scenarios.

 doi:10.20944/preprints202604.1731.v1



Through systematic comparison of the attention heat maps of RHG-DETR against the baseline model RT-DETR, we analyzed in depth the improvement of each innovation module on attention distribution and target focusing capability. In six typical degradation scenarios—mild blur smoke-occluded tanks (row a), moderate blur woodland-camouflaged soldiers (row b), moderate blur combined with rain outdoor-camouflaged tanks (row c), mild blur combined with strong light camouflaged soldiers (row d), moderate blur combined with electromagnetic interference rocket systems (row e), and moderate blur combined with low illumination rocket systems (row f)—the baseline model consistently exhibits inherent deficiencies including widely distributed attention, significant background interference, and insufficient target region activation. As DRHANet, BWAFN, and ASMED are progressively introduced, the heat maps in each scenario show systematic improvement trends: the RHGF module models high-order semantic topology dependencies among components, strengthening target main body responses; the CMRFT node maintains cross-scale semantic consistency under atmospheric degradation; the SWSA sparse window attention constraint confines feature interactions to high-saliency regions, effectively suppressing background noise; SGFF injects spatial geometric topology constraints into the feedforward stage, preserving sub-component structural co-occurrence relationships; and DCNM with MGCA cooperatively stabilize the non-stationary activation distributions under extreme imaging conditions. The complete model (+A+B+C) achieves the highest target region activation intensity, the most precise spatial localization, and the lowest background false activation across all six degradation scenarios, demonstrating clear advantages over each ablation stage. Overall, as the three modules are progressively stacked, the attention heat maps of RHG-DETR consistently exhibit a systematic evolution from scattered interference to precise focusing, fully validating the cooperative effectiveness of each module in enhancing special target attention perception capability under multi-type composite degradation conditions and their practical deployment value.

### *4.5. Comparative Experiments*

#### 4.5.1. Backbone Network Comparison

Conducting a lateral comparison of DRHANet against existing advanced backbone networks under the same neck and decoder configuration is a direct means of measuring its practical competitiveness. We selected representative backbones including StarNet, LSKNet, Mambaout, and UMFormer as comparison objects, using the ResNet backbone in RT-DETR as the baseline, as shown in Table 4 and Figure 12, for a comprehensive comparison of each backbone scheme across detection accuracy and computational efficiency dimensions.

**Table 4.** Comparative experimental results of different backbone networks versus DRHANet.

| Model | GFLOPs | Param/M | P | R | $mAP_{50}$ | $mAP_{50\text{-}95}$ |
|---|---|---|---|---|---|---|
| ResNet (base) | 57.3 | 19.85 | 85.3 | 70.2 | 74.8 | 41.5 |
| StarNet [43] | 49.0 | 27.60 | 83.5 | 66.5 | 74.1 | 41.9 |
| LSKNet [44] | 54.6 | 31.51 | 85.6 | 71.3 | 75.7 | 41.2 |
| Mambaout [45] | 50.1 | 17.35 | 86.0 | 70.8 | 75.2 | 42.0 |
| UMFormer [46] | 63.5 | 32.74 | 84.5 | 69.1 | 74.4 | 41.7 |
| DRHANet (ours) | 46.8 | 14.52 | 86.7 | 72.0 | 75.8 | 42.3 |

 doi:10.20944/preprints202604.1731.v1



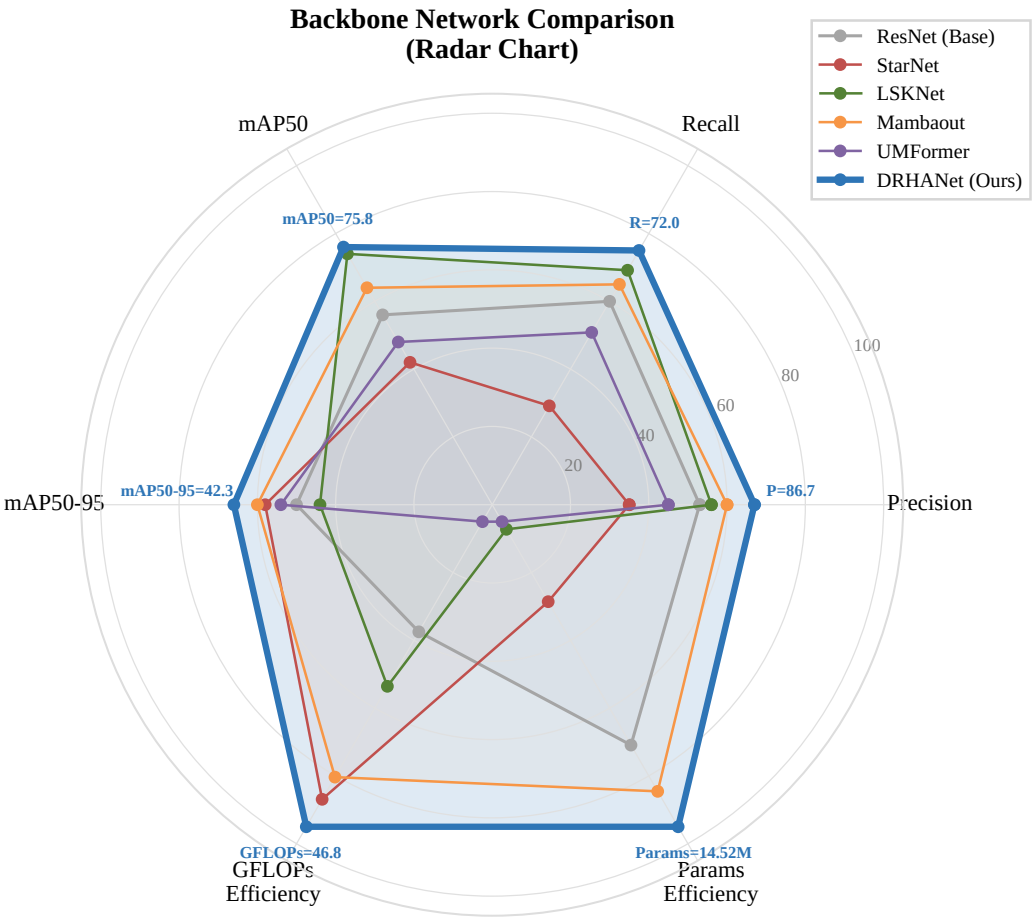


**Figure 12.** Comparative experimental results of different backbone networks versus DRHANet.

The comparison results show that all competing backbone networks exhibit obvious limitations in the overall trade-off between accuracy and efficiency. StarNet has a parameter count as high as 27.6M, far exceeding the baseline, and its $mAP_{50}$ of only 74.1% actually falls 0.7% below the baseline; LSKNet improves $mAP_{50}$ by 0.9% over the baseline but its parameter count expands to 31.51M, incurring a high computational cost; Mambaout's $mAP_{50}$ is comparable to the baseline with limited efficiency advantages; UMFormer's GFLOPs reach 63.5 and its $mAP_{50}$ is only 74.4%, with both performance and efficiency unsatisfactory. By contrast, our proposed DRHANet reduces GFLOPs and parameter count by 18.3% and 26.9% relative to the baseline, respectively, while improving $mAP_{50}$ and $mAP_{50\text{-}95}$ by 1.0% and 0.8%, simultaneously achieving the highest detection accuracy and optimal computational efficiency among all comparison methods. This result fully validates that DRHANet, through dynamic anisotropic receptive field learning and hyper-graph high-order semantic topology modeling, can effectively enhance feature extraction capability for special targets under multi-type degradation conditions while reducing redundant computation, placing us at the forefront in both accuracy and lightweighting.

### 4.5.2. Neck Network Comparison

The design quality of the neck network directly affects the feature fusion effectiveness at multiple scales and the final detection accuracy. Under fixed backbone and decoder configurations, we systematically compared BWAFN against several advanced neck network methods—RFPN, HS-FPN, FreqFFPN, and HFFE—as shown in Table 5.

 doi:10.20944/preprints202604.1731.v1



**Table 5.** Comparative experimental results of different neck networks versus BWAFN.

| Model | GFLOPs | Params/M | P | R | $mAP_{50}$ | $mAP_{50-95}$ |
|---|---|---|---|---|---|---|
| CCFF (base) | 57.3 | 19.85 | 85.3 | 70.2 | 74.8 | 41.5 |
| RFPN [47] | 42.1 | 31.29 | 85.1 | 69.8 | 74.9 | 41.0 |
| HS-FPN [48] | 39.8 | 27.93 | 84.3 | 70.7 | 75.8 | 41.7 |
| FreqFFPN [49] | 65.9 | 17.35 | 84.7 | 70.4 | 75.5 | 41.9 |
| HFFE [50] | 62.6 | 19.32 | 86.2 | 71.1 | 74.1 | 40.8 |
| BWAFN (ours) | 48.6 | 17.11 | 85.9 | 71.5 | 76.5 | 42.0 |

Examining the experimental data of each group reveals that competing neck networks all struggle to simultaneously achieve clear advantages in both accuracy and efficiency. RFPN reduces GFLOPs to 42.1, but its $mAP_{50}$ of only 74.9% is comparable to the baseline and fails to bring effective performance improvement; HS-FPN ties with BWAFN on $mAP_{50}$, but its parameter count of 27.93M far exceeds BWAFN's 17.11M; FreqFFPN improves $mAP_{50}$ by 0.7% over the baseline, but carries a heavier computational burden; HFFE's $mAP_{50}$ and $mAP_{50-95}$ are both below the baseline, exhibiting inferior performance. Our proposed BWAFN achieves the highest $mAP_{50}$ among all comparison methods with 48.6 GFLOPs and 17.11M parameters, improving by 1.7% over the baseline with a 0.5% improvement in $mAP_{50-95}$. This result validates that BWAFN, through the cooperative design of its bidirectional adaptive weighted fusion path, CMRFT multi-receptive-field node transforms, and LSCU lightweight spatial compensation upsampling, can effectively improve the semantic consistency and boundary localization accuracy of neck multi-scale features while maintaining a relatively low computational cost, making us comprehensively superior to existing methods in the combined performance of accuracy and efficiency.

#### 4.5.3. Overall Comparison

To evaluate the comprehensive competitiveness of RHG-DETR from a broader perspective, we conducted an extensive lateral comparison against multiple representative methods among the current mainstream DETR-series detectors, two-stage detectors, and single-stage detectors. As shown in Table 6 and Figure 13, the comparison methods include DETR-series frameworks such as RT-DETR, Swin-DETR, Deformable-DETR, DEIM, and D-Fine, as well as representative detectors including Faster R-CNN, Cascade R-CNN, YOLOv8-M/v10-M/v11-M/v12-M, FBRT-YOLO-M, and GOLD-YOLO-M, with comprehensive evaluation across detection accuracy, computation, parameter count, and inference speed dimensions.

 doi:10.20944/preprints202604.1731.v1



**Table 6.** Overall comparative experimental results of RHG-DETR versus mainstream state-of-the-art methods.

| Model | GFLOPs | Params/M | P | R | $mAP_{50}$ | $mAP_{50-95}$ | FPS |
|---|---|---|---|---|---|---|---|
| *DETR-Based Object Detector* | | | | | | | |
| RT-DETR-R50 | 128.9 | 41.88 | 78.1 | 73.8 | 76.9 | 46.1 | 64.8 |
| RT-DETR-R101 | 247.4 | 74.23 | 84.7 | 71.5 | 77.4 | 46.8 | 45.3 |
| Swin-DETR [51] | 102.6 | 43.71 | 77.0 | 70.5 | 75.6 | 45.7 | 24.8 |
| Deformable-DETR [52] | 173.5 | 42.02 | 81.2 | 69.7 | 73.8 | 44.1 | 21.4 |
| DEIM-M [53] | 271.3 | 47.32 | 83.6 | 71.3 | 76.4 | 45.8 | 23.0 |
| D-Fine-M [54] | 56.37 | 19.19 | 82.0 | 70.1 | 77.2 | 46.5 | 39.9 |
| *Two-Stage Object Detector* | | | | | | | |
| Faster R-CNN [55] | 180.0 | 42.02 | 74.3 | 66.0 | 72.9 | 40.1 | 35.7 |
| Cascade R-CNN [56] | 229.61 | 66.32 | 70.7 | 67.3 | 70.3 | 39.0 | 30.5 |
| *One-Stage Object Detector* | | | | | | | |
| YOLOv8-M [57] | 64.4 | 25.0 | 84.6 | 72.3 | 74.7 | 44.0 | 106.4 |
| YOLOv10-M [7] | 58.9 | 15.32 | 85.0 | 69.8 | 75.2 | 43.3 | 107.8 |
| YOLOv11-M [8] | 67.7 | 20.04 | 87.0 | 70.7 | 76.8 | 45.1 | 113.8 |
| YOLOv12-M [9] | 67.1 | 20.11 | 86.7 | 71.2 | 76.5 | 43.6 | 114.6 |
| FBRT-YOLO-M [58] | 58.7 | 7.36 | 86.2 | 70.9 | 77.3 | 45.9 | 83.2 |
| GOLD-YOLO-M [59] | 87.6 | 41.28 | 87.3 | 70.3 | 75.9 | 42.7 | 69.3 |
| RT-DETR-R18 (base) | 57.3 | 19.88 | 85.3 | 70.2 | 74.8 | 41.5 | 79.6 |
| **RHG-DETR (ours)** | **37.6** | **14.16** | **87.6** | **74.4** | **78.5** | **44.6** | **84.2** |

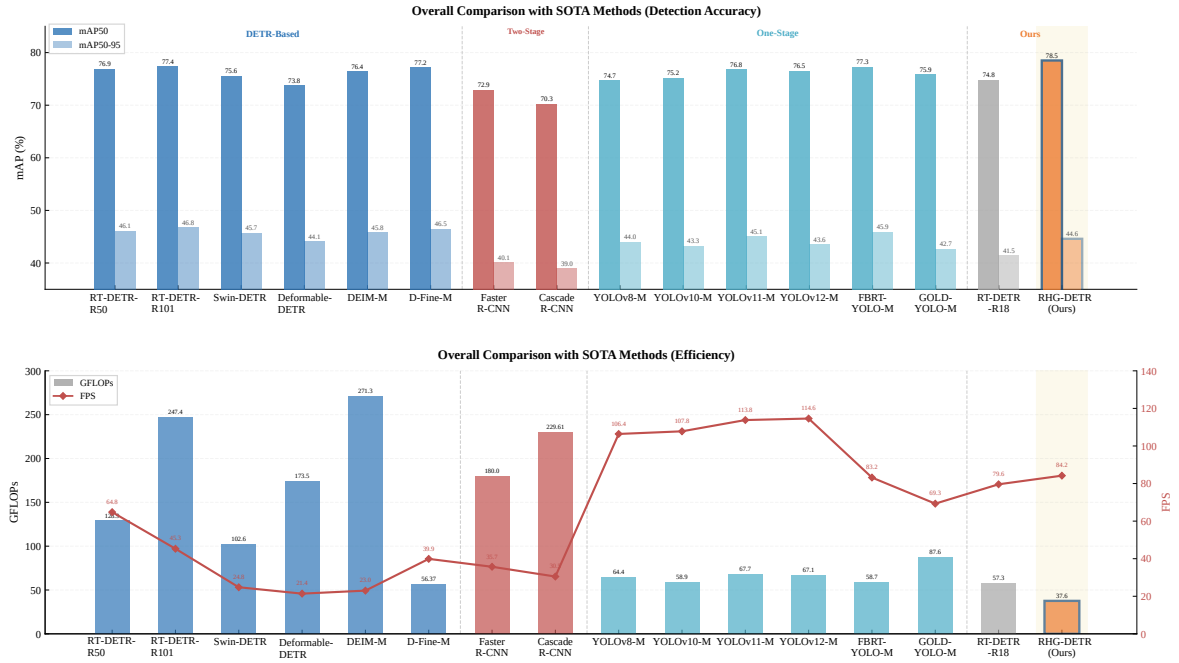


**Figure 13.** Overall comparative experimental results of RHG-DETR versus mainstream state-of-the-art methods.

From the quantitative results, existing DETR-series methods generally face the problem of high computational cost and slow inference speed: RT-DETR-R101 achieves $mAP_{50}$ of 77.4% but at the cost of GFLOPs as high as 247.4 and FPS of only 45.3; Swin-DETR and Deformable-DETR both have $mAP_{50}$ not exceeding 75.6% with similarly insufficient inference speed. Among two-stage detectors, Faster R-CNN and Cascade R-CNN achieve $mAP_{50}$ of 72.9% and 70.3% respectively, lagging behind in both accuracy and speed. YOLO-series methods hold certain advantages in inference speed, with YOLOv11-M and FBRT-YOLO-M reaching $mAP_{50}$ of 76.8% and 77.3% respectively, but still exhibit limitations in the comprehensive trade-off between accuracy and lightweighting. Our proposed RHG-DETR achieves the highest $mAP_{50}$ and $mAP_{50-95}$ among all comparison methods, with precision and recall improving by 2.3% and 4.2% respectively over the baseline RT-DETR-R18, $mAP_{50}$ and

 doi:10.20944/preprints202604.1731.v1



$mAP_{50-95}$ improving by 3.7% and 3.1% respectively, while GFLOPs and parameter count are reduced by 34.4% and 28.8% relative to the baseline, and inference speed is boosted to 84.2 FPS. This result fully demonstrates that RHG-DETR, through the cooperative design of its three core modules, achieves comprehensive leadership in detection accuracy, computational efficiency, and inference speed in the special target detection task under complex degradation conditions.

### *4.6. Cross-Dataset Evaluation*

A robust detection model should not only excel within a specific training domain; its transfer capability across scenes is equally critical. To this end, we selected VisDrone2019 and BDD100K as two public datasets to compare RHG-DETR against the baseline RT-DETR-R18 for validation. As shown in Table 7 and Figure 14, a comprehensive evaluation of precision, recall, $mAP_{50}$, and $mAP_{50-95}$ was conducted for both methods on the two datasets.

**Table 7.** Generalization comparison results of RHG-DETR on the VisDrone2019 and BDD100K datasets.

| Dataset | Model | P | R | $mAP_{50}$ | $mAP_{50-95}$ |
|---|---|---|---|---|---|
| VisDrone2019 | RT-DETR-R18 | 63.4 | 46.8 | 48.4 | 30.0 |
| | RHG-DETR | 64.1 | 48.6 | 50.2 | 31.1 |
| BDD100K | RT-DETR-R18 | 63.3 | 48.1 | 48.8 | 31.9 |
| | RHG-DETR | 64.7 | 49.3 | 50.5 | 33.0 |

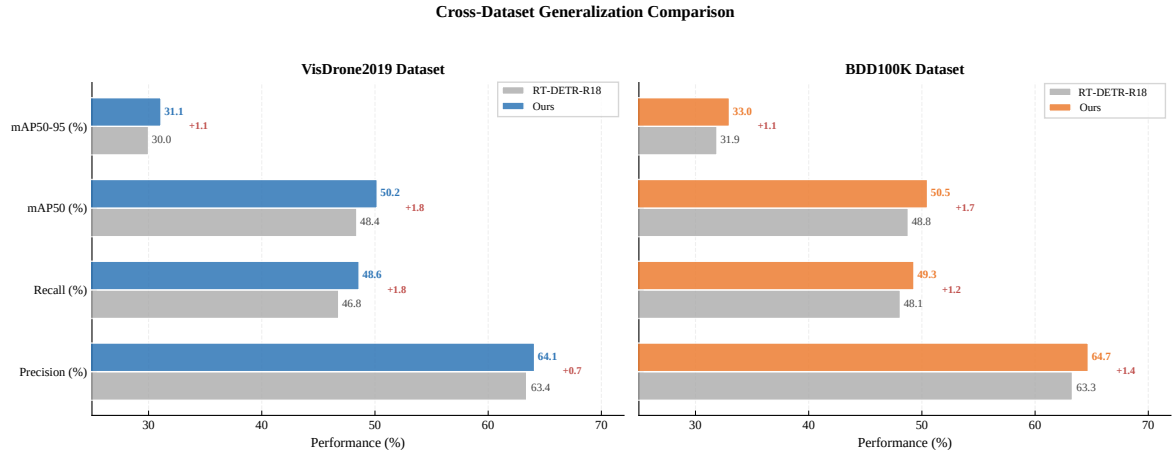


**Figure 14.** Cross-dataset generalization comparison results of RHG-DETR on VisDrone2019 and BDD100K.

The experimental results on both datasets consistently demonstrate that RHG-DETR achieves performance improvements over the baseline, exhibiting good cross-scene generalization capability. On the VisDrone dataset, RHG-DETR's $mAP_{50}$ improves by 1.8% over RT-DETR-R18, $mAP_{50-95}$ improves by 1.1%, and precision and recall improve by 0.7% and 1.8%, respectively; on BDD100K, $mAP_{50}$ likewise improves by 1.7% and $mAP_{50-95}$ by 1.1%, with all metrics exhibiting consistent positive improvement. This stable improvement across datasets fully demonstrates that the dynamic anisotropic receptive field extraction, bidirectional adaptive feature fusion, and sparse encoding mechanism proposed in RHG-DETR possess strong task-agnostic properties, capable of effectively transferring to object detection tasks in different scenes rather than merely overfitting to the specific training domain, further validating the practical advantage of our method in improving detection generalizability.

### *4.7. Visualization Results*

#### 4.7.1. Special Target Detection Visualization

To validate the effectiveness of the proposed RHG-DETR framework in special target detection tasks under complex degradation environments, this section conducts systematic visual comparison experiments across multiple typical challenging scenarios. Through intuitive detection result compari-

 doi:10.20944/preprints202604.1731.v1



son against current mainstream detection methods—YOLOv11m, GOLD-YOLO-M, and RT-DETR-R18 (RT-DETR)—a comprehensive evaluation of RHG-DETR's detection performance across different target categories and degradation types is conducted from three dimensions: detection bounding box localization accuracy, missed detection rate, and false detection rate, as shown in Figure 15.

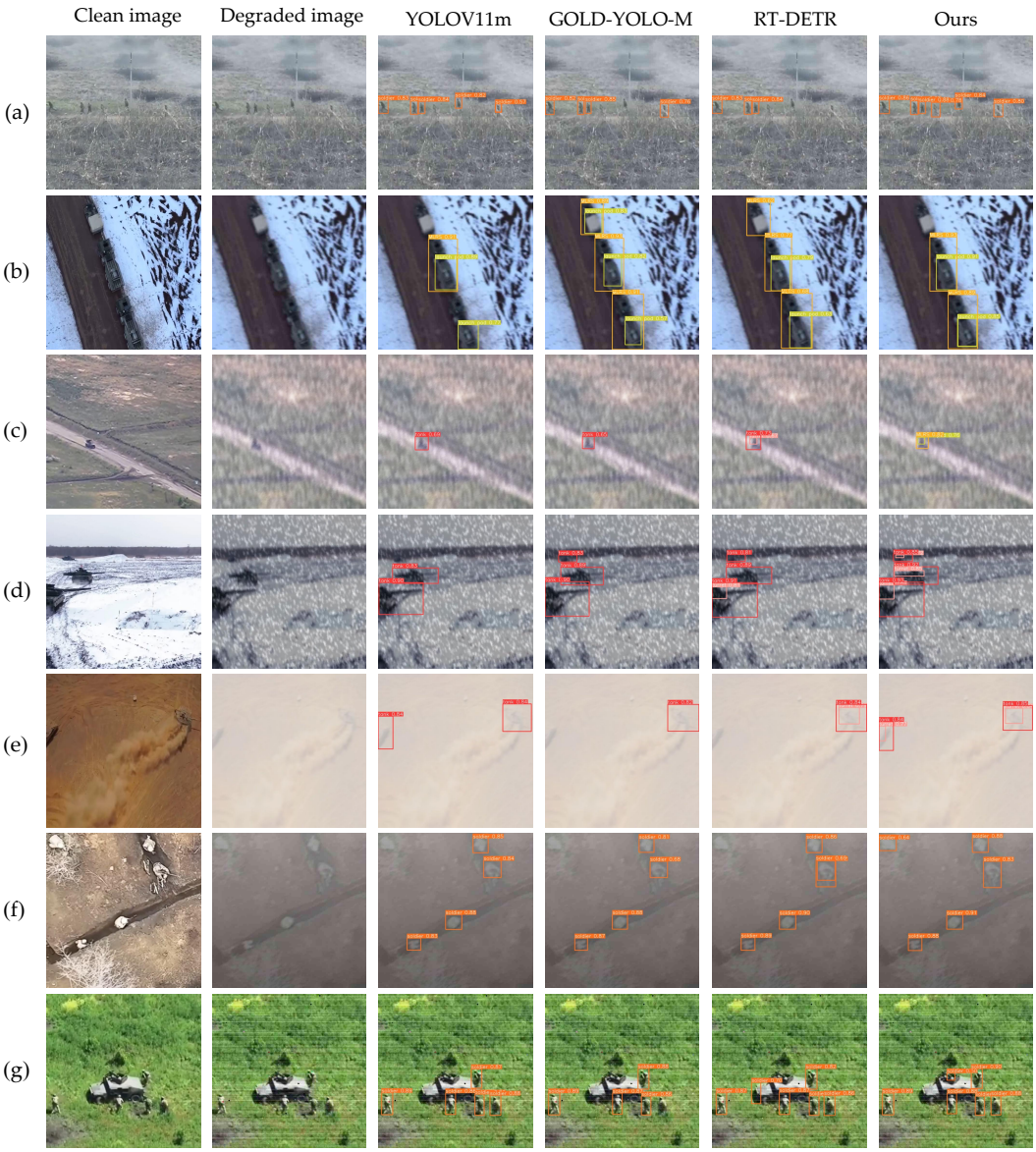


**Figure 15.** Qualitative detection result comparison on the special target dataset under multi-type degradation conditions.

As shown in Figure 15, RHG-DETR exhibits outstanding detection performance across multiple degradation scenarios, significantly outperforming the comparison methods in bounding box localization accuracy, missed detection rate, and false detection rate. In all six typical degradation scenarios—mild blur smoke-occluded tanks (row a), moderate blur woodland-camouflaged soldiers (row b), moderate blur combined with rain outdoor-camouflaged tanks (row c), mild blur combined with strong light camouflaged soldiers (row d), moderate blur combined with electromagnetic interference rocket systems (row e), and moderate blur combined with low illumination rocket systems (row f)—the baseline model consistently exhibits the inherent deficiencies of widely distributed attention, significant background interference, and insufficient target region activation. As DRHANet, BWAFN, and ASMED are progressively introduced, the heat maps in each scenario show systematic improvement trends: RHGF models high-order semantic topology dependencies among components,

 doi:10.20944/preprints202604.1731.v1



strengthening target main body responses; CMRFT nodes maintain cross-scale semantic consistency under atmospheric degradation; SWSA confines feature interactions to high-saliency regions, effectively suppressing background noise; SGFF injects spatial geometric topology constraints into the feedforward stage, preserving sub-component structural co-occurrence relationships; and DCNM with MGCA cooperatively stabilize the non-stationary activation distributions under extreme imaging conditions.

The complete model (+A+B+C) achieves the highest target region activation intensity, the most precise spatial localization, and the lowest background false activation across all six degradation scenarios, demonstrating clear advantages over each ablation stage. Overall, as the three modules are progressively stacked, RHG-DETR's attention heat maps consistently exhibit a systematic evolution from scattered interference to precise focusing, fully validating the cooperative effectiveness of each module in enhancing special target attention perception under multi-type composite degradation and their practical deployment value.

Through analysis of the failure cases in Figure 16, RHG-DETR still exhibits certain advantages over YOLOv11m, GOLD-YOLO-M, and RT-DETR in five extreme degradation scenarios: severely blurred grassland tanks (row a), severe blur combined with moderate fog trench-camouflaged soldiers (row b), moderate blur combined with heavy rain mountainous tanks (row c), moderate blur combined with severe strong-light desert rocket systems (row d), and mild blur combined with severe low-illumination plateau rocket systems (row e). Specifically, in row a, component-level recognition of one tank is achieved; in row b, two soldier instances are detected with superior recall; in row c, two tanks are correctly identified along with the turret structure of one; in row d, two rocket system targets are successfully identified with reasonable localization; and in row e, four rocket system instances are detected under extremely weak illumination, with relatively stronger weak-light perception capability.

Nevertheless, RHG-DETR still exposes clear limitations in these scenarios: sub-component recognition and localization accuracy for turrets and similar parts under heavy blur conditions requires improvement; target recall in composite occlusion scenarios still has room for enhancement; sub-component recognition capability under severe strong-light saturation and extreme low illumination decreases noticeably; and fine localization errors remain under heavy rain combined with blur degradation. Overall, detection robustness under extreme imaging conditions, completeness of sub-component recognition, and localization accuracy are important directions for improvement in future research.

 doi:10.20944/preprints202604.1731.v1



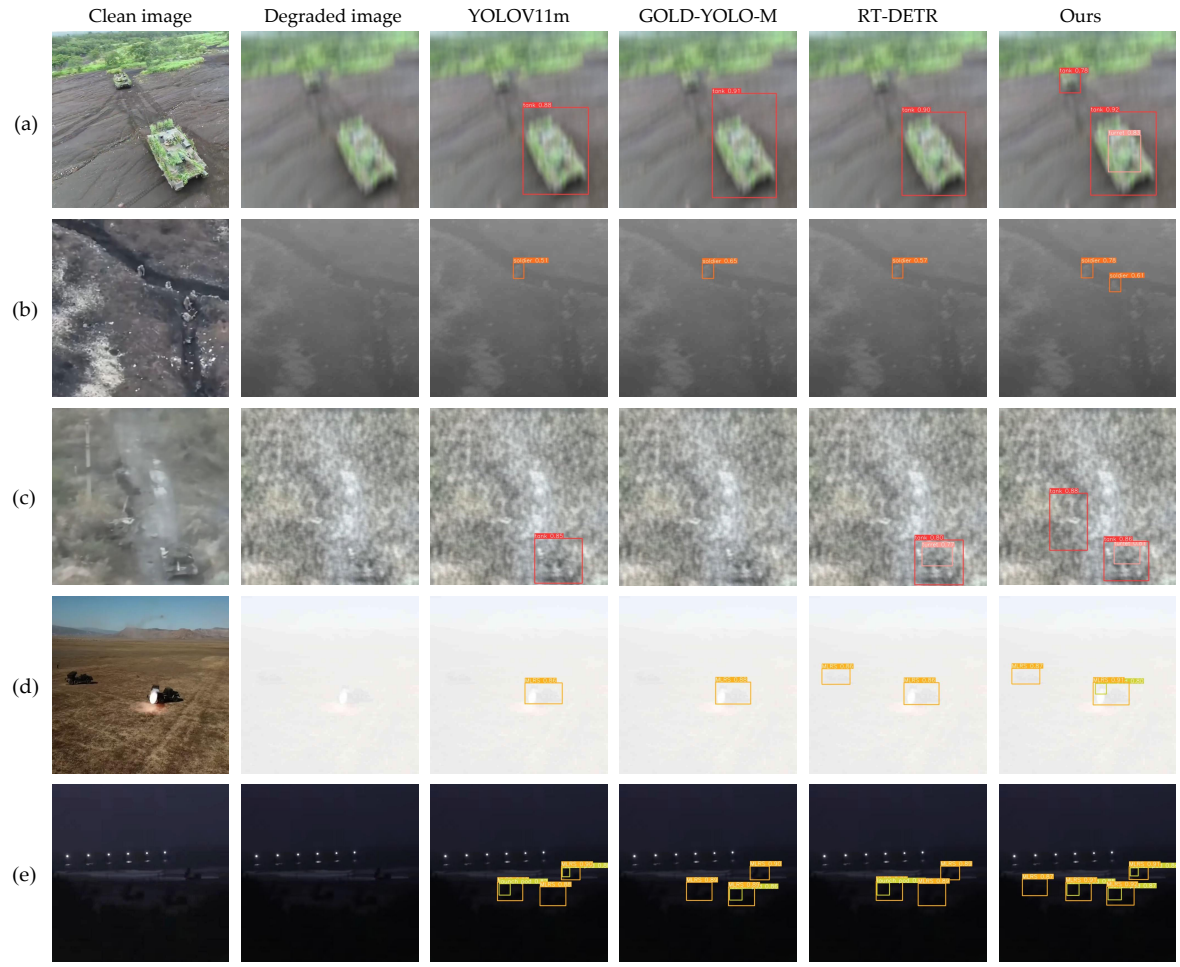


**Figure 16.** Visualization of detection failure cases.

4.7.2. VisDrone2019 Dataset Visualization

To validate the effectiveness and robustness of the proposed RHG-DETR framework in UAV aerial small target detection, this paper conducts systematic visual comparison experiments on the VisDrone2019 UAV aerial dataset. The experiments select representative UAV aerial application scenarios including urban road vehicle detection from high-altitude top-down viewpoints, multi-motion-blur target distribution detection in complex nighttime outdoor traffic environments, and occluded target detection under nighttime low-illumination conditions. Through intuitive detection result comparison against baseline model RT-DETR-R18 (RT-DETR), a comprehensive evaluation of RHG-DETR's detection performance is conducted from three dimensions: multi-scale target detection, blurred target detection, and detection robustness under low-illumination conditions, as shown in Figure 17.

 doi:10.20944/preprints202604.1731.v1




**Figure 17.** Visualization comparison of RT-DETR versus RHG-DETR on the VisDrone2019 dataset.

As shown in Figure 17, RHG-DETR demonstrates significant performance advantages over the baseline method RT-DETR across multiple typical UAV aerial scenarios in the VisDrone2019 dataset. In the high-altitude top-down urban road vehicle detection scenario (first column), RHG-DETR accurately detects all vehicle targets with precise localization, while RT-DETR exhibits obvious missed detections, especially insufficient perception of small-scale targets in distant views; this demonstrates that the anisotropic dynamic receptive field mechanism of DRHANet effectively enhances the backbone network's extraction capability for directional features of small targets, and the bidirectional adaptive weighted fusion path of BWAFN ensures the integrity of shallow spatial details during cross-scale propagation. In the complex traffic scenario with superimposed low illumination and motion blur (second column), RHG-DETR leverages the adaptive perception of anisotropic dynamic receptive fields for directional features of motion-blurred targets, achieving more complete and precise detection coverage; RT-DETR, due to the dual mismatch of fixed receptive fields and the linear normalization assumption, exhibits obvious missed detections and false detections. In the nighttime weak-light occluded target detection scenario (third column), RHG-DETR relies on the effective compression of DCNM for high dynamic range activations and the active suppression of SWSA for background redundant activations, accurately localizing multiple targets despite low-illumination interference; RT-DETR, due to the statistical mismatch between the linear normalization assumption and the non-stationary activation distribution, exhibits significantly reduced target-background activation separability with evidently insufficient robustness.

#### 4.7.3. BDD100K Dataset Visualization

To validate the generalization capability and detection performance of the proposed RHG-DETR framework in heterogeneous traffic scenarios, this paper conducts systematic visual comparison experiments on the BDD100K autonomous driving dataset. The experiments select typical scenarios including complex urban intersection night-vision blurred targets, multi-lane multi-scale targets on highways under low-illumination conditions, and multi-target blur and occlusion, and through systematic comparison against the baseline model RT-DETR, a comprehensive evaluation of RHG-DETR's accuracy and robustness in detecting safety-critical elements such as vehicles, traffic signs, and pedestrians is conducted from both quantitative and qualitative perspectives, as shown in Figure 18.

 doi:10.20944/preprints202604.1731.v1




**Figure 18.** Visualization comparison of RT-DETR versus RHG-DETR on the BDD100K dataset.

As shown in Figure 18, RHG-DETR demonstrates detection performance superior to the baseline model RT-DETR across multiple typical complex traffic scenarios in the BDD100K autonomous driving dataset. In the nighttime urban intersection scenario (first column), RHG-DETR leverages the precise feature focusing of SWSA on high-semantically-responsive regions and the explicit constraint of SGFF on target geometric topological relationships, achieving accurate and complete detection of traffic lights, pedestrians, and vehicles; RT-DETR, due to the indiscriminate aggregation of global attention over background noise, exhibits traffic light missed detections and pedestrian localization offsets. In the low-illumination highway multi-lane scenario (second column), RHG-DETR leverages the adaptive multi-directional receptive field modeling of ADWC and the dynamic maintenance of cross-scale semantic consistency by the BWAFN bidirectional weighted fusion path, significantly outperforming RT-DETR in detection completeness and confidence for distant small targets, effectively suppressing missed detections caused by weak activation signals being diluted by background noise. In the occluded and blurred target scenario (third column), RHG-DETR, through the bounded compression normalization of DCNM for high dynamic range activations and the dynamic coupling control of MGCA at the feature flow level, effectively maintains the separability of target and background activations under blur conditions, outperforming RT-DETR in both detection accuracy and bounding box localization precision. The above visualization results comprehensively validate the technical advantages of RHG-DETR in cross-domain generalization applications from a qualitative perspective across multiple scenes and degradation types, demonstrating that the three core innovations DRHANet, BWAFN, and ASMED possess good cooperative adaptability in out-of-domain complex traffic perception tasks.

## 5. Conclusions

This paper proposes RHG-DETR, a robust detection framework for special target recognition in degraded UAV remote sensing imagery under multi-type degradation conditions. Three synergistic modules are introduced at the backbone, neck, and encoder levels. DRHANet enhances adaptive multi-scale feature extraction and high-order semantic topology modeling of target components through anisotropic dynamic depthwise separable convolution and the Riemannian hyper-graph fusion mechanism, effectively addressing the systematic feature representation degradation caused by fixed receptive fields under blur and low-contrast backgrounds. BWAFN constructs a bidirectional adaptive weighted feature pyramid combined with lightweight spatial compensation upsampling, significantly improving cross-scale semantic consistency under composite degradation conditions. The ASMED encoder suppresses background interference and preserves geometric topology semantics of target

 doi:10.20944/preprints202604.1731.v1



sub-components through the synergistic design of sparse window self-attention and dynamic channel-wise nonlinear modulation. On the self-constructed special target dataset, RHG-DETR achieves an $mAP_{50}$ of 78.5%, outperforming the RT-DETR baseline and other mainstream methods, while reducing computational cost and parameter count by 34.4% and 28.8%, respectively, at an inference speed of 84.2 FPS. Consistent performance improvements on the VisDrone2019 and BDD100K datasets further confirm the framework's strong cross-domain generalization ability. However, the current framework still exhibits limitations under extreme degradation conditions such as heavy blur, severe strong-light saturation, and extreme low illumination, particularly in sub-component recognition and fine-grained localization accuracy. Future work will focus on adaptive preprocessing mechanisms for extreme imaging conditions and joint multi-degradation modeling strategies to further enhance detection robustness in extreme environments.

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

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

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

## References

1. Cheng, B.; Deng, Q.; Gan, W.; Li, Y.; Wu, C.; Wang, Q. A Review of Remote Sensing Image Object Detection: Advances and Challenges in Data, Techniques, and Applications. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 10438–10472. https://doi.org/10.1109/JSTARS.2026.3671199
2. Bartlett, B.; Santos, M.; Dorian, T.; Moreno, M.; Trslic, P.; Dooly, G. Real-Time UAV Surveys with the Modular Detection and Targeting System: Balancing Wide-Area Coverage and High-Resolution Precision in Wildlife Monitoring. *Remote Sens.* **2025**, *17*, 879. https://doi.org/10.3390/rs17050879
3. Kaur, R.; Karmakar, G.; Xia, F. et al. Deep learning: survey of environmental and camera impacts on internet of things images. *Artif. Intell. Rev.* **2023**, *56*, 9605–9638. https://doi.org/10.1007/s10462-023-10405-7
4. Munir, A.; Siddiqui, A.J.; Anwar, S.; El-Maleh, A.; Khan, A.H.; Rehman, A. Impact of Adverse Weather and Image Distortions on Vision-Based UAV Detection: A Performance Evaluation of Deep Learning Models. *Drones* **2024**, *8*, 638. https://doi.org/10.3390/drones8110638
5. Liu, J.; Zhan, J.; Guo, Y.; Li, T.; Zhao, Y.; Zhang, J.; Tang, L.; Li, Y.; Wei, Y.; Cai, W. Learning Geometric-Aware and Weather-Adaptive Semantics in Remote Sensing: Affine Lie Group Enhanced Detector for UAV Road Scenes. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 1–20. https://doi.org/10.1109/TGRS.2026.3655446
6. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
7. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 107984–108011.
8. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725.
9. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524.
10. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Computer Vision – ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.

 doi:10.20944/preprints202604.1731.v1




11. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 17–21 June 2024; pp. 16965–16974.
12. Ding, L.; Huang, H.; Zang, Y. Image Quality Assessment Using Directional Anisotropy Structure Measurement. *IEEE Trans. Image Process.* **2017**, *26*, 1799–1809. https://doi.org/10.1109/TIP.2017.2665972
13. Kadha, V.; Bakshi, S.; Das, S.K. Unravelling Digital Forgeries: A Systematic Survey on Image Manipulation Detection and Localization. *ACM Comput. Surv.* **2025**, *57*, 323. https://doi.org/10.1145/3731243
14. Wang, D.; Yan, Z.; Liu, P. Fine-Grained Interpretation of Remote Sensing Image: A Review. *Remote Sens.* **2025**, *17*, 3887. https://doi.org/10.3390/rs17233887
15. Shang, R.; Zhang, J.; Jiao, L.; Li, Y.; Marturi, N.; Stolkin, R. Multi-scale Adaptive Feature Fusion Network for Semantic Segmentation in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 872. https://doi.org/10.3390/rs12050872
16. Aleissaee, A.A.; Kumar, A.; Anwer, R.M.; Khan, S.; Cholakkal, H.; Xia, G.-S.; Khan, F.S. Transformers in Remote Sensing: A Survey. *Remote Sens.* **2023**, *15*, 1860. https://doi.org/10.3390/rs15071860
17. Yu, W.; Zhang, J.; Liu, D.; Xi, Y.; Wu, Y. An Effective and Lightweight Full-Scale Target Detection Network for UAV Images Based on Deformable Convolutions and Multi-Scale Contextual Feature Optimization. *Remote Sens.* **2024**, *16*, 2944. https://doi.org/10.3390/rs16162944
18. Li, J.; Shi, Y.; Hong, Q.; Jia, Y. A Scale-Aware Multidomain DETR for Small Object Detection in UAV Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–20. https://doi.org/10.1109/TGRS.2025.3624765
19. Pei, Y.; Huang, Y.; Zou, Q.; Zhang, X.; Wang, S. Effects of Image Degradation and Degradation Removal to CNN-Based Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1239–1253. https://doi.org/10.1109/TPAMI.2019.2950923
20. Zhao, Z.; Xiong, B.; Wang, L.; Ou, Q.; Yu, L.; Kuang, F. RetinexDIP: A Unified Deep Framework for Low-Light Image Enhancement. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1076–1088. https://doi.org/10.1109/TCSVT.2021.3073371
21. Ma, J.; Lin, M.; Zhou, G.; Jia, Z. Joint Image Restoration For Domain Adaptive Object Detection In Foggy Weather Condition. In *Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, UAE, 27–30 October 2024; pp. 542–548. https://doi.org/10.1109/ICIP51287.2024.10647560
22. Kang, L.-W.; Chou, K.-L.; Fu, R.-H. Deep Learning-Based Weather Image Recognition. In *Proceedings of the 2018 International Symposium on Computer, Consumer and Control (IS3C)*, Taichung, Taiwan, 6–8 December 2018; pp. 384–387. https://doi.org/10.1109/IS3C.2018.00103
23. Gupta, H.; Kotlyar, O.; Andreasson, H.; Lilienthal, A.J. Robust Object Detection in Challenging Weather Conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 3–8 January 2024; pp. 7523–7532.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016.
25. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q. et al. VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Seoul, Republic of Korea, 27 October–2 November 2019.
26. Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J. et al. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 18–22 June 2018.
27. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. https://doi.org/10.1016/j.isprsjprs.2019.11.023
28. Peng, Y.; Tang, Z.; Zhao, G.; Cao, G.; Wu, C. Motion Blur Removal for Uav-Based Wind Turbine Blade Images Using Synthetic Datasets. *Remote Sens.* **2022**, *14*, 87. https://doi.org/10.3390/rs14010087
29. Nah, S.; Kim, T.H.; Lee, K.M. Deep Multi-Scale Convolutional Neural Network for Dynamic Scene Deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017.
30. Garg, K.; Nayar, S.K. Photorealistic rendering of rain streaks. *ACM Trans. Graph.* **2006**, *25*, 996–1002. https://doi.org/10.1145/1141911.1141985
31. Liu, Y.-F.; Jaw, D.-W.; Huang, S.-C.; Hwang, J.-N. DesnowNet: Context-Aware Deep Network for Snow Removal. *IEEE Trans. Image Process.* **2018**, *27*, 3064–3073. https://doi.org/10.1109/TIP.2018.2806202

 doi:10.20944/preprints202604.1731.v1




32. Narasimhan, S.G.; Nayar, S.K. Vision and the Atmosphere. *Int. J. Comput. Vis.* **2002**, *48*, 233–254. https://doi.org/10.1023/A:1016328200723
33. Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; Wang, Z. Benchmarking Single-Image Dehazing and Beyond. *IEEE Trans. Image Process.* **2019**, *28*, 492–505. https://doi.org/10.1109/TIP.2018.2867951
34. Kim, S.-G.; Lee, E.; Hong, I.-P.; Yook, J.-G. Review of Intentional Electromagnetic Interference on UAV Sensor Modules and Experimental Study. *Sensors* **2022**, *22*, 2384. https://doi.org/10.3390/s22062384
35. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*, 4th ed.; Pearson: Hoboken, NJ, USA, 2018.
36. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep Retinex Decomposition for Low-Light Enhancement. *arXiv* **2018**, arXiv:1808.04560.
37. Wei, K.; Fu, Y.; Zheng, Y.; Yang, J. Physics-Based Noise Modeling for Extreme Low-Light Photography. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 8520–8537. https://doi.org/10.1109/TPAMI.2021.3103114
38. Liu, J.; Xu, D.; Yang, W.; Fan, M.; Huang, H. Benchmarking Low-Light Image Enhancement and Beyond. *Int. J. Comput. Vis.* **2021**, *129*, 1153–1184. https://doi.org/10.1007/s11263-020-01418-8
39. Afifi, M.; Derpanis, K.G.; Ommer, B.; Brown, M.S. Learning Multi-Scale Photo Exposure Correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 20–25 June 2021; pp. 9157–9167.
40. Hendrycks, D.; Dietterich, T.G. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *arXiv* **2019**, arXiv:1903.12261.
41. Michaelis, C.; Mitzkus, B.; Geirhos, R.; Rusak, E.; Bringmann, O.; Ecker, A.S.; Bethge, M.; Brendel, W. Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. *arXiv* **2019**, arXiv:1907.07484.
42. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 13–19 June 2020; pp. 2636–2645. https://doi.org/10.1109/CVPR42600.2020.00271
43. Ma, X.; Dai, X.; Bai, Y.; Wang, Y.; Fu, Y. Rewrite the Stars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 17–21 June 2024; pp. 5694–5703.
44. Li, Y.; Li, X.; Dai, Y.; Hou, Q.; Liu, L.; Liu, Y.; Cheng, M.-M.; Yang, J. LSKNet: A Foundation Lightweight Backbone for Remote Sensing. *Int. J. Comput. Vis.* **2025**, *133*, 1410–1431. https://doi.org/10.1007/s11263-024-02247-9
45. Yu, W.; Wang, X. MambaOut: Do We Really Need Mamba for Vision? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 11–15 June 2025; pp. 4484–4496.
46. Li, L.; Yi, J.; Fan, H.; Lin, H. A Lightweight Semantic Segmentation Network Based on Self-Attention Mechanism and State Space Model for Efficient Urban Scene Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. https://doi.org/10.1109/TGRS.2025.3562185
47. Li, H. Rethinking Features-Fused-Pyramid-Neck for Object Detection. In *Computer Vision – ECCV 2024*; Leonardis, A. et al., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2025; pp. 74–90.
48. Chen, Y.; Zhang, C.; Chen, B.; Huang, Y.; Sun, Y.; Wang, C.; Fu, X.; Dai, Y.; Qin, F.; Peng, Y.; et al. Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases. *Comput. Biol. Med.* **2024**, *170*, 107917. https://doi.org/10.1016/j.compbiomed.2024.107917
49. Chen, L.; Fu, Y.; Gu, L.; Yan, C.; Harada, T.; Huang, G. Frequency-Aware Feature Fusion for Dense Image Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10763–10780. https://doi.org/10.1109/TPAMI.2024.3449959
50. Zhang, Y.; Bao, W.; Yang, Y.; Wan, W.; Xiao, Q.; Zou, X. HAFNet: Hierarchical Attention Fusion Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–16. https://doi.org/10.1109/TGRS.2025.3607732
51. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
52. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2021**, arXiv:2010.04159.
53. Huang, S.; Lu, Z.; Cun, X.; Yu, Y.; Zhou, X.; Shen, X. DEIM: DETR with Improved Matching for Fast Convergence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 11–15 June 2025; pp. 15162–15171.

 doi:10.20944/preprints202604.1731.v1




54. Peng, Y.; Li, H.; Wu, P.; Zhang, Y.; Sun, X.; Wu, F. D-FINE: Redefine Regression Task in DETRs as Fine-grained Distribution Refinement. *arXiv* **2024**, arXiv:2410.13842.
55. Xu, J.; Ren, H.; Cai, S.; Zhang, X. An improved faster R-CNN algorithm for assisted detection of lung nodules. *Comput. Biol. Med.* **2023**, *153*, 106470. https://doi.org/10.1016/j.compbiomed.2022.106470
56. Chai, B.; Nie, X.; Zhou, Q.; Zhou, X. Enhanced Cascade R-CNN for Multiscale Object Detection in Dense Scenes From SAR Images. *IEEE Sensors J.* **2024**, *24*, 20143–20153. https://doi.org/10.1109/JSEN.2024.3393750
57. Yaseen, M. What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector. *arXiv* **2024**, arXiv:2408.15857.
58. Xiao, Y.; Xu, T.; Xin, Y.; Li, J. FBRT-YOLO: Faster and Better for Real-Time Aerial Image Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 8673–8681. https://doi.org/10.1609/aaai.v39i8.32937
59. Wang, C.; He, W.; Nie, Y.; Guo, J.; Liu, C.; Wang, Y.; Han, K. Gold-YOLO: Efficient Object Detector via Gather-and-Distribute Mechanism. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 51094–51112.