SGW-DETR: A Spectral-Guided Graph-Structured Wavelet Transformer for UAV Infrared Object Detection Under Degradation

Kaipeng Wang , Guanglin He * , Yuzhe Fu , Zelong Chen , Hao Zhang

Posted Date: 14 May 2026

doi: 10.20944/preprints202605.0956.v1

Keywords: UAV infrared detection; composite image degradation; spectral feature learning; graph neural network; wavelet decomposition; transformer detection

 doi:10.20944/preprints202605.0956.v1

Article

SGW-DETR: A Spectral-Guided Graph-Structured Wavelet Transformer for UAV Infrared Object Detection Under Degradation

Kaipeng Wang [1], Guanglin He [1,*], Yuzhe Fu [1], Zelong Chen [1] and Hao Zhang [2]

[1] National Key Laboratory of Proximity Detection and Control, Beijing Institute of Technology, Beijing 100081, China
[2] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan 250000, China
* Correspondence: heguanglin@bit.edu.cn; Tel.: +86-150-1096-5661

Highlights

What are the main findings?

- SGW-DETR is a novel detection framework for infrared target detection in UAV remote sensing images under multi-type composite degradation conditions. Three synergistic innovation modules (FDSANet, GSFN, WCFA) respectively enhance frequency-domain adaptive feature extraction, graph-structured multi-scale fusion, and wavelet-guided contrastive decoding, while effectively suppressing background thermal radiation interference.
- A UAV infrared multi-type degradation dataset was constructed containing 4,686 images covering six degradation types (blur, rain, snow, fog, strong light, and electromagnetic interference) with component-level annotations. On this dataset, SGW-DETR achieves an mAP_{50} of 75.2%, with computational cost and parameter count reduced by 16.8% and 9.9% respectively, and inference speed reaching 85.5 FPS, comprehensively outperforming the baseline model RT-DETR and other mainstream methods.

What are the implications of the main findings?

- The study demonstrates that explicit robustness modeling for multi-type composite degradations (including blur, rain, snow, fog, strong light, and electromagnetic interference) can significantly improve UAV infrared target detection performance, thereby breaking the dependence of traditional detectors on standard clear imaging conditions.
- The proposed framework provides a lightweight and practical solution for real-time infrared target perception on resource-constrained UAV platforms, with broad application potential in scenarios such as nighttime reconnaissance and surveillance, adverse weather search and rescue, and traffic monitoring.

Abstract

Infrared object detection from unmanned aerial vehicles (UAVs) is critically challenged by multi-type composite degradation—including noise, blur, and low contrast—which severely undermines feature discriminability and multi-scale target perception. This study proposes SGW-DETR (Spectral-Guided Graph-structured Wavelet Detection Transformer), a novel framework built upon RT-DETR, incorporating three synergistic modules across the backbone, neck, and encoder. FDSANet (Frequency Domain Spectral Awareness Network) replaces the conventional ResNet backbone, integrating the Multi-Scale Frequency Perception Module (MSFPM), Selective Channel Frequency Decomposition (SCFD), and Dynamic Kernel Spectral Modulation (DKSM) to achieve instance-level adaptive spectral feature extraction without degradation-type supervision. The Graph-Structured Fusion Network (GSFN) combines the Adaptive Semantic Fusion Module (ASFM) with the Graph Structure Perception Module (GSPM), employing Gaussian kernel soft membership and two-stage message passing to explicitly model spatial topological dependencies among object components. The Wavelet-guided

 doi:10.20944/preprints202605.0956.v1

Contrast Feature Aggregation module (WCFA) restructures the Attention-based Intra-scale Feature Interaction (AIFI) encoder via a Haar-based Frequency Decomposition Unit (HFDU), decomposing features into foreground-edge and background-thermal components and achieving hierarchical foreground–background decoupling through nested dual-path causal contrastive attention. A UAV infrared degradation dataset comprising 4,686 images spanning six degradation types with component-level annotations was constructed for evaluation. SGW-DETR achieves 75.2% mAP_{50}, outperforming RT-DETR by 3.5%, while simultaneously reducing GFLOPs and parameter count by 16.8% and 9.9% at an inference speed of 85.5 FPS. Sustained performance gains on M3FD and IndraEye benchmarks further demonstrate the framework's cross-domain generalization capability, offering practical value for UAV-based surveillance, search-and-rescue, and border monitoring under adverse imaging conditions.

Keywords: UAV infrared detection; composite image degradation; spectral feature learning; graph neural network; wavelet decomposition; transformer detection

1. Introduction

UAV-based infrared object detection has broad application value in the field of remote sensing, enabling the execution of reconnaissance surveillance, search-and-rescue operations, and traffic monitoring tasks under conditions where visible-light imaging is severely limited [1,2]. By capturing the thermal energy radiated by targets themselves rather than relying on ambient illumination, infrared sensors possess inherent advantages for night operations and adverse weather conditions. However, UAV infrared detection systems face a fundamental challenge in real-world environments: imaging quality can be severely degraded by multiple physical interference factors, including atmospheric scattering, optical blur induced by platform vibration, extreme illumination, and electromagnetic interference in complex environments [3,4]. Under these composite degradation conditions, the discriminability of thermally radiating targets is substantially reduced, posing serious challenges for achieving robust and reliable object detection [5]. In particular, the aerial perspective of UAVs means that targets commonly exhibit multi-scale structures and partial occlusion, making the spatial topological relationships between targets and their components difficult for conventional detection frameworks to model effectively—a limitation that places even higher demands on robust target perception [6].

The rapid development of deep learning has driven continuous advances in infrared object detection. Convolutional neural network-based detectors (e.g., the YOLO series [7,8]) and Transformer-based end-to-end detection frameworks (e.g., RT-DETR [9]) have both achieved excellent results on standard benchmark datasets. Nevertheless, the feature representations learned by these models are adapted to specific data distributions during training, and they lack explicit frequency-domain adaptive modulation mechanisms to cope with the spectral distribution shifts caused by multi-type degradation at inference time, resulting in limited robustness when extracting discriminative features from targets [10]. In UAV-captured infrared scenes, background thermal radiation (e.g., heated road surfaces and building facades) further contaminates feature representations, causing persistent deterioration of foreground targets across multiple scales [11]. From an aerial perspective, the ubiquitous multi-scale and partially occluded targets are disproportionately affected because the discriminative high-frequency boundary information inherent to such targets is already sparse or incomplete, and thus particularly susceptible to suppression by degradation factors such as motion blur, atmospheric interference, and sensor noise [12].

Existing research on degradation-robust detection can be organized into two paradigms. The first is the restore-then-detect pipeline approach, which employs image restoration networks such as Restormer [13] and FFA-Net [14] as preprocessing modules prior to detection; MWFormer [15] further integrates multi-weather unified restoration into a single Transformer framework. However, a fundamental optimization misalignment exists between the restoration loss targeting visual quality and the task objective targeting detection performance, and gains in restoration quality do not

 doi:10.20944/preprints202605.0956.v1

necessarily translate into improvements in detection accuracy [16]. The second paradigm is end-to-end degradation-aware detection, which embeds degradation adaptability directly into the detection backbone through attention mechanisms or dynamic convolutions [17,18]. A recent example, MODE [19], introduces multi-modal guided mechanisms to improve adaptability to multiple degradation types, yet the degradation types covered still focus primarily on haze, rain, and snow, making it difficult to address the practical demands of real-world conditions where blur, strong illumination, electromagnetic interference, and other degradation types coexist.

Furthermore, existing multi-scale feature fusion methods treat features at each scale as independent entities, integrating cross-layer information via simple concatenation or element-wise summation, without the capacity for explicit modeling of spatial topological dependencies among object components. In infrared scenes where thermal radiation distributions are irregular and target component discriminability declines with increasing degradation intensity, this structural deficiency is further amplified [20,21].

To systematically address these issues, this paper proposes SGW-DETR, a UAV infrared object detection framework designed for multi-type composite degradation. The study targets cars (including the car front), trucks (including the truck front), and persons, with simulated degradation types covering blur, rain, snow, fog, strong light, and electromagnetic interference. Three synergistically designed innovative modules are introduced at the backbone, neck, and encoder stages of the RT-DETR architecture, respectively, constructing a complete degradation-robust detection system from three dimensions: frequency-domain adaptive feature extraction, graph-structured multi-scale fusion, and frequency-domain contrastive decoding attention. The main contributions of this paper are summarized as follows:

- FDSANet is proposed as a frequency-domain adaptive backbone network. Channel-level frequency decomposition and residual spectral modulation are achieved through the Residual Frequency Spectral Module (RFSM) within MSFPM; differentiated frequency-domain modeling with semantic-level awareness is realized through SCFD; and instance-level dynamic kernel spectral modulation is accomplished through DKSM—all without requiring degradation-type supervision, maintaining stable multi-scale discriminative feature representations under multi-type composite infrared degradation.
- GSFN is proposed, integrating ASFM and GSPM. Graph-structure modeling of spatial dependencies among target components is introduced, and cross-scale semantic fusion is calibrated adaptively through attention, significantly improving target feature discriminability under thermal radiation and degradation clutter conditions.
- WCFA is proposed, achieving hierarchical decoupling of foreground saliency enhancement and background thermal interference suppression under physically interpretable frequency-domain priors, through Haar frequency decomposition via HFDU and nested dual-path causal contrastive attention.

The remainder of this paper is organized as follows. Section 2 reviews related work on UAV-perspective infrared degradation detection, frequency-domain feature learning and graph-structured perception modeling, and infrared degradation detection datasets. Section 3 details the design of the proposed method. Section 4 presents experimental results and ablation studies. Section 5 concludes the paper.

2. Related Work

2.1. UAV-Perspective Infrared Degradation Object Detection

Research on robust UAV infrared detection under degradation conditions is relatively sparse and has largely been confined to single degradation types. Wu et al. [22] proposed UIU-Net, which achieves multi-level multi-scale representation learning and global-local contrast enhancement for infrared small targets through a "U-Net embedded within U-Net" architecture combined with resolution-preserving deep supervision and interactive cross-attention modules. Yuan et al. [23] proposed SCTransNet, which

 doi:10.20944/preprints202605.0956.v1

encodes global semantic differences effectively using spatial-channel cross-Transformer blocks with spatially embedded single-head channel cross-attention and complementary feed-forward networks on the long skip connections of U-Net, achieving infrared small-target detection performance that surpasses prior methods on multiple public datasets. Zhang et al. [24] proposed a method based on PicoDet that introduces a lightweight LCNet backbone, integrates squeeze-and-excitation modules, and improves the feature pyramid structure, achieving a 31 fps increase in frame rate and a 7% improvement in average precision for UAV infrared small-target detection on the HIT-UAV dataset. Randieri et al. [25] systematically reviewed the latest advances in obstacle and aerial-vehicle detection technology for UAVs operating under adverse conditions such as fog, rain, smoke, low light, and motion blur, analyzing the applicability of different sensors and the performance trade-offs of various methods, and identifying lightweight, adaptive, and all-weather real-time detection systems as future research directions. Wang et al. [26] proposed the JFD^3 framework, which uses a weight-sharing dual-branch architecture wherein a clean branch guides a blurry branch with feature-level supervision, combined with a frequency structure guidance module and a feature-consistency self-supervised loss, achieving excellent detection performance and real-time efficiency on the self-constructed IRBlurUAV infrared UAV motion blur dataset; however, this work remains focused on a single motion-blur degradation type. Collectively, these studies confirm that systematic research addressing multi-type degradation scenarios—covering blur, rain, snow, fog, strong light, and electromagnetic interference—in UAV infrared object detection remains scarce, which forms the core motivation of this paper.

2.2. Frequency-Domain Feature Learning and Graph-Structured Perception Modeling

The synergistic design of frequency-domain feature representations and graph-structured perception capabilities constitutes a critical pathway for improving detector discriminability in composite-degradation infrared scenes. On the side of frequency-domain feature learning, Chi et al. [27] proposed Fast Fourier Convolution (FFC), which achieves non-local receptive fields and cross-scale fusion and can directly replace standard convolutions in existing networks. Chen et al. [28] proposed dynamic convolution, which aggregates multiple parallel convolution kernels dynamically via attention mechanisms to enhance model representational capacity without increasing network depth or width, providing an important foundation for dynamic kernel spectral modeling. In the domain of wavelet-space feature learning, a Haar wavelet-based downsampling (HWD) module that maximizes information preservation while reducing spatial resolution is readily integrated into various CNN architectures [29], offering a physically interpretable prior for foreground–background frequency-domain decoupling in degraded scenes.

On the side of graph-structured perception modeling, Kipf and Welling [30] proposed Graph Convolutional Networks (GCN), realizing efficient graph-structured feature propagation with spectral-domain convolution operators and establishing the neighborhood aggregation paradigm for graph-based learning. Veličković et al. [31] further proposed Graph Attention Networks (GAT), achieving adaptive structure-aware aggregation through learnable attention weights. However, these methods rely on predefined fixed graph topologies or hard-threshold adjacency relationships, making graph structural stability difficult to guarantee when features are perturbed by degradation. Wang et al. [32] proposed Dynamic Graph Convolutional Neural Networks (DGCNN), which adaptively reconstruct the neighborhood graph at each layer in feature space and capture dynamic local topological structures with EdgeConv operators, laying the groundwork for a dynamic graph construction paradigm in which soft membership replaces hard thresholds. In infrared imaging scenes, the irregular spatial distribution of target thermal radiation and the degradation-induced decline in local discriminability make the spatial topological associations between target bodies and components difficult to capture through conventional fusion strategies. To address this structural deficiency, this paper introduces Gaussian-kernel soft membership functions in place of hard-threshold adjacency partitioning to dynamically construct graph structures, and completes feature propagation of spatial topological relationships through two-stage message passing from nodes to structure edges and back, achieving robust explicit

 doi:10.20944/preprints202605.0956.v1

modeling of spatial dependencies between targets and their components under multi-type composite degradation.

2.3. Infrared Degradation Detection Datasets

Existing UAV-perspective infrared object detection datasets—including HIT-UAV-Infrared-Thermal-Dataset [33], LLVIP [34], DroneVehicle Dataset [35], and RGBTDronePerson [36]—each emphasize different aspects of scene diversity and annotation scale, but their annotation types are uniformly singular: none includes component-level annotations, and none provides systematic coverage of multi-type composite degradation conditions such as blur, rain, snow, fog, strong light, and electromagnetic interference. To this end, this paper constructs a multi-type degradation infrared object detection dataset from the UAV perspective, covering three categories: cars (including the car front), trucks (including the truck front), and persons. Original images were collected from HIT-UAV-Infrared-Thermal-Dataset [33] (977 images), LLVIP [34] (983 images), DroneVehicle [35] (1,749 images), and RGBTDronePerson [36] (977 images), totaling 4,686 images, to ensure diversity in environments and target categories. The dataset was re-annotated with component-level detection labels: cars are labeled "c", the car front (the region from the foremost edge of the front end to the front door gap) is labeled "c_f", trucks are labeled "t", the truck front (the cab and the frame directly below it) is labeled "t_f", and persons are labeled "p".

The complete dataset is split into a training set (3,280 images), test set (937 images), and validation set (469 images) in a 7:2:1 ratio using a stratified sampling strategy to ensure consistent distributions across subsets. The degradation simulation dataset consists of two parts: single blur degradation, which applies mild, moderate, and severe blur processing to all clean images in a 4:4:2 ratio; and composite degradation, which superimposes five additional degradation types—rain, fog, snow, strong light, and electromagnetic interference—on top of blur, with 500 samples per type processed at three intensity levels (mild, moderate, severe) to simulate complex imaging conditions where multiple real-world interference sources coexist.

The specific degradation generation methods are as follows: motion blur is simulated via directional kernel convolution to model the relative motion caused by UAV platform vibration; defocus blur is modeled with a Gaussian defocus kernel to simulate sensor focal-length degradation [37,38]; rain streak effects are synthesized with a physics-driven rendering method that parameterizes raindrop trajectory direction, density, and transparency [39]; snowfall effects are synthesized by randomly scattering snowflake particles of varying sizes and transparency [40]; fog is generated using the Koschmieder atmospheric scattering model [41,42], producing physically consistent scattering images based on scene depth estimation and transmittance map computation; electromagnetic interference is generated by superimposing periodic stripe noise, salt-and-pepper noise, and random block artifacts [43,44]; and strong light effects are produced by saturating brightness to make the intensity values in the corresponding regions reach their upper limits [44,45]. Following established construction conventions for degraded-image detection benchmarks [46,47], these strategies are designed to systematically reproduce typical visual degradation scenarios encountered during actual deployment of target recognition systems, thereby enhancing the dataset's practical value for evaluating algorithm robustness.

As shown in Figure 1, single degradation at different intensities was applied to the images, demonstrating the degradation effects of different types and intensities on infrared targets. The first column shows clean images, the second column shows mildly degraded images, the third column shows moderately degraded images, and the fourth column shows severely degraded images.

Partial examples from the self-constructed dataset are shown in Figure 2. Red markings indicate persons; dark-blue markings indicate trucks, and light-blue markings indicate the truck front; yellow markings indicate cars, and orange markings indicate the car front. This dataset is restricted to academic use and can be provided upon reasonable request; please contact the corresponding author to apply.

 doi:10.20944/preprints202605.0956.v1

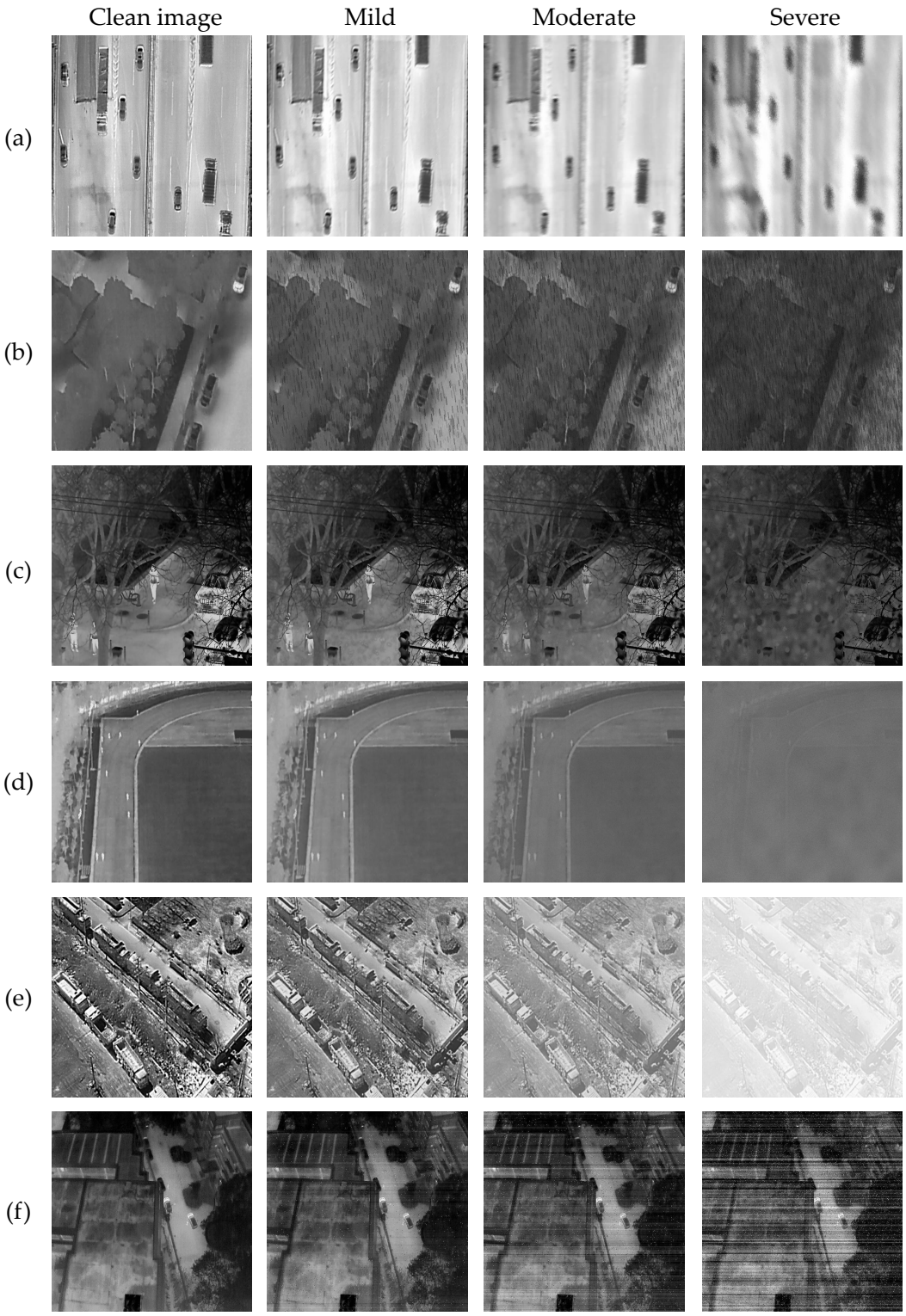

Figure 1. (**a**) Infrared images without degradation and with three intensity levels of infrared blur degradation. (**b**) Infrared images without degradation and with three intensity levels of infrared rain degradation. (**c**) Infrared images without degradation and with three intensity levels of infrared snow degradation. (**d**) Infrared images without degradation and with three intensity levels of infrared fog degradation. (**e**) Infrared images without degradation and with three intensity levels of infrared strong-light degradation. (**f**) Infrared images without degradation and with three intensity levels of infrared electromagnetic interference degradation.

 doi:10.20944/preprints202605.0956.v1

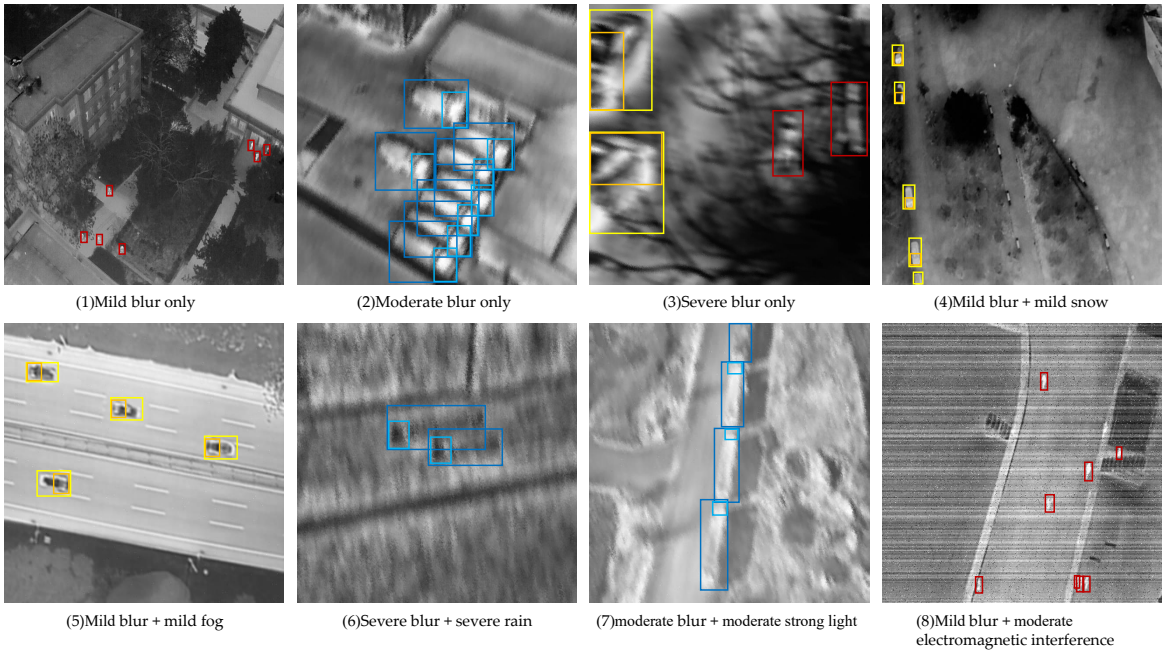

Figure 2. Examples from the self-constructed dataset, including images subjected only to blur degradation and images subjected to multi-source composite degradation.

3. Methods

As illustrated in Figure 3, SGW-DETR takes RT-DETR as the baseline and introduces synergistically designed innovative modules at three levels: backbone, neck, and encoder. At the backbone level, FDSANet replaces the static ResNet backbone with frequency-domain adaptive convolutional units; through the progressive frequency-domain modulation of MSFPM, SCFD, and DKSM, the backbone at each scale stage can maintain stable representations of discriminative target features even under the frequency-domain distribution drift induced by multi-type composite degradation. The neck-level redesign starts from explicit modeling of the spatial relationships among object components: after GSFN replaces the original Cross-scale CNN Feature Fusion (CCFF) structure, ASFM dynamically calibrates the semantic contributions of multi-scale feature branches across branches, while GSPM captures the topological dependencies between targets and their components through two-stage message passing under a soft membership graph framework; the combined effect of the two modules significantly improves the model's perception of occluded and thermally blurred targets. At the encoder level, WCFA restructures the AIFI module: fixed orthogonal Haar filters physically decompose features into low-frequency background structure components and high-frequency foreground edge components, after which a nested dual-path causal contrastive attention mechanism completes foreground saliency enhancement and background thermal interference suppression in sequence, with a causal precedence constraint between the two paths ensuring stability in the foreground–background decoupling process. The three modules jointly constitute an end-to-end infrared object detection framework; the structural design and operating principles of each component are detailed in the subsections that follow.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 May 2026 doi:10.20944/preprints202605.0956.v1

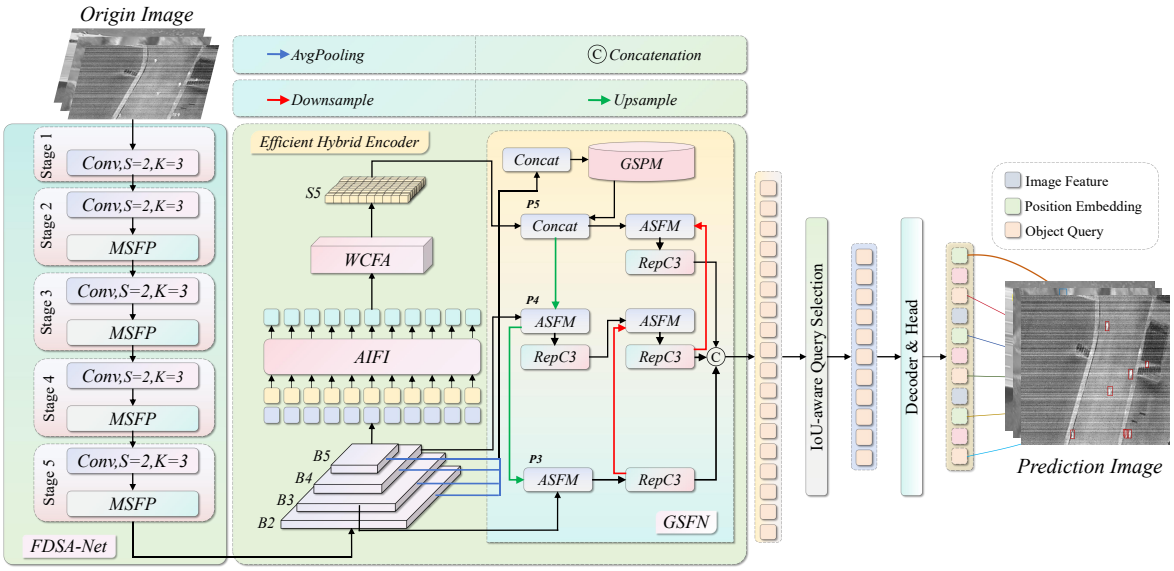

Figure 3. Overall SGW-DETR framework structure, illustrating the synergistic module design at the three levels of FDSANet backbone, GSFN neck, and WCFA encoder.

3.1. Frequency Domain Spectral Awareness Network

The ResNet backbone relied upon by conventional RT-DETR builds its feature hierarchy with static spatial convolution kernels whose weights are fixed in parameter space once training is complete, rendering the model incapable of responding to the frequency-domain distribution drift induced at inference time by blur, rain and snow, haze, low illumination, strong light, and electromagnetic interference in infrared images. This systematic degradation of discriminative boundaries—for human contours, car front structures, and local key components of trucks—accumulates progressively with increasing degradation intensity. To address these problems, FDSANet is proposed, embedding frequency-domain adaptive convolutional units into the multi-scale feature extraction stages of the backbone and replacing static spatial responses with dynamic frequency-domain weight modulation, so that feature representations maintain stable discriminative structures under heterogeneous degradation distributions. The structure of FDSANet is shown in Figure 4.

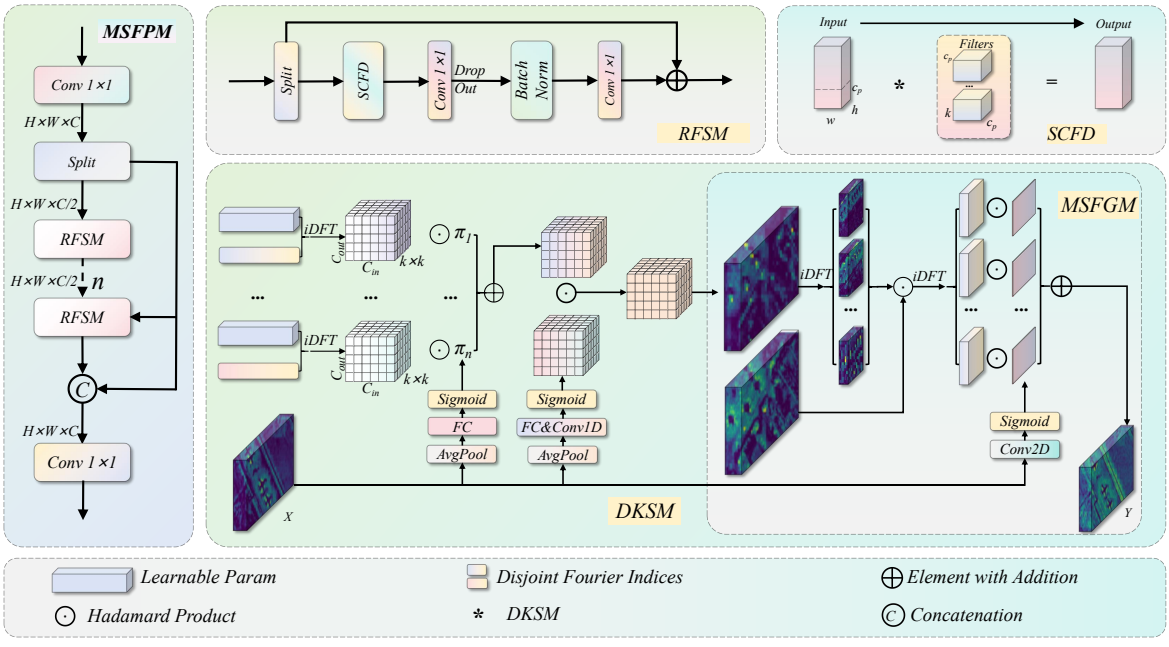

Figure 4. FDSANet overall structure, including the three submodules MSFPM, SCFD, and MSFGM and their frequency-domain adaptive processing flow.

 doi:10.20944/preprints202605.0956.v1

MSFPM replaces the original Bottleneck with RFSM units within a cross-stage branch framework, extending the modeling space for channel-level feature refinement from the spatial domain to the frequency domain. In the original C2f, Bottleneck units in each branch can only capture low-order statistical features within the local receptive field of fixed convolution kernels, lacking any explicit modeling capacity for the frequency-domain energy shifts caused by degradation. After replacement by RFSM, each branch performs parameterized modulation of the amplitude and phase components of input features in the spectral domain; the frequency-domain responses of multiple branches are synthesized in the sense of the operator tensor product into the output representation of the l-th stage:

$$Z_l = \phi\left(\int_{\mathcal{Y}^{(l)}} \bigotimes_{j=1}^{n} \mathcal{F}_j^{(l)}\left(Y_0^{(l)}; \theta_j^{(l)}\right) \mathrm{d}\mu\left(Y^{(l)}\right) \right) \tag{1}$$

where $\mathcal{Y}^{(l)}$ is the function space spanned by the branch outputs at stage l, $\mu(Y^{(l)})$ is the measure on this space, $\mathcal{F}_j^{(l)}\left(\cdot; \theta_j^{(l)}\right)$ is the j-th frequency-domain transform operator parameterized by $\theta_j^{(l)}$, and ϕ is the final aggregation map. Compared with the passive accumulation of spatial local responses by the original Bottleneck, this design equips the feature maps at each scale stage with the explicit capacity to suppress spectral components of degradation noise in the frequency domain, maintaining both high-frequency detail preservation and low-frequency structural stability of target semantic features under different degradation intensities, thereby improving the discriminability of multi-scale feature representations.

RFSM couples local frequency-domain spatial mixing with channel nonlinear transformation in residual form; the inclusion of the residual path allows the network to preserve the semantic structure of the original features through identity mapping when frequency-domain modulation gain is insufficient, preventing the over-suppression of target-discriminative spectral components by deep frequency-domain transforms. After the input passes through SCFD for frequency-domain spatial mixing of active channels, multi-layer channel transforms expand and compress the frequency-domain response nonlinearly; the accumulated gain along the residual path with respect to degradation intensity δ is expressed as:

$$\mathbf{x}_l^{\mathrm{out}} = \mathbf{x}_l + \int_0^{\delta} \sum_{r=1}^{R} \frac{\partial^2 \mathcal{M}_r^{(l)}\left(\mathcal{S}^{(l)}(\mathbf{x}_l; s)\right)}{\partial \mathbf{x}_l \, \partial s} \cdot \lambda_r^{(l)}(s) \, \mathrm{d}s \tag{2}$$

where $\mathcal{S}^{(l)}(\cdot; s)$ is the local frequency-domain mixing operator of SCFD at stage l under degradation intensity s, $\mathcal{M}_r^{(l)}$ is the r-th layer channel transform, $\lambda_r^{(l)}$ is the residual gain function with respect to degradation intensity, and R is the number of transform layers. The second-order mixed partial derivative explicitly characterizes the sensitivity of the frequency-domain modulation gain to changes in degradation intensity, allowing the network to maintain the spectral completeness of target semantic frequency components through the residual path even when the degradation distribution shifts, effectively suppressing spectral leakage of discriminative information in deep frequency-domain transforms.

SCFD constrains the scope of channel-level frequency-domain transforms via complementary partitioning of the frequency-domain support set; the design rationale is that channel features at different depths of the backbone exhibit significant semantic-level differentiation—high-semantic channels concentrate their frequency-domain energy in low-frequency structural components, while low-semantic channels respond more strongly to high-frequency texture details, and applying a uniform frequency-domain transform to both leads to misaligned modulation of semantic frequency components. By decomposing the complete frequency domain Ω into an active subdomain Ω_1 and a passive subdomain Ω_2—applying adaptive frequency-domain transforms to active channels while preserving the original frequency-domain responses for passive channels—semantically differentiated frequency-domain modeling is achieved:

 doi:10.20944/preprints202605.0956.v1

$$\mathcal{F}^{l}(\mathbf{x}_l) = \int \Big(\mathbb{I}_{\Omega_1}(\omega) \cdot \mathcal{F}_l^{\mathrm{FD}}(\mathbf{x}_l, \omega) + \mathbb{I}_{\Omega_2}(\omega) \cdot \mathbf{x}_l(\omega) \Big) \mathrm{d}\omega \tag{3}$$

where $\mathbb{I}_{\Omega_1}(\omega)$ and $\mathbb{I}_{\Omega_2}(\omega)$ are indicator functions over the corresponding subdomains, and $\mathcal{F}_l^{\mathrm{FD}}$ is the adaptive transform operator of stage l over the active frequency band. DKSM migrates the convolution kernel weights entirely to the frequency domain, with a kernel-space modulation network jointly predicting channel, filter, spatial, and kernel four-dimensional attention from the global statistics of the input, allowing the spectral structure of convolution kernels to undergo instance-level adaptive reconstruction as the input degradation type changes; its joint modulation response over the product measure space is:

$$\mathbf{y}_l = \exp\!\left(\int_{\Gamma^l} \ln \gamma_l(\omega)\, \mathrm{d}\mu_{\Gamma}^{l}(\omega) \right) \odot \mathcal{F}^{-1}\!\left[\prod_{k=1}^{K} \int_{\Omega_k} \mathbf{W}_l^{k}(\omega) \cdot \boldsymbol{\tau}_l^{k}(\omega)\, \mathrm{d}\omega \right] \tag{4}$$

where Γ^l is the product measure space composed of channel, filter, and spatial attention at stage l, μ_{Γ}^{l} is the counting measure on this space, $\mathbf{W}_l^{k}(\omega)$ is the weight distribution of the k-th component in the frequency domain Ω_k, $\boldsymbol{\tau}_l^{k}(\omega)$ is the frequency-domain kernel attention predicted by DKSM, and K is the number of parameter decompositions. Multi-Scale Frequency Gain Module (MSFGM) further applies selective gain constraints to different frequency bands on this basis, enhancing the spectral response of convolution kernels in the frequency bands where target semantic frequency components concentrate and suppressing it in the frequency bands dominated by degradation noise, thereby improving the consistency of boundary response precision and category confidence estimation for the detection head on human bodies, car fronts, and local key components of trucks under multi-degradation conditions.

By threading the frequency-domain adaptive mechanism through all levels of backbone feature extraction, FDSANet enables the detection model to maintain stable discriminative boundaries for human bodies, car fronts, and local key components of trucks under the cross-interference of multi-type degradation, achieving significant improvements over the original RT-DETR baseline in both feature robustness and target localization accuracy.

3.2. Graph-Structured Fusion Network

Conventional feature fusion strategies treat multi-scale features as mutually independent entities, integrating cross-layer information via simple concatenation or element-wise summation, without the ability to effectively model the intrinsic spatial topological dependencies between spatial regions. In infrared imaging scenes, thermal radiation features exhibit irregular spatial distributions, and target discriminability in specific regions declines substantially under composite degradation conditions including blur, rain, snow, fog, low illumination, strong illumination, and electromagnetic interference, further amplifying the structural deficiencies of existing fusion paradigms. To this end, GSFN is proposed to replace the original CCFF neck structure, introducing GSPM for explicit modeling of structural dependencies between spatial nodes and designing ASFM for cross-scale semantic integration, so that the network simultaneously possesses topological relationship awareness and cross-level feature co-representation capability under multi-type degradation scenarios. The GSFN overall structure and ASFM detailed structure are shown in Figure 5.

The overall flow of GSFN begins with a multi-level semantic aggregation stage: feature maps from different depths of the backbone are spatially aligned and fused by ASFM into a unified multi-scale representation, then fed into GSPM for spatial topological dependency modeling; the output features are projected to the target dimensionality via linear projection and passed layer by layer to the detection decoder. The weighted aggregation of semantics across levels can be expressed as:

$$\mathbf{F}_{\mathrm{out}} = \int_{\mathcal{L}} \alpha(l) \cdot \mathcal{T}(\mathbf{F}(l)) \mathrm{d}l \tag{5}$$

where $\mathbf{F}(l)$ is the feature representation at continuous level l, $\alpha(l)$ is the level-importance density function learned by the attention mechanism, $\mathcal{T}$ is the composite transform operator encompassing

 doi:10.20944/preprints202605.0956.v1

graph-structure propagation and linear projection, and the integral spans all levels in the participating fusion level set $\mathcal{L}$.

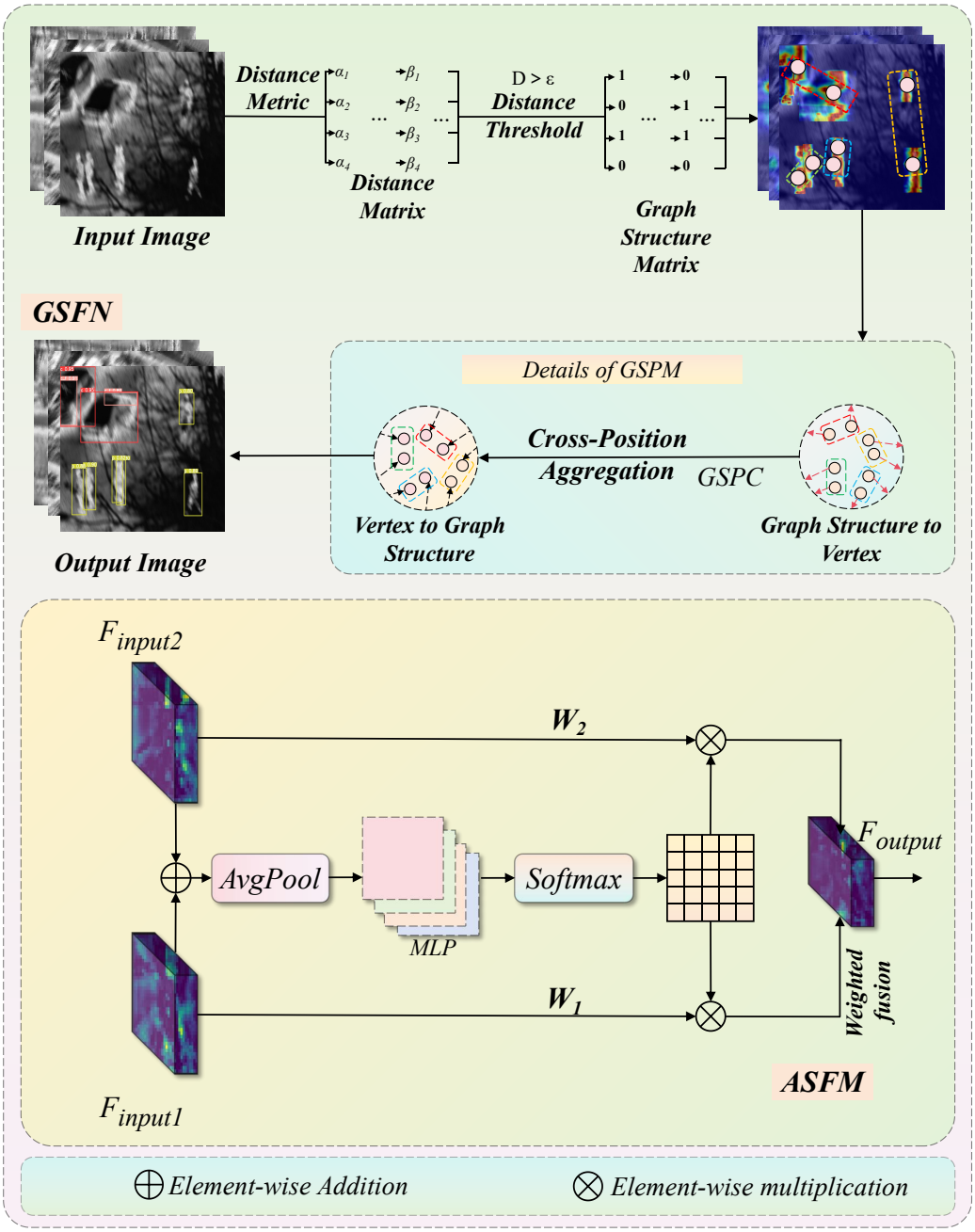

Figure 5. GSFN overall structure and ASFM detailed structure, illustrating the multi-scale feature adaptive semantic fusion and graph-structured perception propagation flow.

ASFM replaces naive concatenation with attention-driven weighted summation, dynamically calibrating the contributions of each input feature branch. Given K channel-aligned input feature maps, ASFM sums all branches and extracts global context; per-branch attention weights are generated via a lightweight bottleneck structure and cross-branch normalization, and the fusion output is:

$$\mathbf{F}_{\mathrm{ASFM}} = \sum_{k=1}^{K} \frac{\alpha_k\left(\mathbf{F}^k\right)}{\sum_{i=1}^{K} \alpha_i\left(\mathbf{F}^i\right)} \odot \mathbf{F}^k \tag{6}$$

where $\mathbf{F}^k$ is the channel-aligned feature of the k-th branch, and α is the function mapping each branch's feature to its importance score. This mechanism ensures that semantically dominant features

 doi:10.20944/preprints202605.0956.v1

are reinforced while branches corrupted by strong illumination or electromagnetic interference are effectively suppressed.

GSPM dynamically constructs a graph structure on the flattened set of spatial feature nodes, capturing spatially correlated node groups as structure edges to model dependencies. Structure edge membership is determined by a Gaussian-kernel soft membership function:

$$h_e = \int_{\mathcal{V}} \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{x}_j)^2}{2\sigma^2}\right) \mathrm{d}\mu_{\mathcal{V}}(j) \tag{7}$$

where $d(\cdot,\cdot)$ is the distance metric in feature space, $\mu_{\mathcal{V}}$ is the counting measure on the node set $\mathcal{V}$, and σ is the bandwidth parameter controlling the scope of structure edge membership; soft membership in place of hard thresholds confers stronger robustness against feature perturbations on the graph structure. Graph Structure Propagation Convolution (GSPC) completes information propagation through two-stage message passing from nodes to structure edges and from structure edges back to nodes; the Node-Aggregation Path Accumulator (NAPA) performs normalized accumulation on source node features propagated along associated paths, with the node aggregation representation characterized by a weighted integral over associated paths:

$$\mathbf{x}_i^{\mathrm{agg}} = \int_{\mathcal{P}(i,j)} \frac{w(i,j)}{\int_{\mathcal{P}} w(i,j)\,\mathrm{d}\mu} \cdot \mathbf{x}_j \,\mathrm{d}\mu_{\mathcal{P}}(j) \tag{8}$$

where $w(i,j)$ is the association strength function between nodes i and j along the graph path, and the denominator normalizes the association strengths of all neighboring nodes. After graph-structure propagation, the feature map undergoes linear projection for channel compression and alignment; this projection, acting along the channel axis, is equivalent to a continuous linear integral:

$$\mathbf{F}^{\mathrm{proj}}(c') = \int_0^{C_{\mathrm{in}}} w(c', c) \cdot \mathbf{F}(c)\,\mathrm{d}c, \quad c' \in \{1, \ldots, C_{\mathrm{out}}\} \tag{9}$$

where $w(c', c)$ is the projection weight function from input channel c to output channel c', providing dimensionally unified inputs for subsequent level-by-level feature fusion at a very low parameter cost.

By systematically embedding graph-structure-aware branches and adaptive semantic fusion paths, GSFN endows the network with explicit perception of associations among target components such as car fronts, truck contours, and human body configurations, establishes semantically adaptive cross-scale information channels between pyramid levels, and significantly improves the model's feature discriminability for small targets, occluded targets, and thermally blurred targets under multi-type degradation conditions.

3.3. Wavelet-guided Contrast Feature Aggregation

Under multi-type degradation conditions, the thermal radiation from complex backgrounds and degradation noise continuously contaminate target responses in feature space, yet the homogeneous treatment of input features by the AIFI module prevents it from distinguishing the fundamental spectral differences between foreground targets and background interference. To address this, the Wavelet-guided Contrast Feature Aggregation module (WCFA) is proposed, as shown in Figure 6; it explicitly decomposes features into low-frequency background structure components and high-frequency foreground edge components via HFDU, and achieves hierarchical decoupling of foreground saliency enhancement and background thermal interference suppression through a dual-path contrastive attention mechanism. Frequency-domain decomposition is implemented by HFDU. This component applies a fixed orthogonal Haar filter bank to the input features, projecting the spatial response of each channel onto four complementary frequency subbands; the sum of the three high-frequency subbands constitutes the foreground component F_ψ, and the low-frequency approximation constitutes the background component B_ϕ, with the decomposition expressed as:

 doi:10.20944/preprints202605.0956.v1

$$F_\psi = \sum_{d\in\{H,V,D\}} \int X(u)\cdot h_d(x-u)\,\mathrm{d}u, \qquad B_\phi = \int X(u)\cdot h_\phi(x-u)\,\mathrm{d}u \tag{10}$$

where h_ϕ is the low-frequency approximation filter kernel; $\{h_H, h_V, h_D\}$ correspond to the horizontal, vertical, and diagonal high-frequency filter kernels, respectively; and the four together constitute a complete orthogonal Haar wavelet basis. The filter kernel weights remain fixed throughout the optimization process, preserving the physical interpretability of the frequency-domain decomposition without introducing additional learnable parameters. The low-frequency component B_ϕ encodes the spatially smooth distribution of background thermal radiation in the infrared image, while the high-frequency component F_ψ concentrates the directional edge responses highly correlated with target contours, jointly providing semantically explicit frequency-domain priors for the subsequent contrastive attention.

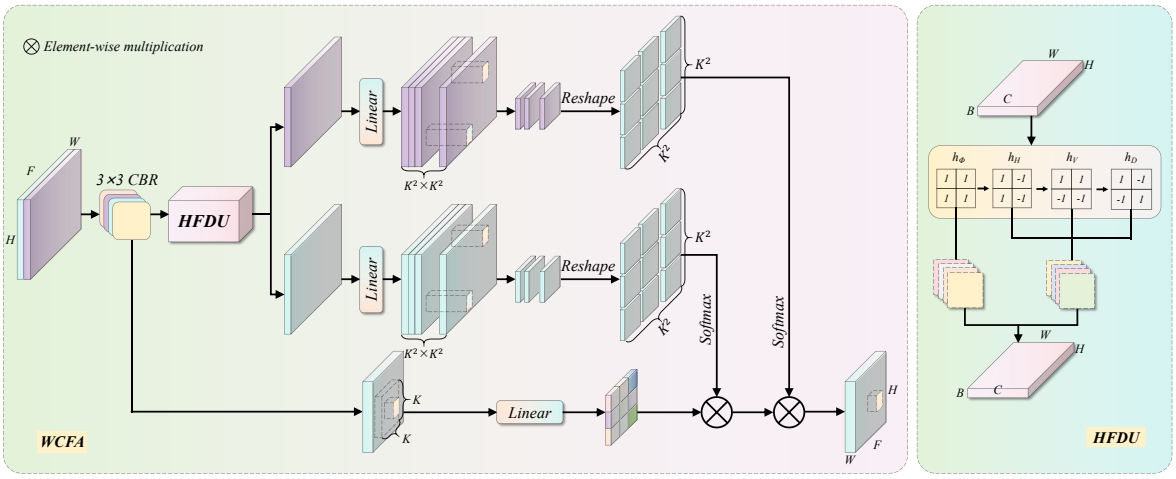

Figure 6. WCFA structure and HFDU detailed structure, illustrating the foreground–background decoupling flow of Haar frequency decomposition and the nested dual-path causal contrastive attention mechanism.

WCFA adopts a nested dual-path attention structure in which background attention is applied only after foreground refinement is complete, thereby establishing a deterministic causal dependency between the two paths. The overall output of this hierarchical contrastive aggregation is expressed as:

$$Z_{\text{out}} = \Phi_{\text{out}}\left(\sum_{i=1}^{N_h}\int_\Omega \mathcal{A}_{\text{bg}}^{(i)}(B_\phi;\mathbf{x})\cdot\left[\sum_{j=1}^{N_h}\int_\Omega \mathcal{A}_{\text{fg}}^{(j)}(F_\psi;\mathbf{x}')\cdot V^{(j)}(\mathbf{x}')\,\mathrm{d}\mathbf{x}'\right]^{(i)}\mathrm{d}\mathbf{x}\right) \tag{11}$$

where $\mathcal{A}_{\text{fg}}^{(j)}(\cdot)$ and $\mathcal{A}_{\text{bg}}^{(i)}(\cdot)$ are the contrastive attention kernels generated by the j-th and i-th attention heads based on the foreground and background components, respectively; $V^{(j)}(\mathbf{x}')$ is the value feature under the corresponding head; Ω denotes the feature spatial domain; and Φ_{out} is the output convolutional refinement operator. The nested integral structure mathematically establishes the causal precedence constraint of foreground enhancement over background suppression, ensuring that the feature transforms of the two contrastive paths are completed sequentially along a deterministic causal chain.

WCFA embeds frequency-domain-aware hierarchical contrastive attention into the encoder stage of the detector, forming a structural coupling between foreground target edge response enhancement and background thermal radiation interference suppression, equipping the model with the structural capacity to maintain stable target feature representations under complex degradation imaging conditions.

 doi:10.20944/preprints202605.0956.v1

4. Experiments

4.1. Datasets

M3FD [48] is a large-scale multi-scene benchmark dataset for infrared–visible fusion object detection tasks. The dataset contains 4,200 pairs of infrared and visible images subjected to rigorous spatiotemporal alignment, covering complex road scenes including daytime, overcast, and nighttime conditions, with annotation categories spanning six classes including people, cars, and buses, providing bounding box annotations for a cumulative 33,603 object instances. The rich scene coverage of M3FD across multiple illumination conditions and complex backgrounds makes it an important benchmark for evaluating the cross-scene generalization capability of detection models; in this paper it serves as the primary test set for cross-dataset generalization validation.

IndraEye [49] is an electro-optical–infrared bimodal dataset designed specifically for multi-sensor aerial perception tasks from UAV perspectives. The dataset contains 5,612 images and 145,666 annotated instances, with data simultaneously covering both visible-light and infrared modalities, supporting multi-modal learning and cross-modal domain adaptation research. Data collection spans multiple observation angles, flight altitudes, background scenes, and time periods, covering multiple sites across the Indian subcontinent, with 13 categories including persons, cars, buses, and tractors. This dataset aims to advance the development of robust visual systems under challenging conditions including low light. In this paper it serves as the secondary test set for cross-dataset generalization validation.

4.2. Implementation Details and Training Configuration

All experiments were conducted on Ubuntu 22.04 with a software environment built on Python 3.11 and PyTorch 2.3.0, using a single NVIDIA GeForce RTX 3090 GPU (CUDA 12.1). During training, the AdamW optimizer was adopted with an initial learning rate and weight decay coefficient both set to 0.0001; each batch contains 8 samples, and the model was trained for a total of 300 epochs. To ensure stability and reproducibility of experimental results, all runs used the same random seed, and other hyperparameters follow the default configuration of RT-DETR. For inference efficiency evaluation, frames per second (FPS) was adopted as the metric, also measured on the RTX 3090 GPU. Input image resolution was uniformly resized to 640×640 pixels, with a batch size of 1. To ensure fairness of comparison across methods, inference timing for all tested models excluded data loading and post-processing stages.

4.3. Evaluation Metrics

To validate the effectiveness of the proposed SGW-DETR framework, several widely recognized quantitative evaluation metrics in the field of object detection were selected. For detection accuracy, Precision (P) and Recall (R) were adopted as basic evaluation indicators, where precision measures the proportion of true positives among the model's positive predictions, reflecting the detector's ability to suppress false alarms, and recall measures the proportion of actual target instances successfully detected, capturing the detector's performance in controlling missed detections. Building on these, the mean Average Precision at an Intersection over Union (IoU) threshold of 0.5—denoted mAP_{50} or mAP@0.5—was adopted as the primary performance evaluation metric, providing comprehensive assessment of localization accuracy and classification capability under relatively relaxed overlap criteria. Additionally, to more stringently evaluate bounding box localization quality, the mean Average Precision across IoU thresholds from 0.5 to 0.95 in steps of 0.05—denoted $mAP_{50\text{-}95}$ or mAP@0.5:0.95—was also computed. For computational efficiency, giga floating-point operations (GFLOPs) and model parameter count (Params) were adopted, measuring computational complexity and storage overhead respectively, which are important references for evaluating practical deployment value on resource-constrained remote sensing platforms. Drawing on these metrics, SGW-DETR's performance on car and other target detection tasks in degraded remote sensing images was comprehensively evaluated across both detection accuracy and computational efficiency dimensions.

 doi:10.20944/preprints202605.0956.v1

4.4. Ablation Study

4.4.1. FDSANet Ablation Study

FDSANet comprises three submodules—MSFPM, SCFD, and MSFGM—each serving a distinct functional role in the frequency-domain perception chain. To clarify the independent contribution of each component to final detection performance, an incremental ablation experiment was conducted on the infrared dataset by introducing the three components sequentially, quantitatively analyzing each module's specific function under multi-degradation imaging conditions. The experimental results are shown in Table 1.

Table 1. Incremental ablation results for FDSANet submodules.

ID	MSFPM	SCFD	MSFGM	P	R	mAP_{50}	mAP_{50-95}
1.base				81.0	66.4	71.7	39.9
2	✓			81.6	65.8	72.1	40.1
3	✓	✓		81.2	66.9	72.3	40.0
4.ours	✓	✓	✓	82.4	67.2	72.5	40.3

Examining the progressive changes across rows in Table 1, the three submodules exhibit distinct gain characteristics on different evaluation metrics. When MSFPM is introduced alone, mAP_{50} improves by 0.4% over the baseline, establishing initial multi-scale frequency-domain perception capacity. After adding SCFD, recall further improves by 1.1%, with differentiated frequency-domain modeling of active channels effectively enhancing coverage of target edge responses. After MSFGM is fully incorporated, precision improves by 1.2%, and mAP_{50} and mAP_{50-95} improve by 0.8% and 0.4% over the baseline, respectively; selective frequency-band gain constraints reinforce target semantic features while suppressing degradation noise spectral components, and the three-module combination drives simultaneous improvement in precision and recall.

4.4.2. ASFM Ablation Study

The core design of ASFM lies in the combination of cross-branch attention and cross-branch normalization; whether this combination outperforms other fusion strategies still requires experimental confirmation. With all other modules held constant, a systematic comparative experiment was conducted across five fusion strategies: simple concatenation, element-wise summation, SE channel attention, softmax weighted summation, and ASFM. The results are shown in Table 2.

Table 2. Comparative results for different feature fusion strategies in ASFM.

Fusion Strategy	Attention Type	Cross-Branch Norm	P	R	mAP_{50}	mAP_{50-95}
Concatenation	None	—	80.8	66.1	71.5	39.5
Element-wise sum	None	—	81.2	65.4	71.2	39.3
SE channel attention	Channel	—	81.5	67.4	72.3	39.8
Softmax weighted sum	Cross-branch	—	81.3	66.5	71.9	40.0
ASFM (Ours)	Cross-branch	✓	81.8	69.1	73.6	41.4

Comparing the quantitative metrics across fusion strategies, the absence of inter-branch interaction capacity proves to be the root cause limiting the performance ceiling of simple fusion methods: simple concatenation and element-wise summation achieve mAP_{50} values of only 71.5% and 71.2%, respectively. SE channel attention operates only within individual branches and cannot suppress

 doi:10.20944/preprints202605.0956.v1

corrupted features across branches, resulting in limited mAP_{50} improvement. Softmax weighting introduces cross-branch interaction but lacks normalization constraints, making noise contamination difficult to eliminate. ASFM, which simultaneously incorporates cross-branch attention and normalization mechanisms, improves recall by 3.0% over the simple concatenation baseline, with mAP_{50} and $mAP_{50\text{-}95}$ improving by 2.1% and 1.9% respectively, validating the complementary gains of the two synergistic design elements under complex degradation scenarios.

4.4.3. Overall Ablation Study of the Three Innovative Modules

Using RT-DETR-R18 as the baseline, eight ablation configurations spanning single-module, dual-module, and full three-module combinations were designed, with GFLOPs and parameter count simultaneously recorded for each configuration, to comprehensively evaluate the independent contributions and synergistic effects of each module along both computational cost and detection accuracy dimensions. The experimental results are shown in Table 3 and Figure 7.

Table 3. Overall ablation results for FDSANet, GSFN, and WCFA.

ID	FDSANet	GSFN	WCFA	GFLOPs	Params/M	P	R	mAP_{50}	$mAP_{50\text{-}95}$
1.base				57.0	19.87	81.0	66.4	71.7	39.9
2	✓			43.8	13.21	82.4	67.2	72.5	40.3
3		✓		61.2	21.92	81.8	69.1	73.6	41.4
4			✓	59.3	22.71	82.6	67.8	72.9	40.7
5	✓	✓		45.1	14.95	83.2	69.8	74.3	41.9
6	✓		✓	46.1	16.03	82.9	68.5	73.7	41.2
7		✓	✓	63.5	24.75	82.3	70.4	74.8	42.3
8.ours	✓	✓	✓	47.4	17.91	83.7	71.1	75.2	42.8

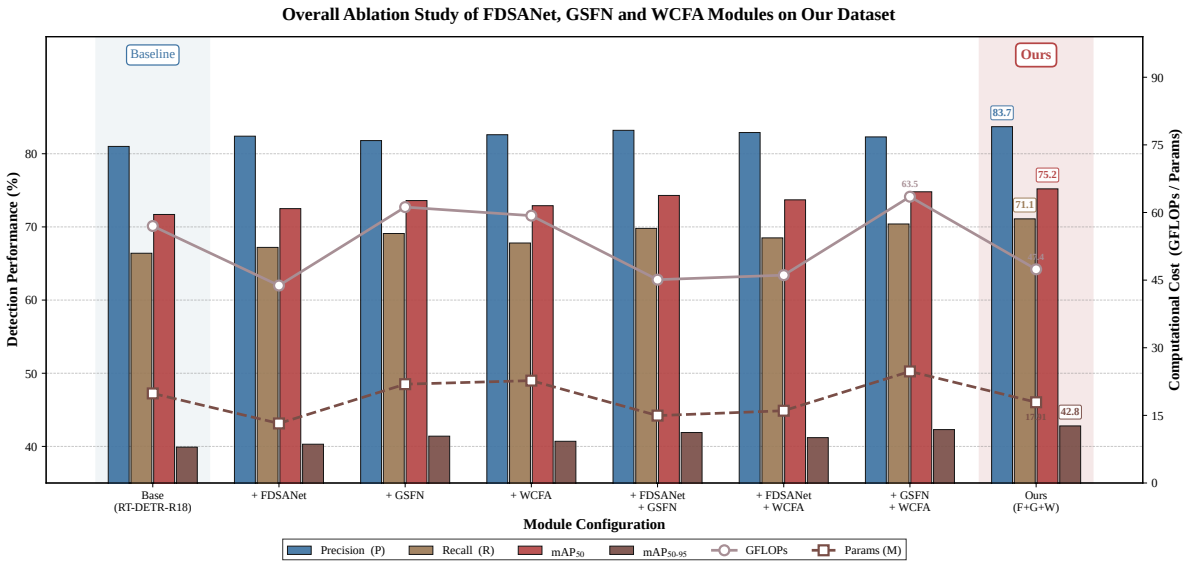

Figure 7. Visualization of the overall ablation study for FDSANet, GSFN, and WCFA. Bar charts (left axis) show the four accuracy metrics; line graphs (right axis) simultaneously present changes in GFLOPs and Params. The proposed method achieves optimal accuracy at lower computational cost.

Examining Table 3 row by row, the three modules contribute in different directions of accuracy, and dual-module combinations universally exhibit nonlinear synergistic gains. When FDSANet is introduced alone, it reduces GFLOPs by 23.2% and parameter count by 33.5% while improving mAP_{50} by 0.8%: its frequency-domain dynamic modulation mechanism replaces static convolution responses with dynamic weights, allowing the backbone to maintain stable discriminative feature structures across multi-type degradation distributions, thereby preserving feature discriminability at substantially reduced computational cost. When GSFN is introduced alone, recall improves by 2.7% and mAP_{50}

 doi:10.20944/preprints202605.0956.v1

by 1.9%: the explicit modeling of spatial dependencies by graph-structure relational edges breaks the isolated treatment of local target regions by conventional concatenation fusion, effectively mitigating missed detections in multi-degradation scenarios. WCFA, by decomposing features into foreground edge components and background thermal radiation components via fixed orthogonal Haar filters and then achieving hierarchical decoupling through nested dual-path contrastive attention, contributes a 1.6% improvement in precision, with particularly prominent effectiveness in suppressing complex thermal-radiation background interference.

Dual-module combination results further reveal intrinsic complementarity among modules; the FDSANet-plus-GSFN combination achieves additional mAP_{50} gains beyond the sum of the individual gains of each module, indicating a mutually reinforcing synergistic mechanism between frequency-domain robust backbone features and graph-structure fusion representations. The final three-module configuration, SGW-DETR, reduces GFLOPs by 16.8% and parameters by 9.9% while improving mAP_{50} and $mAP_{50\text{-}95}$ over the baseline by 3.5% and 2.9%, respectively, with precision and recall improving by 2.7% and 4.7%, comprehensively demonstrating the integrated advantages of the three innovative modules working in concert.

To validate the effectiveness of the SGW-DETR framework under multi-type composite degradation and to reveal the improvements each module makes to attention distributions, an attention heatmap visualization comparison experiment was designed, selecting six typical composite degradation detection scenarios; the results are shown in Figure 8. The first column shows clean images, the second column shows degraded images, and the remaining columns correspond to the attention heatmaps of four ablation configurations—RT-DETR-R18 (Base), +A, +A+B, and +A+B+C—respectively. The six degradation scenarios cover: person detection on a playground under mild blur (row a), car detection on a highway under moderate blur (row b), car and truck detection under mild blur combined with mild electromagnetic interference (row c), complex campus car detection under mild blur combined with moderate strong light (row d), urban overpass truck detection under moderate blur combined with moderate rain (row e), and urban person detection under moderate blur combined with moderate snow (row f). The baseline model exhibits consistent deficiencies across all scenarios, including scattered attention energy, insufficient target activation intensity, and reduced foreground–background separability.

With the progressive introduction of FDSANet (+A), GSFN (+A+B), and WCFA (+A+B+C), the attention heatmaps show systematic improvement. After introducing FDSANet (+A), DKSM's instance-level dynamic kernel spectral modulation equips the backbone with adaptive response to frequency-domain distribution drift, with attention converging significantly toward target regions; SCFD's differentiated frequency-domain modeling of high- and low-semantic channels effectively preserves discriminative edge responses under degradation conditions. After stacking GSFN (+A+B), GSPM's soft membership graph construction and two-stage message passing explicitly model the spatial topological dependencies between target components, with component-level activation responses trending toward completeness; ASFM dynamically suppresses interfered feature branches, eliminating semantic inconsistencies in cross-scale fusion. The complete model +A+B+C decomposes features into high-frequency foreground edges and low-frequency background thermal radiation components via WCFA's HFDU and accomplishes foreground saliency enhancement and background thermal interference suppression through nested dual-path causal contrastive attention, achieving the highest target activation intensity, most precise spatial localization, and lowest background misactivation across all six degradation scenarios. Overall, the stepwise introduction of the three modules drives a systematic evolution of SGW-DETR's attention heatmaps from scattered interference to precise focus, fully validating the effectiveness and practical deployment value of the synergistic improvement in target perception capability by FDSANet, GSFN, and WCFA under multi-type composite degradation conditions.

 doi:10.20944/preprints202605.0956.v1

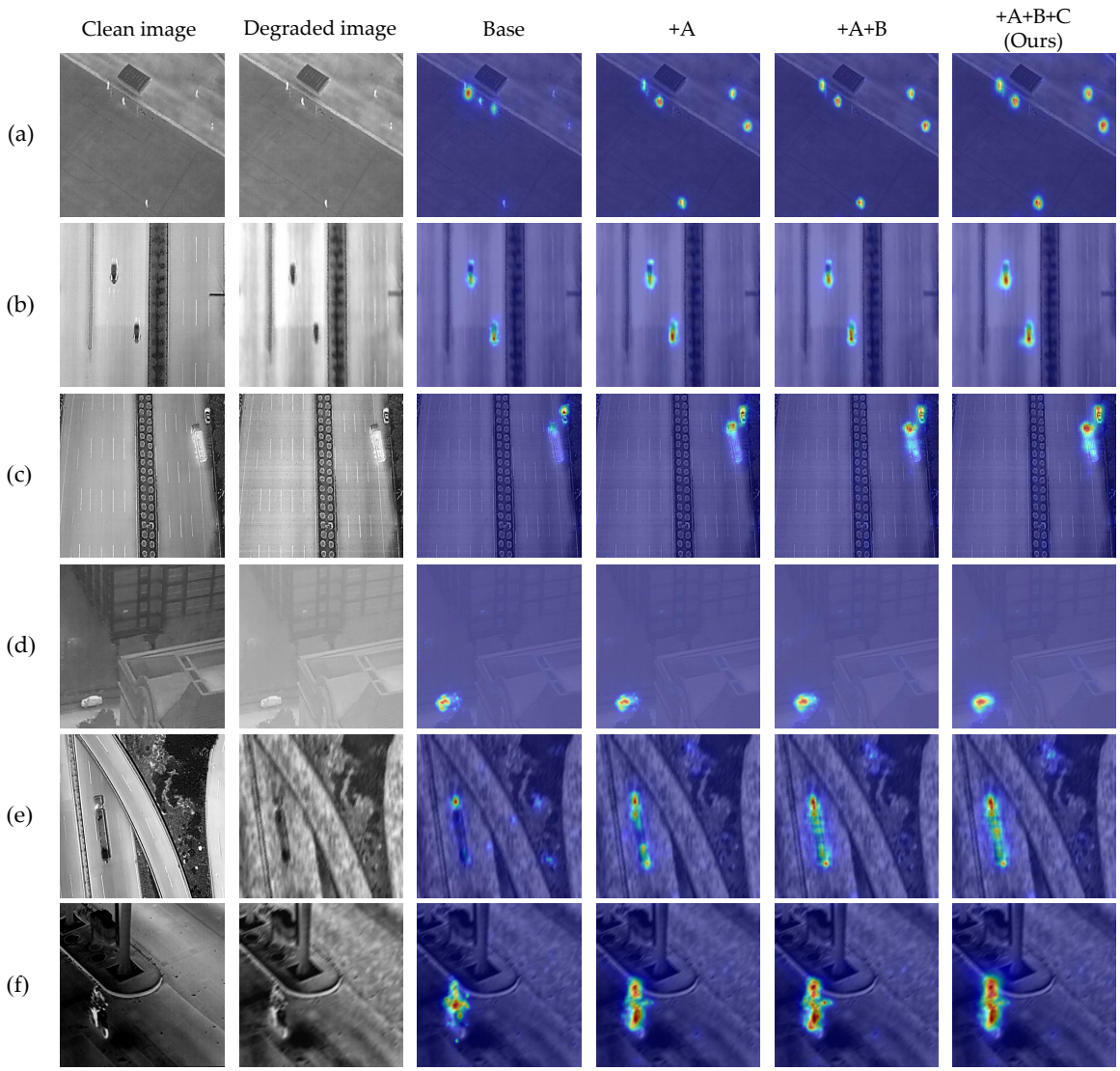

Figure 8. Attention heatmap visualization comparison between SGW-DETR and RT-DETR. (**a**) Heatmap visualization under mild blur degradation. (**b**) Heatmap visualization under moderate blur degradation. (**c**) Heatmap visualization under mild blur combined with mild electromagnetic interference. (**d**) Heatmap visualization under mild blur combined with moderate strong-light degradation. (**e**) Heatmap visualization under moderate blur combined with moderate rain degradation. (**f**) Heatmap visualization under moderate blur combined with moderate snow degradation.

4.5. Comparative Experiments

4.5.1. Backbone Comparison

The design intent of FDSANet is to achieve more robust feature extraction from degraded infrared images at lower computational cost, and its competitiveness must be verified in lateral comparison against similar backbone solutions. FAENet [50], MobileMamba [51], MambaOut [52], and MobileNetV4 [53] were selected as comparison schemes, and lateral comparison experiments were conducted under a unified evaluation configuration. The results are shown in Table 4.

Surveying the accuracy and efficiency data across all backbone schemes, the trade-off between lightweight design and high accuracy manifests as a clear see-saw relationship in most methods: MobileNetV4 and MobileMamba, despite compressing computational load, decrease mAP_{50} by 1.1% and 0.6% respectively relative to the baseline; FAENet and MambaOut achieve slight accuracy advantages over the baseline but exhibit increased GFLOPs and parameter count. FDSANet, while reducing GFLOPs by 23.2% and parameter count by 33.5%, improves mAP_{50} and $mAP_{50\text{-}95}$ by 0.8% and 0.4% respectively, achieving the best accuracy–efficiency balance among all comparison schemes,

 doi:10.20944/preprints202605.0956.v1

validating the practical effectiveness of the frequency-domain adaptive modulation design in infrared degradation imaging scenarios.

Table 4. Lateral comparison results for different backbone networks.

Model	GFLOPs	Params/M	P	R	mAP_{50}	mAP_{50-95}
ResNet (base)	57.0	19.87	81.0	66.4	71.7	39.9
FAENet [50]	60.8	22.14	81.9	67.0	72.2	40.1
MobileMamba [51]	53.4	17.95	80.5	65.7	71.1	39.2
MambaOut [52]	50.4	17.88	82.1	67.1	72.3	40.2
MobileNetV4 [53]	41.6	12.75	79.4	64.9	70.6	38.7
FDSANet (Ours)	43.8	13.21	82.4	67.2	72.5	40.3

4.5.2. Neck Fusion Network Comparison

The neck fusion network directly determines the quality of multi-scale feature integration; whether the graph-structure relationship modeling introduced by GSFN can stand out among mainstream fusion methods requires thorough lateral comparison. HS-FPN [54], PST [55], NSFPN [56], FAAFusion [57], and DPCF [58] were incorporated into the comparative experiment, with all methods tested connected to the same backbone and decoder configuration. The results are shown in Table 5.

Table 5. Comparative results for different neck feature fusion networks.

Model	GFLOPs	Params/M	P	R	mAP_{50}	mAP_{50-95}
CCFF (base)	57.0	19.87	81.0	66.4	71.7	39.9
HS-FPN [54]	40.3	28.05	81.6	66.9	72.1	40.2
PST [55]	64.8	24.35	82.7	67.8	73.0	40.8
NSFPN [56]	52.1	17.54	80.4	65.7	71.2	39.4
FAAFusion [57]	59.2	20.18	82.2	67.3	72.7	40.5
DPCF [58]	49.6	15.82	79.7	64.9	70.4	38.8
GSFN (Ours)	61.2	21.92	81.8	69.1	73.6	41.4

The comparison data show that existing fusion methods generally face a difficult trade-off between lightweight design and high recall: lightweight schemes such as DPCF, while compressing computational load, cause mAP_{50} to drop by 1.3% below the baseline; PST achieves relatively high mAP_{50} among comparison methods but its GFLOPs and parameter count are 3.6 and 2.43M higher than GSFN, respectively. GSFN improves recall by 2.7% over the baseline, with mAP_{50} and mAP_{50-95} improving by 1.9% and 1.5% respectively; the graph-structure spatial topological dependency modeling provides substantial gains in perception capability for occluded and small targets under multi-degradation scenarios, achieving the best overall accuracy among all comparison schemes.

4.5.3. Comprehensive Comparison with Mainstream SOTA Models

Placing SGW-DETR within a broader detection method landscape is a necessary step in evaluating its practical application value. A large-scale comparative experiment encompassing 16 mainstream methods—including single-stage detectors, two-stage detectors, and DETR-series models—was constructed, with accuracy, computational load, and inference speed comprehensively measured for each method under a unified dataset and evaluation protocol. The experimental results are shown in Table 6 and Figure 9.

 doi:10.20944/preprints202605.0956.v1

Table 6. Comprehensive performance comparison with mainstream SOTA detection models.

Model	GFLOPs	Params/M	P	R	mAP_{50}	mAP_{50-95}	FPS
One-Stage Object Detector							
Gold-YOLO-S [59]	45.9	21.59	81.2	65.7	62.8	41.2	103.2
YOLOv11-M [7]	67.7	20.03	81.6	64.1	64.1	42.2	112.7
YOLOv13-M [60]	63.2	18.42	83.1	66.5	66.3	40.6	115.3
Two-Stage Object Detector							
Faster R-CNN [61]	180.5	41.87	76.3	60.7	67.4	35.6	34.2
Cascade R-CNN [62]	227.7	69.14	77.9	62.3	68.8	37.1	31.6
DETR-Based Object Detector							
RT-DETR-R34	91.5	31.09	82.2	68.1	71.3	41.8	71.2
RT-DETR-R50	135.6	42.26	82.8	69.2	73.4	41.6	60.1
RT-DETRv2-R18 [63]	60.2	19.94	81.7	67.4	70.8	41.2	74.6
RT-DETRv2-R50 [63]	135.5	42.32	83.3	69.8	74.1	42.1	31.4
LW-DETR-M [64]	43.2	18.39	81.8	67.5	72.4	40.3	101.5
D-FINE-M [65]	56.6	19.24	84.3	70.4	73.8	41.4	75.7
DEIM-L [66]	91.3	30.91	85.8	72.9	73.8	41.9	49.7
RT-DETRv4-M [67]	56.8	19.20	85.0	71.4	74.3	41.1	75.9
DEIMv2-M [68]	50.1	17.66	85.9	72.6	73.8	40.3	89.7
RF-DETR-S [69]	61.9	32.02	84.6	71.2	72.1	42.1	68.6
RT-DETR-R18 (base)	57.0	19.87	81.0	66.4	71.7	39.9	78.1
SGW-DETR (Ours)	47.4	17.91	83.7	71.1	75.2	42.8	85.5

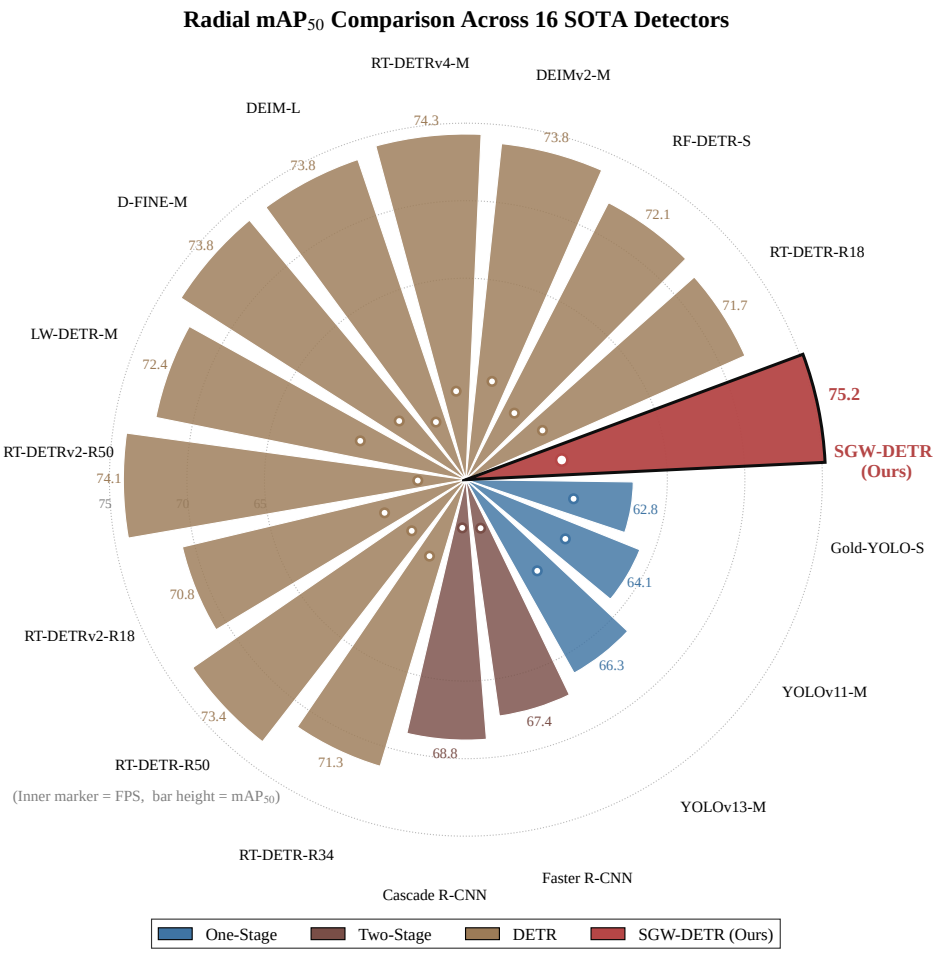

Figure 9. Radial mAP_{50} comparison with 16 mainstream SOTA detection models. Sector height represents mAP_{50}; colors distinguish detector categories; the inner white dot represents FPS.

 doi:10.20944/preprints202605.0956.v1

Surveying all comparison results in Table 6, different categories of detectors exhibit markedly different trade-off paths between accuracy and efficiency, and SGW-DETR demonstrates unique comprehensive advantages within this spectrum. The traditional two-stage methods Faster R-CNN and Cascade R-CNN, despite GFLOPs 133.1 and 180.3 higher than SGW-DETR respectively, still achieve mAP_{50} values 7.8% and 6.4% below SGW-DETR, indicating structural limitations of their static region proposal mechanisms in infrared degradation scenarios where high computational cost fails to yield accuracy advantages. Compared with the baseline RT-DETR-R18, SGW-DETR reduces GFLOPs by 16.8% and parameters by 9.9% while improving mAP_{50} and $mAP_{50\text{-}95}$ by 3.5% and 2.9%, respectively, with FPS increasing from 78.1 to 85.5 for a 9.5% inference speed improvement; the dual boost in lightweight design and accuracy jointly reflects the synergistic contributions of the frequency-domain backbone replacement and graph-structure fusion.

Against RT-DETRv2-R18 of comparable scale, SGW-DETR improves mAP_{50} and $mAP_{50\text{-}95}$ by 4.4% and 1.6%, making the structural design advantages even more pronounced under a similar parameter budget. Compared with the recently strong-accuracy D-FINE-M, SGW-DETR reduces GFLOPs by 16.2% while improving both mAP_{50} and $mAP_{50\text{-}95}$ by 1.4% and raising FPS by 13.0%, demonstrating that the proposed method retains detection capability surpassing current strong baselines while maintaining real-time inference performance. Notably, compared with RT-DETRv4-M and DEIMv2-M with similar parameter counts, SGW-DETR's mAP_{50} improves by 0.9% and 1.4% respectively, further confirming that the comprehensive gains of the three innovative modules do not stem from parameter scale advantages but rather from the structural discriminative capability formed by the synergy of frequency-domain adaptive representations, graph-structure relationship modeling, and wavelet contrastive attention. Among all participating comparison schemes, SGW-DETR achieves the best mAP_{50} and $mAP_{50\text{-}95}$ at the lowest computational cost, fully demonstrating that the synergistic design of the three innovative modules significantly strengthens detection capability in complex infrared degradation scenarios while meeting real-time inference demands.

4.6. Cross-Dataset Generalization Validation

A detection model's practical value is not only reflected in performance on benchmark datasets, but more importantly in whether it can maintain stable detection capability across data distribution gaps. SGW-DETR and the baseline RT-DETR-R18 were subjected to lateral comparison testing on the two representative public infrared datasets M3FD and IndraEye under identical hyperparameter configurations. The experimental results are shown in Table 7.

Table 7. Cross-dataset generalization validation results.

Dataset	Model	P	R	mAP_{50}	$mAP_{50\text{-}95}$
M3FD (primary)	RT-DETR-R18	0.861	0.819	0.853	0.571
	SGW-DETR	0.875	0.831	0.865	0.580
IndraEye (secondary)	RT-DETR-R18	0.924	0.923	0.941	0.732
	SGW-DETR	0.932	0.937	0.955	0.741

The comparison results from both datasets converge on the same conclusion: the performance gains of SGW-DETR are not the product of fitting to a specific data distribution, but stem from cross-domain robust feature representation capability inherent to the model structure itself. On M3FD, precision and recall improve over the baseline by 1.4% and 1.2%, respectively, with mAP_{50} and $mAP_{50\text{-}95}$ improving by 1.2% and 0.9%; on IndraEye, which has greater scene diversity, recall improves by 1.4%, mAP_{50} and $mAP_{50\text{-}95}$ improve by 1.4% and 0.9%, and precision likewise improves by 0.8%. These results demonstrate that the dynamic modulation capability of the frequency-domain adaptive backbone for degraded features, the spatial dependency modeling capability of the graph-structure neck, and the foreground–background decoupling capability of the wavelet contrastive attention collectively constitute a robust feature representation system with cross-domain generalization potential,

 doi:10.20944/preprints202605.0956.v1

and the model's adaptability to infrared imaging degradation characteristics shows promising domain generalization potential.

4.7. Visualization Results

4.7.1. Infrared Degradation Object Detection Visualization

To validate the detection performance of the proposed SGW-DETR framework on infrared targets in complex degradation environments, systematic visual comparison experiments were conducted across multiple typical scenarios. Specifically, YOLOv13-M, YOLOv11-M, and RT-DETR-R18 were selected as mainstream comparison methods for intuitive effect comparison. SGW-DETR's performance across different target categories and degradation types was comprehensively evaluated from the three dimensions of detection box localization accuracy, missed detection rate, and false detection rate. The relevant visualization results are shown in Figure 10.

As shown in Figure 10, SGW-DETR achieves substantially better detection performance than YOLOv13-M, YOLOv11-M, and RT-DETR-R18 across all six typical composite degradation scenarios—person detection among small woodland targets under mild blur (row a), suburban car and truck detection under moderate blur (row b), campus car detection under mild blur combined with moderate fog (row c), highway detection under mild blur combined with moderate electromagnetic interference (row d), urban intersection detection under moderate blur combined with moderate rain (row e), and factory truck detection under moderate blur combined with moderate strong light (row f). The comparison methods exhibit characteristic deficiencies across scenarios: row a shows missed detections and false detections in all three methods; row b shows widespread bounding box localization drift and insufficient component-level association awareness; row c shows large-scale misactivation on low-contrast targets; row d shows significantly elevated missed detection rates due to periodic stripe noise; row e shows that background rain streak interference aggravates missed detection and false detection of occluded persons; and row f shows notably deteriorated truck localization precision in brightness-saturated regions.

SGW-DETR achieves the best results across all these scenarios. FDSANet, through DKSM's instance-level dynamic kernel spectral modulation and SCFD's selective channel frequency-domain decomposition, maintains stable extraction of discriminative edge features under the frequency-domain distribution drift caused by heterogeneous degradation, effectively guaranteeing target detection rates in rows a and c. GSFN, through GSPM's soft membership graph construction and two-stage message passing that explicitly models spatial topological dependencies between target components, improves component-level missed detection in rows b and d, while ASFM's cross-branch attention calibration further improves bounding box localization accuracy in rows b and f. WCFA, by decomposing features into high-frequency foreground edges and low-frequency background thermal radiation components via HFDU and completing foreground enhancement and background suppression through nested dual-path causal contrastive attention, effectively reduces background misactivation in the composite rainy scene of row e, achieving more complete target coverage. The synergistic action of the three modules allows SGW-DETR to comprehensively surpass comparison methods across all six degradation scenarios with the highest localization accuracy, lowest missed detection rate, and lowest false detection rate.

 doi:10.20944/preprints202605.0956.v1

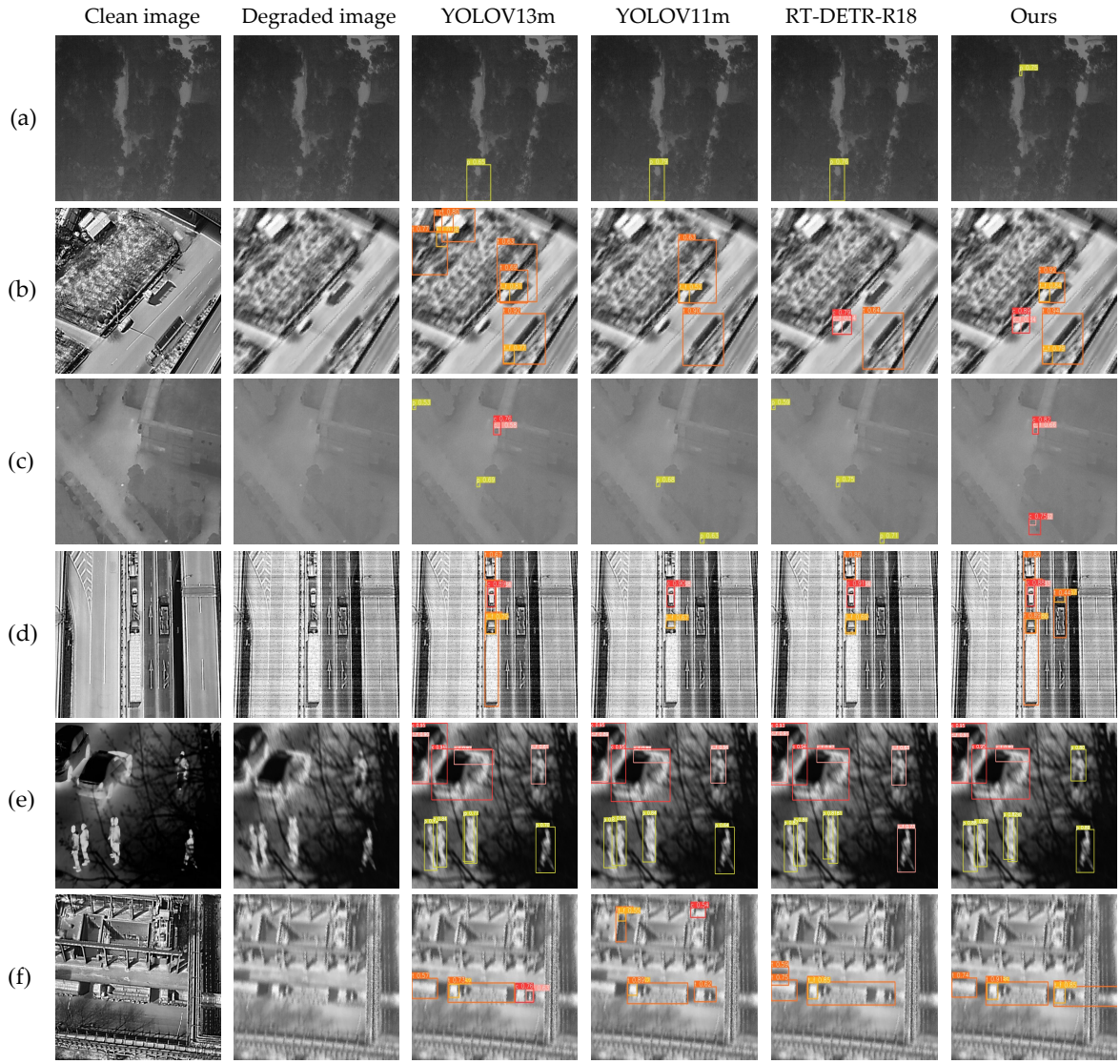

Figure 10. Qualitative detection result comparisons on the self-constructed multi-type degradation infrared object dataset. (**a**) Degradation condition: mild blur. (**b**) Degradation condition: moderate blur. (**c**) Degradation condition: mild blur combined with moderate fog. (**d**) Degradation condition: mild blur combined with moderate electromagnetic interference. (**e**) Degradation condition: moderate blur combined with moderate rain. (**f**) Degradation condition: moderate blur combined with moderate strong light.

Analysis of the failure cases in Figure 11 reveals that SGW-DETR still demonstrates certain advantages over the comparison methods in five extreme degradation scenarios: severe-blur highway car and truck detection (row a), mild-blur combined with severe-snow campus person detection (row b), mild-blur combined with severe strong-light factory car and truck detection (row c), moderate-blur combined with severe-fog urban person detection (row d), and moderate-blur combined with severe-rain wooded-area car detection (row e). In row a, all comparison methods exhibit component-level missed detections, while SGW-DETR maintains complete association between the target body and front labels via GSPM graph topology modeling. In row b, all three comparison methods fail to detect targets, whereas SGW-DETR achieves the only effective person detection by using DKSM dynamic kernel spectral modulation to adaptively compensate for frequency-domain energy shifts caused by snow degradation. In row c, SGW-DETR effectively detects some cars by explicitly decoupling high-frequency foreground edges from low-frequency background saturation regions through WCFA's nested dual-path causal contrastive attention. In row d, SGW-DETR maintains basic perception of extremely low-contrast targets by dynamically suppressing fog-contaminated feature branches via ASFM's cross-branch attention normalization mechanism. In row e, SCFD's selective channel

 doi:10.20944/preprints202605.0956.v1

frequency-domain decomposition partially restores car contour responses under rain streak occlusion. Nevertheless, SGW-DETR still exhibits clear limitations under these extreme degradation scenarios: multi-target recall under severe blur leaves considerable room for improvement; severe snow coverage causes significant declines in detection accuracy and recall; and detection performance under severe strong light, severe fog, and severe rain combined with blur all have substantial room for improvement. Improving detection robustness under extreme imaging conditions, ensuring completeness of component-level recognition, and increasing target recall rate are the core directions that future research must break through.

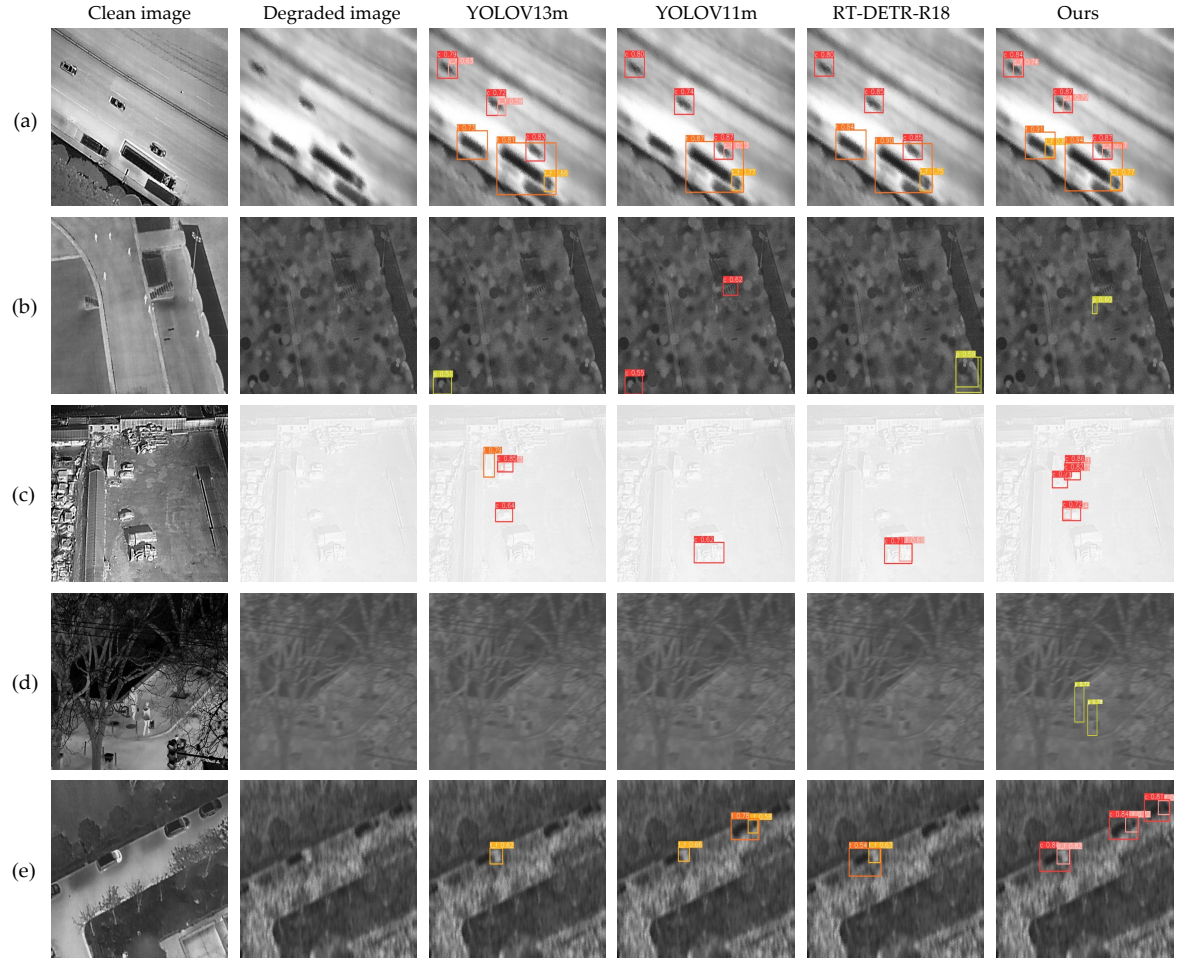

Figure 11. Failure case analysis. (**a**) Degradation condition: severe blur. (**b**) Degradation condition: mild blur combined with severe snow. (**c**) Degradation condition: mild blur combined with severe strong light. (**d**) Degradation condition: moderate blur combined with severe fog. (**e**) Degradation condition: moderate blur combined with severe rain.

4.7.2. M3FD Dataset Visualization Analysis

To validate the cross-domain generalization capability of SGW-DETR in heterogeneous infrared traffic scenarios, three typical scenarios were selected on the M3FD dataset—multi-lane multi-scale highway vehicle detection (first column), low-illumination urban road small-target detection (second column), and dense-occlusion urban building area detection (third column)—and compared visually with RT-DETR-R18, as shown in Figure 12. In the highway scenario, near-distance large-target trucks and far-distance small-target cars coexist; RT-DETR-R18 exhibits insufficient perception of small-scale targets, with missed detections and misclassification of trucks as cars. SGW-DETR, relying on MSFGM's multi-scale frequency-band gain constraints and GSPM's graph topology modeling, explicitly captures the spatial structural associations of targets at different scales in multi-lane scenes, with localization accuracy and target coverage both significantly superior to the baseline. In the

 doi:10.20944/preprints202605.0956.v1

low-illumination urban road scenario, extremely weak illumination causes far-distance small-target thermal radiation responses to nearly disappear into background noise; RT-DETR-R18 fails to detect distant occluded vehicles. SGW-DETR, by decomposing features into high-frequency foreground and low-frequency background components via WCFA's HFDU and performing foreground enhancement and background noise suppression through nested dual-path causal contrastive attention, successfully detects all vehicle targets completely. In the dense-occlusion urban scenario, RT-DETR-R18 experiences people missed detection due to highly aliased target boundary features; SGW-DETR, relying on ASFM's cross-branch attention calibration and GSPM topology modeling, effectively distinguishes the discriminative edge responses of each occluded instance, achieving complete detection of car and person targets. These results demonstrate that the synergistic action of FDSANet, GSFN, and WCFA endows SGW-DETR with robust detection capability in heterogeneous scenarios including multi-scale, low-illumination, and dense-occlusion conditions, qualitatively supporting the quantitative conclusions of cross-dataset generalization performance.

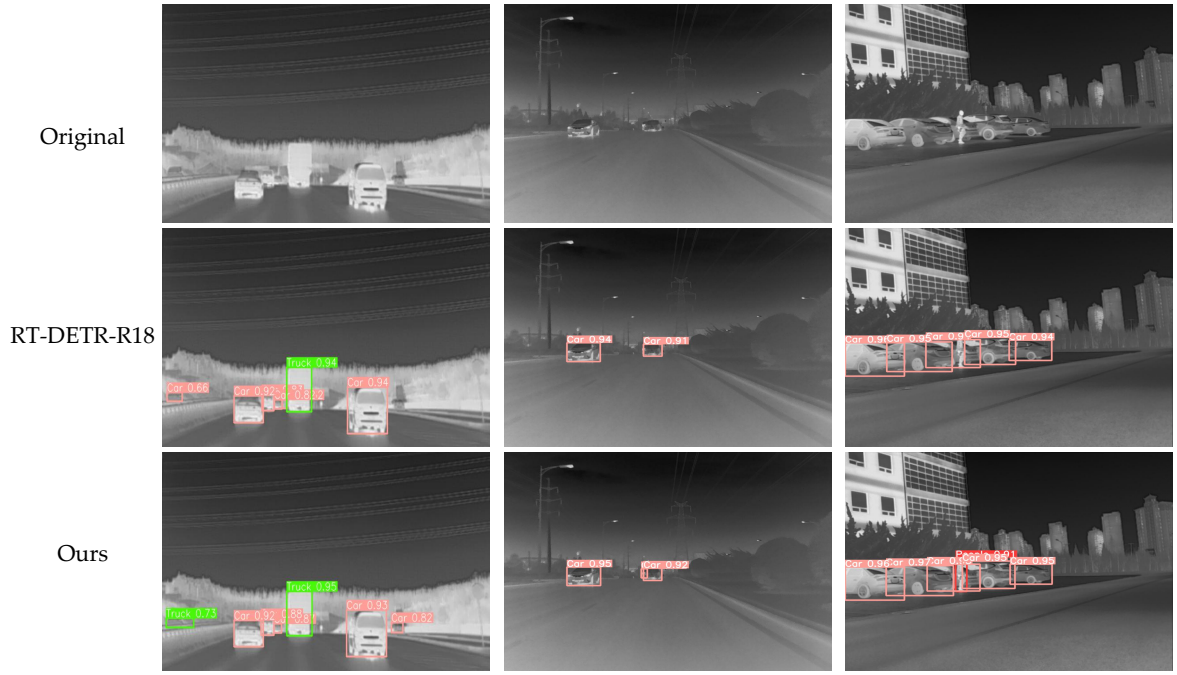

Figure 12. Visualization comparison of RT-DETR-R18 and SGW-DETR on the M3FD dataset.

4.7.3. IndraEye Dataset Visualization Analysis

To validate the generalization capability and detection performance of the proposed SGW-DETR framework on targets from heterogeneous UAV aerial perspectives, systematic visualization comparison experiments were conducted on the IndraEye dataset. The experiments selected typical scenarios including multi-category multi-scale targets on urban roads, multi-type occluded targets in complex road environments, and motion-blurred targets on urban roads, providing comprehensive evaluation of SGW-DETR's accuracy and robustness in detecting targets such as cars, motorcycles, and persons through systematic comparison with the baseline RT-DETR-R18 from both quantitative and qualitative dimensions. The visualization results are shown in Figure 13.

As shown in Figure 13, SGW-DETR demonstrates detection performance superior to the baseline RT-DETR-R18 across multiple typical complex UAV aerial scenes on the IndraEye dataset. In the multi-category multi-scale target scenario (first column), SGW-DETR achieves precise complete detection of highly discriminative target regions across scales by leveraging FDSANet's frequency-domain adaptive spectral modulation mechanism and DKSM's dynamic kernel spectral modulation, accurately recognizing cars, motorcycles, and persons at different scales; RT-DETR-R18, constrained by static spatial convolution kernels that cannot respond to spectral distribution shifts induced by heterogeneous imaging conditions, misclassifies cars as motorcycles at small scales. In the complex road environment

 doi:10.20944/preprints202605.0956.v1

multi-type occlusion target scenario (second column), SGW-DETR leverages GSPM's graph-structure modeling and ASFM's adaptive semantic fusion to explicitly model spatial dependencies between occluded target components, significantly outperforming RT-DETR-R18 in both detection completeness and confidence of densely occluded targets, effectively suppressing missed detections caused by weak activation signals being masked by adjacent target interference. In the urban road motion blur target scenario (third column), SGW-DETR, through HFDU's Haar frequency-domain decomposition and WCFA's nested dual-path causal contrastive attention mechanism, effectively maintains the separability of foreground edge activation and background thermal radiation components at the feature level under motion blur conditions, outperforming RT-DETR-R18 in both target recall rate and category confidence dimensions. These visualization results comprehensively validate SGW-DETR's technical advantages across multiple scenarios and degradation types in cross-domain generalization applications from a qualitative perspective, demonstrating that the three core innovative modules FDSANet, GSFN, and WCFA possess good collaborative adaptive capability in out-of-domain complex UAV infrared perception tasks.

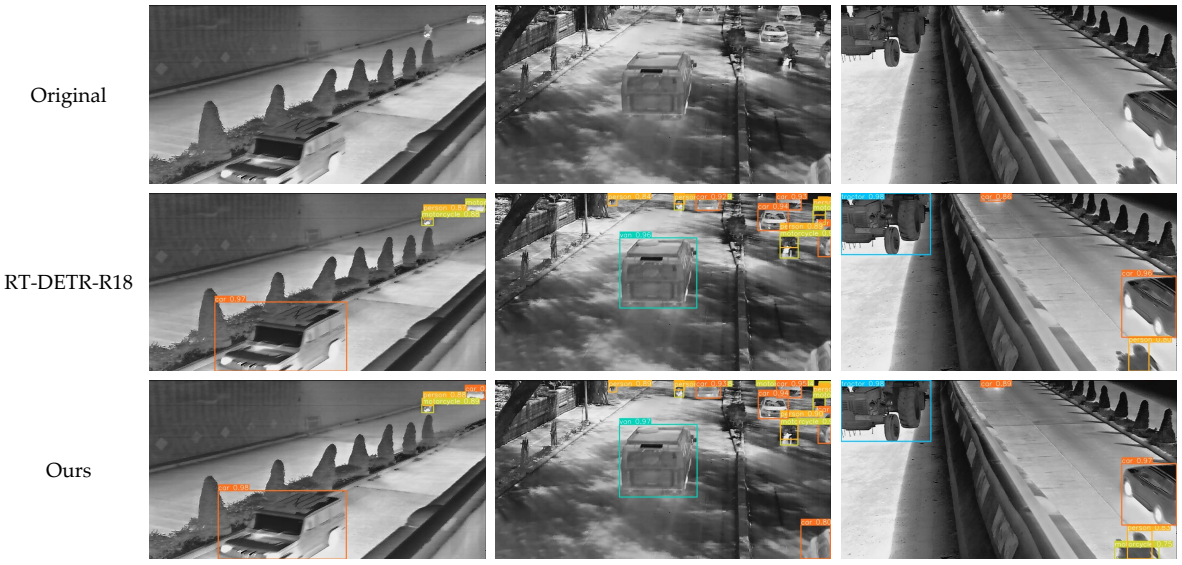

Figure 13. Visualization comparison of RT-DETR-R18 and SGW-DETR on the IndraEye dataset.

5. Conclusions

This paper presents SGW-DETR, a robust detection framework for infrared object recognition in UAV remote sensing images under multi-type composite degradation conditions, introducing three synergistically innovative modules at the backbone, neck, and encoder levels respectively. FDSANet enhances adaptive multi-scale spectral feature extraction capability through the residual frequency-domain spectral modulation of MSFPM, semantics-aware differentiated frequency-domain modeling of SCFD, and instance-level dynamic kernel spectral modulation of DKSM, effectively resolving the systematic degradation of feature representations caused by static convolution weights under composite degradation distributions including blur, rain, snow, fog, strong light, and electromagnetic interference. GSFN constructs a graph-structured feature pyramid with adaptive semantic fusion, explicitly modeling the spatial topological dependencies between object components through soft membership graph construction and two-stage message passing, significantly improving recall capability for occluded and thermally blurred targets. The WCFA encoder module effectively decouples foreground target responses from background thermal radiation clutter through the synergistic design of Haar frequency-domain decomposition and nested dual-path causal contrastive attention, suppressing background thermal interference while preserving foreground edge semantics. On the self-constructed UAV infrared multi-type degradation dataset, SGW-DETR achieves 75.2% mAP_{50}, comprehensively surpassing the RT-DETR baseline and other mainstream methods while reducing computational cost and parameter count by 16.8% and 9.9%, respectively, at an inference speed of

 doi:10.20944/preprints202605.0956.v1

85.5 FPS. Sustained performance improvements on M3FD and IndraEye datasets further demonstrate the proposed framework's strong cross-domain generalization capability.

However, the current framework still exhibits clear limitations under extreme degradation conditions such as severe snow coverage, severe strong-light saturation, and extreme dense fog, particularly in terms of component-level recognition completeness and target recall rate in multi-target scenarios. Future work will focus on the following directions: exploring self-supervised or few-shot learning paradigms to reduce dependence on annotated degradation data; investigating physics-driven degradation generation models to further enhance training data diversity; and further studying the deployability of the proposed algorithm on resource-constrained UAV platforms.

Author Contributions: Conceptualization, K.W. and G.H.; methodology, K.W.; software, K.W.; validation, G.H. and Y.F.; formal analysis, K.W. and Z.C.; investigation, K.W. and H.Z.; 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.

Informed Consent Statement: Not applicable.

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

Acknowledgments: The authors thank the providers of the HIT-UAV-Infrared-Thermal-Dataset, LLVIP, DroneVehicle, RGBTDronePerson, M3FD, and IndraEye datasets for making their data publicly available.

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

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned Aerial Vehicle
SGW-DETR	Spectral-Guided Graph-structured Wavelet Detection Transformer
RT-DETR	Real-Time Detection Transformer
FDSANet	Frequency Domain Spectral Awareness Network
MSFPM	Multi-Scale Frequency Perception Module
RFSM	Residual Frequency Spectral Module
SCFD	Selective Channel Frequency Decomposition
DKSM	Dynamic Kernel Spectral Modulation
MSFGM	Multi-Scale Frequency Gain Module
GSFN	Graph-Structured Fusion Network
ASFM	Adaptive Semantic Fusion Module
GSPM	Graph Structure Perception Module
GSPC	Graph Structure Propagation Convolution
NAPA	Node-Aggregation Path Accumulator
WCFA	Wavelet-guided Contrast Feature Aggregation
HFDU	Haar-based Frequency Decomposition Unit
AIFI	Attention-based Intra-scale Feature Interaction
CCFF	Cross-scale CNN Feature Fusion
GCN	Graph Convolutional Network
GAT	Graph Attention Network
DGCNN	Dynamic Graph Convolutional Neural Network
FFC	Fast Fourier Convolution
HWD	Haar Wavelet-based Downsampling
mAP	Mean Average Precision

 doi:10.20944/preprints202605.0956.v1

IoU	Intersection over Union
FPS	Frames per Second
GFLOPs	Giga Floating-Point Operations

References

1. Qin, Y.; Kishk, M.A.; Alouini, M.-S. A Survey on Integrating UAVs Into Public Safety Networks: Advantages and Potential Risks. *IEEE Open J. Commun. Soc.* **2025**, *6*, 9343–9372. https://doi.org/10.1109/OJCOMS.2025.3607987
2. Jia, P.; Tian, D.; Wang, Y.; Yao, D.; Xu, R.; Sun, J.; Sun, H.; Zhang, B. Airborne Optical Imaging Technology: A Road Map in CIOMP. *Light Sci. Appl.* **2025**, *14*, 119. https://doi.org/10.1038/s41377-025-01776-3
3. Dadrass Javan, F.; Samadzadegan, F.; Toosi, A.; van der Meijde, M. Unmanned Aerial Geophysical Remote Sensing: A Systematic Review. *Remote Sens.* **2025**, *17*, 110. https://doi.org/10.3390/rs17010110
4. Jiang, C.; Zhou, X.; Chen, H.; Liu, T. UAV Positioning Using GNSS: A Review of the Current Status. *Drones* **2026**, *10*, 91. https://doi.org/10.3390/drones10020091
5. Liu, W.; Pang, J.; Zhang, B.; Wang, J.; Liu, B.; Tao, D. See Degraded Objects: A Physics-Guided Approach for Object Detection in Adverse Environments. *IEEE Trans. Image Process.* **2025**, *34*, 2198–2212. https://doi.org/10.1109/TIP.2025.3551533
6. Hong, J.; Wang, T.; Han, Y.; Wei, T. Multi-Target Tracking for Satellite Videos Guided by Spatial-Temporal Proximity and Topological Relationships. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–20. https://doi.org/10.1109/TGRS.2025.3539462
7. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725.
8. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524.
9. 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.
10. Dai, Y.; Zhang, T.; Zhou, W.; Kuang, K.; Long, K.; Lu, X.; Wang, S. A Semantic-Enhanced Transformer with Adaptive Fusion for Road Damage Detection. *Comput. Aided Civ. Infrastruct. Eng.* **2025**, *40*, 6391–6418. https://doi.org/10.1111/mice.70154
11. Deng, S.; Ma, Q.; Deng, S.; Chen, Z.; Lian, R.; Li, B.; Liu, A.; Xu, K.; Li, X.; Hu, H. NeRI: Implicit Neural Representation for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. https://doi.org/10.1109/TGRS.2025.3633281
12. Zhu, Z.; Yang, Y.; Qi, G.; Li, S.; Li, H.; Liu, Y. Seeing Clearly and Detecting Precisely: Perceptual Enhancement and Focus Calibration for Small-Object Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2026**, 1–15. https://doi.org/10.1109/TNNLS.2026.3651289
13. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.-H. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 18–24 June 2022; pp. 5728–5739.
14. Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; Jia, H. FFA-Net: Feature Fusion Attention Network for Single Image Dehazing. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 11908–11915. https://doi.org/10.1609/aaai.v34i07.6865
15. Zhu, R.; Tu, Z.; Liu, J.; Bovik, A.C.; Fan, Y. MWFormer: Multi-Weather Image Restoration Using Degradation-Aware Transformers. *IEEE Trans. Image Process.* **2024**, *33*, 6790–6805. https://doi.org/10.1109/TIP.2024.3501855
16. Sun, S.; Ren, W.; Wang, T.; Cao, X. Rethinking Image Restoration for Object Detection. In *Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 4461–4474.
17. Duan, T.; Li, G.; Zhang, H.; Sun, R.; Zhang, B.; Xun, Y.; Sun, J.; Mansurova, M. NER: Noise-Aware Enhancement-Assisted Real-Time Detection Transformer for Pavement Damage Detection in Hazy and Low-Light Environment. *Expert Syst. Appl.* **2026**, *323*, 132469. https://doi.org/10.1016/j.eswa.2026.132469
18. Ci, W.; Hou, H.; Lu, S.; Deng, H.; Ma, J. Multi-Module Enhanced Low-Light Object Detection: From Image Enhancement to Adaptive Feature Fusion. *Adv. Eng. Inform.* **2026**, *69*, 104097. https://doi.org/10.1016/j.aei.2025.104097
19. Nguyen, T.-D.; Le, D.-T. MODE: A Model-Agnostic Framework for Object Detection under Adverse Weather Conditions. *Pattern Recognit.* **2026**, *173*, 112868. https://doi.org/10.1016/j.patcog.2025.112868

 doi:10.20944/preprints202605.0956.v1

20. Kong, Q.; Zhou, H.; Wu, Y. NormFuse: Infrared and Visible Image Fusion with Pixel-Adaptive Normalization. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 2190–2192. https://doi.org/10.1109/JAS.2022.106112
21. Liu, Y.-B.; Liu, W.; Huang, H.-Y. O-Net: Bidirectional Encoding and Decoding Architecture for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–11. https://doi.org/10.1109/TGRS.2025.3629412
22. Wu, X.; Hong, D.; Chanussot, J. UIU-Net: U-Net in U-Net for Infrared Small Object Detection. *IEEE Trans. Image Process.* **2023**, *32*, 364–376. https://doi.org/10.1109/TIP.2022.3228497
23. Yuan, S.; Qin, H.; Yan, X.; Akhtar, N.; Mian, A. SCTransNet: Spatial-Channel Cross Transformer Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. https://doi.org/10.1109/TGRS.2024.3383649
24. Zhang, Q.; Zhou, L.; An, J. Real-Time Recognition Algorithm of Small Target for UAV Infrared Detection. *Sensors* **2024**, *24*, 3075. https://doi.org/10.3390/s24103075
25. Randieri, C.; Ganesh, S.V.; Raj, R.D.A.; Yanamala, R.M.R.; Pallakonda, A.; Napoli, C. Aerial Autonomy Under Adversity: Advances in Obstacle and Aircraft Detection Techniques for Unmanned Aerial Vehicles. *Drones* **2025**, *9*, 549. https://doi.org/10.3390/drones9080549
26. Wang, X.; Fang, H.; Li, Q.; Wang, L.; Chang, Y.; Yan, L. Blur-Robust Detection via Feature Restoration: An End-to-End Framework for Prior-Guided Infrared UAV Target Detection. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 10181–10189. https://doi.org/10.1609/aaai.v40i12.37986
27. Chi, L.; Jiang, B.; Mu, Y. Fast Fourier Convolution. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 4479–4488.
28. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic Convolution: Attention over Convolution Kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 13–19 June 2020.
29. Xu, G.; Liao, W.; Zhang, X.; Li, C.; He, X.; Wu, X. Haar Wavelet Downsampling: A Simple but Effective Downsampling Module for Semantic Segmentation. *Pattern Recognit.* **2023**, *143*, 109819. https://doi.org/10.1016/j.patcog.2023.109819
30. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907.
31. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *arXiv* **2018**, arXiv:1710.10903.
32. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic Graph CNN for Learning on Point Clouds. *ACM Trans. Graph.* **2019**, *38*, 146. https://doi.org/10.1145/3326362
33. Suo, J.; Wang, T.; Zhang, X.; Chen, H.; Zhou, W.; Shi, W. HIT-UAV: A High-Altitude Infrared Thermal Dataset for Unmanned Aerial Vehicle-Based Object Detection. *Sci. Data* **2023**, *10*, 227. https://doi.org/10.1038/s41597-023-02066-6
34. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. LLVIP: A Visible-Infrared Paired Dataset for Low-Light Vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Montreal, QC, Canada, 11–17 October 2021; pp. 3496–3504.
35. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-Based RGB-Infrared Cross-Modality Vehicle Detection via Uncertainty-Aware Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713. https://doi.org/10.1109/TCSVT.2022.3168279
36. Zhang, Y.; Xu, C.; Yang, W.; He, G.; Yu, H.; Yu, L.; Xia, G.-S. Drone-Based RGBT Tiny Person Detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *204*, 61–76. https://doi.org/10.1016/j.isprsjprs.2023.08.016
37. Carmichael, S.; Bhat, M.; Ramanagopal, M.; Buchan, A.; Vasudevan, R.; Skinner, K.A. TRNeRF: Restoring Blurry, Rolling Shutter, and Noisy Thermal Images with Neural Radiance Fields. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Tucson, AZ, USA, 28 February–4 March 2025; pp. 7980–7990. https://doi.org/10.1109/WACV61041.2025.00775
38. Wang, X.; Fang, H.; Li, Q.; Wang, L.; Chang, Y.; Yan, L. Blur-Robust Detection via Feature Restoration: An End-to-End Framework for Prior-Guided Infrared UAV Target Detection. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 10181–10189. https://doi.org/10.1609/aaai.v40i12.37986
39. Garg, K.; Nayar, S.K. Photorealistic Rendering of Rain Streaks. *ACM Trans. Graph.* **2006**, *25*, 996–1002. https://doi.org/10.1145/1141911.1141985
40. 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/preprints202605.0956.v1

41. 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
42. 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
43. 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
44. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*, 4th ed.; Pearson: Hoboken, NJ, USA, 2018.
45. 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, 19–25 June 2021; pp. 9157–9167.
46. Hendrycks, D.; Dietterich, T.G. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *arXiv* **2019**, arXiv:1903.12261.
47. 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.
48. Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; Luo, Z. Target-Aware Dual Adversarial Learning and a Multi-Scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 18–24 June 2022; pp. 5802–5811.
49. Manjunath, D.; Gurunath, P.; Udupa, S.; Gandhamal, A.; Madhu, S.; Sikdar, A.; Sundaram, S. Indra-Eye: Infrared Electro-Optical UAV-Based Perception Dataset for Robust Downstream Tasks. *arXiv* **2024**, arXiv:2410.20953.
50. Liu, Z.; Fang, T.; Lu, H.; Zhang, W.; Lan, R. MASFNet: Multiscale Adaptive Sampling Fusion Network for Object Detection in Adverse Weather. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. https://doi.org/10.1109/TGRS.2025.3558541
51. He, H.; Zhang, J.; Cai, Y.; Chen, H.; Hu, X.; Gan, Z.; Wang, Y.; Wang, C.; Wu, Y.; Xie, L. MobileMamba: Lightweight Multi-Receptive Visual Mamba Network. *arXiv* **2024**, arXiv:2411.15941.
52. Yu, W.; Wang, X. MambaOut: Do We Really Need Mamba for Vision? *arXiv* **2024**, arXiv:2405.07992.
53. Qin, D.; Leichner, C.; Delakis, M.; Fornoni, M.; Luo, S.; Yang, F.; Wang, W.; Banbury, C.; Ye, C.; Akin, B.; et al. MobileNetV4: Universal Models for the Mobile Ecosystem. In *Computer Vision—ECCV 2024*; Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2025; pp. 78–96.
54. Shi, Z.; Hu, J.; Ren, J.; Ye, H.; Yuan, X.; Ouyang, Y.; He, J.; Ji, B.; Guo, J. HS-FPN: High Frequency and Spatial Perception FPN for Tiny Object Detection. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 6896–6904. https://doi.org/10.1609/aaai.v39i7.32740
55. Hu, J.; Bai, T.; Wu, F.; Peng, Z.; Zhang, Y. P^2HCT: Plug-and-Play Hierarchical C2F Transformer for Multi-Scale Feature Fusion. *arXiv* **2026**, arXiv:2505.12772.
56. Yuan, M.; Meng, D.; Xi, Z.; Zhao, T.; Zhao, S.; Dai, Y.; Wei, X. Seeing Through the Noise: Improving Infrared Small Target Detection and Segmentation from Noise Suppression Perspective. *arXiv* **2026**, arXiv:2508.06878.
57. Gu, C.; Chen, L.; Gu, L.; Fu, Y. Fourier Angle Alignment for Oriented Object Detection in Remote Sensing. *arXiv* **2026**, arXiv:2602.23790.
58. Xu, W.; Zheng, S.; Wang, C.; Zhang, Z.; Ren, C.; Xu, R.; Xu, S. SAMamba: Adaptive State Space Modeling with Hierarchical Vision for Infrared Small Target Detection. *Inf. Fusion* **2025**, *124*, 103338. https://doi.org/10.1016/j.inffus.2025.103338
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*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 51094–51112.
60. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733.
61. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.

 doi:10.20944/preprints202605.0956.v1

62. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 18–22 June 2018.
63. Lv, W.; Zhao, Y.; Chang, Q.; Huang, K.; Wang, G.; Liu, Y. RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer. *arXiv* **2024**, arXiv:2407.17140.
64. Chen, Q.; Su, X.; Zhang, X.; Wang, J.; Chen, J.; Shen, Y.; Han, C.; Chen, Z.; Xu, W.; Li, F.; et al. LW-DETR: A Transformer Replacement to YOLO for Real-Time Detection. *arXiv* **2024**, arXiv:2406.03459.
65. 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.
66. 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, 10–17 June 2025; pp. 15162–15171.
67. Liao, Z.; Zhao, Y.; Shan, X.; Yan, Y.; Liu, C.; Lu, L.; Ji, X.; Chen, J. RT-DETRv4: Painlessly Furthering Real-Time Object Detection with Vision Foundation Models. *arXiv* **2025**, arXiv:2510.25257.
68. Huang, S.; Hou, Y.; Liu, L.; Yu, X.; Shen, X. Real-Time Object Detection Meets DINOv3. *arXiv* **2026**, arXiv:2509.20787.
69. Robinson, I.; Robicheaux, P.; Popov, M.; Ramanan, D.; Peri, N. RF-DETR: Neural Architecture Search for Real-Time Detection Transformers. *arXiv* **2026**, arXiv:2511.09554.