symmetry

MDPI

Article

DCM-DETR: A Lightweight Framework for Robust Infrared Small UAV Detection

Linlin Li [1], Jingyao Sun [1,*] and Haochen Hu [2,*]

[1] Xi'an Reseach Institute of High Technology, Xi'an 710025, China
[2] Xi'an Key Laboratory of Human-Machine Integration and Control Technology for Intelligent Rehabilitation, School of Computer Science, Xijing University, Xi'an 710123, China
* Correspondence: jingyao309@yeah.net (J.S.); 2508540402049@stu.xijing.edu.cn (H.H.)

Abstract

Small unmanned aerial vehicle (UAV) detection in low-altitude infrared imagery remains challenging due to extremely small targets, weak contrast, scarce appearance cues, and heavy background clutter, which often leads to missed detections, clutter-induced false alarms, and localisation drift. To address these issues, we propose Directional Context Modelling DETR (DCM-DETR), an end-to-end detector that strengthens weak target evidence via directional context modelling and scale-consistent feature aggregation. Specifically, we build a Directional Receptive-Field Enhancement (DRFE) backbone with C2f-APC units, introducing asymmetric padding to enlarge receptive fields while preserving faint target cues. We further design an Infrared-Enhanced Encoder (IEE), where a CSA-Block jointly captures directional context and local details to steer global interactions towards target-relevant regions. To suppress noise propagation and alleviate cross-scale misalignment, we employ Hierarchical Gated Fusion (HGF) and Residual Alignment (RA), enabling selective semantic modulation and consistent multi-scale alignment. Moreover, we incorporate a Magnitude-Aware Linear Attention AIFI (MALA-AIFI) module to enhance low-SNR responses with linear complexity. Experiments on SIRST-UAVB show that DCM-DETR improves mAP_{50} by 36.58% over YOLOv8n and by 1.09% over RT-DETR, while reducing parameters by 25.1M. On IRSTD, it yields a 2.01% gain in mAP_{50} and boosts speed from 47.43 FPS to 93.45 FPS. These results demonstrate that DCM-DETR achieves a strong accuracy–efficiency trade-off for infrared small UAV detection in cluttered low-altitude scenes.

Keywords: infrared small target detection; UAV detection; end-to-end detection; DETR; attention mechanism; multi-scale fusion

1. Introduction

With the rapid advancement of information technology and the continuous reduction in manufacturing costs, low-altitude unmanned platforms, particularly small unmanned aerial vehicles (UAVs), have gained widespread popularity in recent years [1–3]. Owing to their operational flexibility, low cost, and high task efficiency, small UAVs have been extensively deployed in applications such as news gathering, geographical surveying, search and rescue, equipment inspection, fire patrol, and environmental monitoring. However, the rapid proliferation of UAVs has also raised serious security and privacy concerns. Non-cooperative UAVs may intrude into sensitive or restricted areas and violate citizens' privacy, while improper operation can lead to secondary damage or safety incidents. In addition, unauthorised UAV flights pose substantial threats to airport operations and the

Check for updates

Received: 21 January 2026
Revised: 11 February 2026
Accepted: 14 February 2026
Published: 24 February 2026

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

safety of the surrounding airspace. Malicious actors may further exploit UAVs for illegal surveillance or other disruptive activities. Consequently, developing effective, reliable, and deployable small UAV detection systems for complex low-altitude environments has become a critical prerequisite for public safety and airspace management [4].

Infrared target detection technology identifies and locates targets based on the differences in infrared radiation between the target and its surrounding environment. Due to its insensitivity to lighting changes and ability to perform all-weather imaging, infrared detection is particularly suitable for UAV early warning and situational awareness in low-visibility conditions [5,6]. Existing infrared small target detection techniques can generally be classified into model-driven and data-driven approaches [7,8]. Traditional model-driven methods are based on signal processing hypotheses, leveraging the local high contrast or sparsity of small targets in infrared images to enhance saliency and suppress the background. Representative methods include Local Contrast Measurement (LCM) and its variants (e.g., NLCM) [9,10], morphological filtering (e.g., Top-Hat transform) [11], and low-rank sparse decomposition frameworks (e.g., IPI, NIPPS) [12,13]. These methods usually offer good computational efficiency, but their performance often hinges on strong priors and hand-tuned parameters, leading to limited robustness under dynamic clutter and cross-scene variations [14]. Meanwhile, with the development of deep learning, especially convolutional neural networks (CNNs), the data-driven paradigm has been widely introduced into IRST. Typical object detection frameworks include two-stage detectors (e.g., R-CNN and its variants) [15–17] and one-stage detectors (e.g., YOLO, SSD) [18–25]. The former generates candidate regions and then performs classification and regression, while the latter unifies localisation and classification in a single forward pass, offering higher real-time performance and being more suitable for time-sensitive tasks. In recent years, detection frameworks based on deep convolutional networks and transformers have made significant progress in general object detection [26,27]. The end-to-end DETR series, by using set-based matching and query mechanisms, avoids cumbersome post-processing steps [28]. RT-DETR further balances efficiency and accuracy, providing a feasible path for engineering deployment [29].

However, directly applying existing end-to-end detection frameworks to infrared small UAV detection still faces several inherent challenges [30]. First, the discriminative cues of extremely small targets rely on weak local energy peaks that are easily suppressed by successive downsampling and receptive field expansion, causing targets to vanish in deep feature representations [31]. Second, in heavily cluttered backgrounds, global interaction mechanisms tend to model background statistics rather than target-specific differences, which leads to attention focusing on texture structures or thermal noise and produces false responses [32]. Third, small target localisation depends on high-resolution shallow features, while deeper features suppress background interference. Naive multi-scale feature fusion may, therefore, introduce misalignment and noise propagation, resulting in localisation instability and degraded detection performance.

To address the above challenges, we propose Directional Context Modelling DETR (DCM-DETR), an end-to-end framework for infrared small UAV detection built upon the RT-DETR baseline. DCM-DETR systematically enhances backbone representation, cross-scale fusion, and encoder modelling to improve robustness under complex backgrounds. To alleviate the dilution of extremely small targets in deep representations, a Directional Receptive-Field Enhancement (DRFE) backbone is designed based on the Cross-Stage Partial Fusion with Asymmetric Padding Convolution(C2f-APC) unit. By constructing multi-directional displacement views via asymmetric padding, DRFE enlarges the effective receptive field while avoiding aggressive downsampling, thereby better preserving weak target cues. To mitigate false responses induced by background-dominated interactions

and cross-scale feature misalignment, an Infrared Enhancement Encoder (IEE) is introduced. The encoder integrates CSA-Block for saliency recalibration and further employs HGF and RA to suppress noise propagation and enforce scale-consistent alignment across scales. In addition, a Magnitude-Aware Linear Attention AIFI (MALA-AIFI) module is incorporated into the encoding stage to model global dependencies with linear complexity while enhancing sensitivity to weak target responses. Through these designs, DCM-DETR produces cleaner, well-aligned, and weak-target-sensitive feature representations while preserving end-to-end inference efficiency, thereby enabling more robust infrared small UAV detection. Notably, C2f-APC, CSA-Block, HGF, and RA are modules proposed in this work, while MALA-AIFI incorporates the MALA [33] module into AIFI. The main contributions of this work are summarised as follows:

- DCM-DETR is presented as a lightweight end-to-end detection framework for infrared small UAV detection. By enhancing backbone representation, encoder-stage global modelling, and cross-scale fusion alignment, the framework alleviates feature dilution of extremely small targets, suppresses background interference, and improves localisation stability, achieving a favorable trade-off between detection accuracy and computational efficiency.
- A DRFE backbone based on C2f-APC is developed to strengthen the representation of extremely small targets under cluttered backgrounds. APC expands the effective receptive field through multi-directional displacement views without increasing downsampling, enabling stable preservation of weak target cues with controlled computational overhead.
- An IEE composed of CSA-Block, HGF, and RA is proposed to mitigate background-dominated global interactions, cross-scale misalignment, and noise propagation. The encoder performs saliency recalibration, selective semantic modulation, and scale-consistent feature alignment, thereby enhancing small target separability and localisation robustness.
- MALA-AIFI is incorporated for efficient global dependency modelling. By adopting linear-complexity attention and magnitude correction, MALA-AIFI improves sensitivity to weak target responses and enhances the discriminability and stability of encoded features while maintaining real-time performance.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 details the architecture and key components of DCM-DETR. Section 4 presents experimental results and analysis. Section 5 concludes the paper and discusses future directions.

2. Related Work

2.1. Deep Learning-Based Infrared Small Target Detection

In recent years, infrared detection has evolved along two main directions. One is represented by YOLO-based CNN detectors, which focus on lightweight architectures and multi-scale feature fusion for efficient deployment. The other is represented by end-to-end DETR- or Transformer-based frameworks, which enhance the modelling of complex scenes through alignment, object queries, and global interaction mechanisms. In the one-stage CNN approach represented by YOLO [19–25], for detecting infrared UAV targets in aerial and complex ground backgrounds, ITD-YOLOv8 [34] improves multi-scale and small target detection capability with a lightweight backbone, neck structure reconstruction, and attention embedding, all while maintaining a low computational cost. For dealing with stronger background interference and weak textures, IUAV-YOLO [35] further incorporates global attention and spatial context modelling to improve target-background separability. To address moving small targets, Multi-YOLOv8 [36] enhances robustness

by using multi-input strategies, including the current frame combined with optical flow and background suppression. GT-YOLO [37] improves detection performance in complex scenes by optimising feature fusion, enhancing small target detection, and handling dense targets effectively. EGISD-YOLO [38] proposes an infrared ship target detection network guided by edge information, which demonstrates higher accuracy and robustness in the infrared ship detection task, especially in complex backgrounds and low-contrast scenarios, through dense CSP structures, deconvolution channel attention modules, and edge-guided structures. YOLO-MS [39] introduces a deep learning-based infrared small target detection method that combines image super-resolution technology and multi-scale observation.

Simultaneously, Transformer-based and DETR-based approaches have rapidly advanced in infrared detection. The main motivation is to reduce reliance on post-processing through global modelling and query mechanisms, while enhancing the expression ability for complex occlusions and dense targets. CAGTDet [40] uses a cascade alignment-guided Transformer, which mines complementary information after region alignment. DPDETR [41] decouples categories and different modality positions explicitly to mitigate localisation bias caused by modality misalignment. DAMSDet [42] further introduces competitive query selection and adaptive fusion/cross-attention to adapt to dynamic changes in complementarity across different scenes. These studies reveal that the core challenge of applying Transformers to infrared detection is not just global modelling but also the stability of spatial alignment and query selection across modalities and scales.

In the IRST or small UAV scenario, end-to-end DETR models have begun to deeply integrate task priors. ISTD-DETR [43] combines super-resolution preprocessing with an improved RT-DETR and enhances weak target responses through feature layer expansion and dedicated encoding/augmentation modules. Additionally, works such as PHSI-RTDETR [44] explore real-time deployment strategies for RT-DETR in UAV infrared small-target detection, focusing on lightweight backbones, cross-scale interactions, and loss function design. IR-DETR [45] improves infrared small target encoding discriminability through multi-layer feature fusion and lightweight interactions, demonstrating the engineering needs for multi-scale interactions and efficient modelling. IRTransUNet [46] embeds a Transformer into a U-Net style encoder-decoder structure, maintaining pixel-level detail pathways while introducing long-range dependencies to alleviate insufficient small target details and background interference. GM-DETR [47] addresses the issues of dense distribution and insufficient texture in small UAV group targets by introducing finer-grained cross-layer fusion and supervised sampling sparsification to improve query quality.

2.2. Infrared Small Target Feature Extraction and Encoding

Infrared small target detection focuses on feature extraction driven by edge and shape priors, background modelling, clutter suppression, multi-scale feature fusion, and cross-scale alignment. To address the information loss caused by the sparse textures and shape similarity to bright spots in infrared small targets, research has started emphasising explicit modelling of edges and shape structures. The Edge-aware gated transformer [48] enhances fine-grained structural representation through edge guidance and gating mechanisms. MEAC [49] couples edge cues with multi-scale context at the convolutional module level, using attention to adaptively focus on key regions. PConv [50] replaces traditional CNN backbone network layers with Pinwheel-shaped convolutions, better fitting the Gaussian spatial distribution characteristics of infrared small targets.

ALCNet [51] embeds the concept of local contrast in a learnable manner, achieving attention modulation and long-range context interaction. RPCANet [52] incorporates the traditional low-rank and sparse decomposition paradigm into a network-based and interpretable framework. By employing adaptive region growing and threshold analysis,

RPCANet effectively suppresses complex background clutter. PBT [53], with its progressive background-aware approach, separates target candidate responses from background context during encoding and progressively corrects them during decoding, representing a typical approach where background priors constrain global interactions to reduce false responses.

To address the inherent trade-off between high-resolution localisation required by small targets and semantic denoising relying on deep features, recent studies have focused on more refined multi-scale fusion strategies. GANet [54] exploits Transformer-based attention and adaptive fusion modules to model long-range dependencies between small targets and background regions, and further alleviates sample imbalance through data augmentation. HCF-Net [55] enhances multi-scale representations by employing multi-branch attention mechanisms and selective fusion modules. STASPPNet [56] introduces a Swin Transformer combined with multi-scale dilated pyramid pooling to improve scale coverage. ST-Trans [57] jointly models spatial features and temporal associations, and also releases a sequence-based infrared small target dataset to facilitate further research.

In summary, existing YOLO-based methods mainly emphasise lightweight architectures and multi-scale fusion, whereas DETR- and Transformer-based approaches focus on object queries and global modelling. However, for infrared low-altitude UAV small target detection, several limitations remain. Weak target representations are often diluted by successive downsampling, leading to unstable recall. Global interactions in cluttered backgrounds tend to be dominated by background statistics, resulting in false responses. Multi-scale fusion may introduce feature misalignment and noise propagation, causing localisation drift. Moreover, the conflict between high-resolution feature demands and attention computation cost makes it difficult to achieve a favorable balance between accuracy and efficiency. Motivated by these challenges, this work proposes DCM-DETR, which improves the robustness and deployability of infrared small UAV detection through directional receptive field enhancement, saliency recalibration, gated alignment fusion, and magnitude-aware linear attention.

3. Methodologies

3.1. Overall Structure

This work adopts RT-DETR as the baseline framework. RT-DETR is an end-to-end object detection approach derived from the DETR paradigm, which directly produces detection results through set-based prediction and object query mechanisms, thereby eliminating the need for conventional non-maximum suppression (NMS). This design effectively reduces inference overhead and provides an advantage in latency control compared with CNN-based detectors such as YOLO. The RT-DETR framework comprises an efficient backbone network, a hybrid encoder, and a query selection module, enabling strong feature representation and competitive detection accuracy while maintaining real-time performance [58].

To address the inadequate representation of infrared small UAV targets under weak contrast, strong clutter, and feature dilution, DCM-DETR introduces directional receptive fields, linear attention, and magnitude-aware mechanisms into feature extraction, global encoding, and cross-scale fusion, as illustrated in Figure 1. Specifically, the input infrared image is first processed by the DRFE backbone, which consists of convolutional layers and C2f-APC units to generate multi-scale features (P1–P5) under progressive downsampling. Through asymmetric padding convolution, C2f-APC introduces directional displacement views while preserving spatial resolution, thereby enhancing target–background separability and suppressing structured clutter. The backbone features are then fed into the encoder. High-level features from P5 are processed by MALA-AIFI to model global interactions with linear complexity, where magnitude correction enhances sensitivity to weak target

responses. Meanwhile, intermediate features from P3, P4, and S5 are refined by stacked CSA-Blocks to enhance the spatial coherence and saliency of small-target regions. In addition, mid and high-level features are progressively fused through HGF and RA, which adaptively modulate low-level details under high-level semantic guidance and enforce scale-consistent feature alignment. Through this design, the encoder preserves fine-grained small target structures while maintaining global contextual information for effective clutter suppression. Finally, the fused features are fed into the decoder and head to produce end-to-end regression and classification predictions.

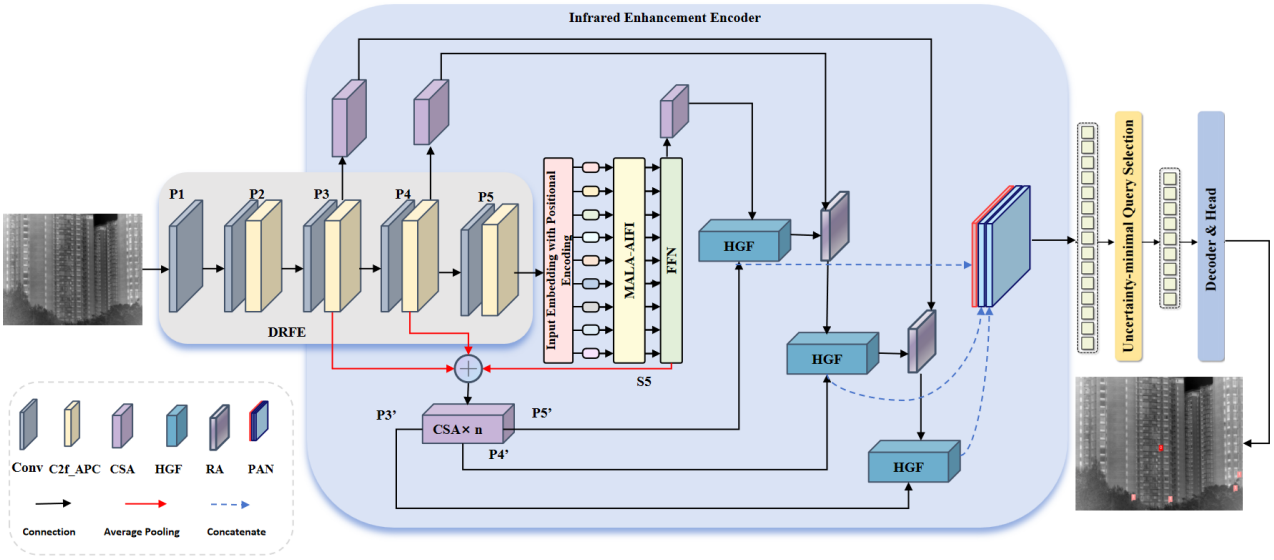


Figure 1. The overall architecture of the proposed DCM-DETR. The DRFE backbone extracts multi-scale features (P1–P5) and incorporates MALA-AIFI, CSA-Blocks, and HGF/RA cross-layer fusion during the encoding phase to enhance weak small target representation. Finally, end-to-end regression and classification predictions are performed through the decoder.

3.2. Directional Receptive-Field Enhancement (DRFE) Backbone

3.2.1. Backbone Architecture

In infrared small UAV target detection, the discriminative cues for small targets heavily rely on local weak brightness peaks and neighborhood differences. However, conventional square convolutions are easily interfered with by background spikes in strong, cluttered environments. On the other hand, stacking convolutions or increasing downsampling to obtain a larger receptive field further reduces resolution, causing the small targets to be averaged out in deep features, resulting in both false negatives and false positives. To address these issues, we propose the DRFE backbone, as shown in Figure 2. The backbone is constructed with C2f-APC as the core, where multi-layer feature extraction is achieved by alternating Conv and C2f-APC modules at each stage. Specifically, the APC employs four types of asymmetric padding to generate directional displacement perspectives. After feature extraction using shared 3 × 3 convolutions, the features are concatenated and fused using grouped 3 × 3 convolutions. This approach explicitly introduces directional global context and enriches the effective receptive field, without significantly increasing downsampling, thereby preserving small target cues and improving detection robustness.

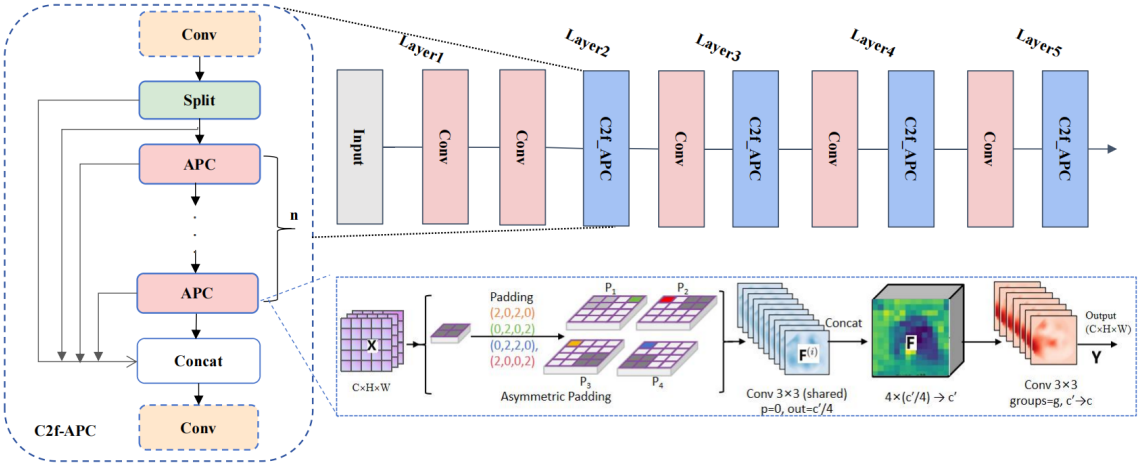


Figure 2. Overall architecture of DRFE. The backbone is composed of alternating Conv and C2f-APC modules. C2f-APC, with APC as the core, generates displacement perspectives through four-direction asymmetric padding. The features are extracted via shared 3 × 3 convolutions, concatenated, and fused, enhancing the directional receptive field to improve the representation of infrared small UAV targets.

The C2f-APC module in this paper is built on the C2f architecture, enhancing feature representation with low computational cost through branching, stacking, and aggregation. APC replaces the bottleneck unit in the original C2f to improve global modelling capability for infrared small targets. Unlike directional convolutions that modify kernel shapes (e.g., strip/anisotropic kernels) or expand receptive fields via aggressive downsampling, APC constructs multi-directional displacement views using four asymmetric padding patterns and applies shared 3 × 3 convolutions. Directional context is introduced by spatial shifting rather than direction-specific kernels, enlarging the effective receptive field with negligible parameter growth and without reducing feature-map resolution, which helps preserve weak local peaks of extremely small infrared targets.

3.2.2. Asymmetric Padding Convolution (APC)

To enhance the separability of infrared small UAV targets in complex backgrounds, we propose Asymmetric Padding Convolution (APC), as shown in the lower right of Figure 2. The core idea is to expand the effective receptive field and inject global context through directional displacement perspectives, without introducing additional downsampling. Unlike conventional 3 × 3 convolutions with symmetric padding, APC first applies four types of asymmetric padding to the input features $X \in \mathbb{R}^{B\times C\times H\times W}$, creating four spatially offset feature views:

$$X^{(i)} = P_i(X), \quad i \in \{1,2,3,4\}.$$

here $P_i(\cdot)$ represents four padding modes (left, right, top, bottom), which generate complementary displacement receptive fields in the horizontal and vertical directions, thus explicitly encoding directional information. Then, for each view, a shared-weight 3 × 3 convolution is applied to perform local feature extraction and channel compression:

$$\mathbf{F}^{(i)} = \mathrm{Conv}^{\mathrm{sh}}_{3\times 3}\left(\mathbf{X}^{(i)}\right), \quad \mathbf{F}^{(i)} \in \mathbb{R}^{B\times \frac{c}{4}\times H\times W}.$$

The shared convolution ensures that the different displacement branches are aligned in the same feature space, avoiding representation drift caused by independent branch

parameters. Then, the four branches are concatenated along the channel dimension to aggregate the directional context.

$$\mathbf{F} = \mathrm{Concat}\Big(\mathbf{F}^{(1)}, \mathbf{F}^{(2)}, \mathbf{F}^{(3)}, \mathbf{F}^{(4)}\Big), \quad \mathbf{F} \in \mathbb{R}^{B\times c'\times H\times W}.$$

Finally, a convolution of groups 3 × 3 is used to perform cross-branch fusion and channel reorganisation, producing enhanced features.

$$\mathbf{Y} = \mathrm{Conv}^{g}_{3\times 3}(\mathbf{F}), \quad \mathbf{Y} \in \mathbb{R}^{B\times C\times H\times W}.$$

where g is the number of groups, used to control the computational cost while maintaining expressive capability.

The advantage of APC lies in its asymmetric padding, which can be equivalently seen as introducing multi-directional displacement sampling on the feature plane. This allows a single convolution layer to aggregate a broader range of directional context, effectively suppressing strip textures and spike noise in infrared backgrounds, while enhancing the separability of weak targets. At the same time, APC achieves complementary integration of multi-directional information through shared convolutions, channel reduction, and grouped fusion, completing directional enhancement with minimal additional parameters and computational cost. Thus, it combines both lightweight design and improved expressive power, further increasing the consistency of small target responses and localisation stability.

3.3. Infrared Enhancement Encoder (IEE) for Small Target

To address the bottleneck in infrared small UAV detection, where weak signals are often drowned out by background statistics during encoding, cross-scale fusion leads to misalignment between semantics and details, and background highlights cause false responses that affect query and regression stability, we propose a strategy of first suppressing clutter, then aligning and fusing, and finally enhancing stability. This results in the construction of the IEE, as shown in the blue section of Figure 1, with key components illustrated in Figure 3.

First, we introduce the CSA-Block as the core of the encoding representation mechanism. Before global interactions, feature saliency is recalibrated to shift attention aggregation from background dominance to the reinforcement of target-related differences, thereby improving the response separability for weak small targets and suppressing clutter spikes. Next, the encoder employs HGF to guide low-level details with high-level semantics, selectively injecting semantic information through gating, thus avoiding noise diffusion and smoothing of details caused by simple addition/concatenation. Simultaneously, RA explicitly aligns resolution and adapts channels during cross-scale backflow and fusion, as shown in Figure 3c, eliminating spatial misalignment and response drift caused by scale inconsistencies. Based on this design, the proposed encoder can output cleaner, aligned, and more sensitive multi-scale features for weak targets, providing stable discriminatory cues for subsequent query selection and decoder regression. This significantly enhances the robustness and localisation accuracy of infrared UAV small target detection.

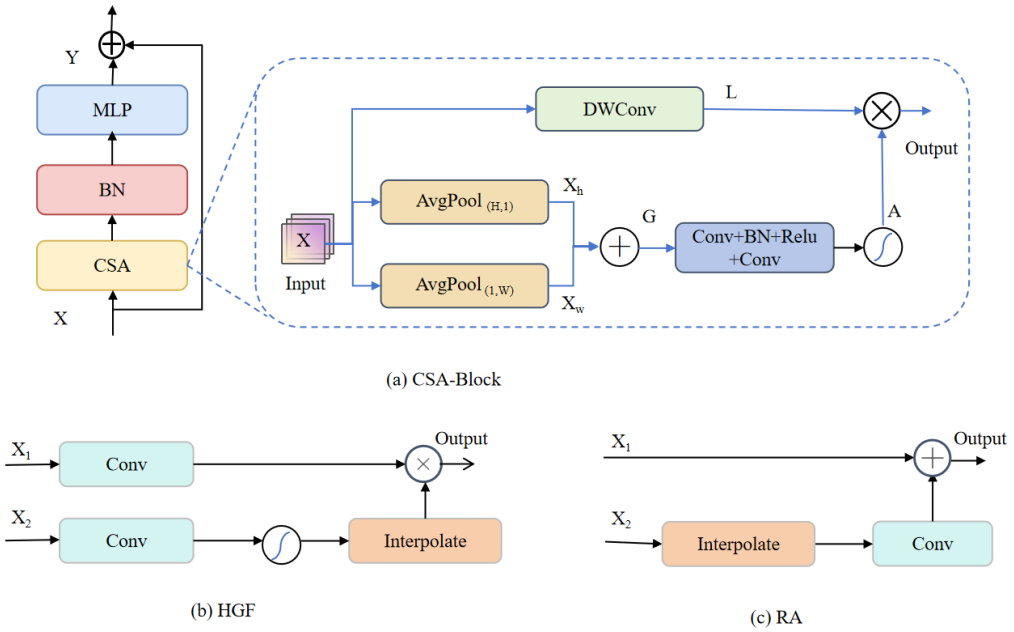


Figure 3. Diagram of the encoder's key components. (**a**) The CSA-Block is a residual module comprising CSA, BN, and an MLP. It uses directional pooling and depthwise convolution to form contextual spatial attention, and applies gated enhancement to local features. (**b**) HGF derives gating weights from high-level features, aligns them, and modulates low-level features for hierarchical gated fusion. (**c**) RA aligns and adapts cross-scale features, followed by residual addition to maintain scale consistency during fusion.

3.3.1. Context Spatial Attention Block (CSA-Block)

We design the CSA-Block, shown in Figure 3a, as a lightweight contextual constraint mechanism. It explicitly aggregates long-range context in the row and column directions to suppress structured background while preserving local details to maintain small target localisation and morphological sensitivity. Compared with Coordinate Attention (CA), which mainly encodes coordinate information for channel re-weighting, CSA generates position-wise spatial gates by fusing row/column context via pooling and broadcasting addition, and directly modulates local DWConv features to preserve tiny peak cues. Compared with CBAM, which applies sequential channel-then-spatial attention using channel-compressed pooling, CSA injects explicit directional (row/column) long-range statistics and refines them with lightweight stripe convolutions, making it more effective for suppressing structured clutter. Compared with Criss-Cross Attention (CCA), which relies on pairwise affinity-based long-range interactions, CSA avoids dense affinity computation and uses statistics-driven directional gating, achieving lower cost and reducing background-dominated global mixing. Given the input features $X \in \mathbb{R}^{B\times C\times H\times W}$, the overall mapping of CSA-Block is as follows:

$$\mathbf{Y} = \mathbf{X} + \mathbf{F}(\mathbf{X}), \quad \mathbf{F}(\mathbf{X}) = \mathrm{MLP}(\mathrm{BN}(\mathrm{CSA}(\mathbf{X}))).$$

Here, $\mathrm{CSA}(\cdot)$ denotes the contextual spatial attention operator proposed in this paper, BN stands for Batch Normalisation, and MLP implements convolutional channel mixing. CSA decomposes input features into a local representation branch and a directional context gating branch, and fuses them via gating to suppress clutter while preserving fine details. The local branch employs depthwise separable convolutions to capture local structures:

$$\mathbf{L} = \mathrm{DWConv}_{k\times k}(\mathbf{X}).$$

here, k represents the size of the square convolution kernel. In practice, we set $k = 3$ as the default in CSA-Block, consistent with our implementation. A small k provides sufficient local structural sensitivity for point-like and tiny targets while keeping the computation lightweight; increasing k may introduce over-smoothing and blur weak peak responses, which is undesirable for low-SNR infrared small targets. This branch is more sensitive to point-like and small-scale textures, which helps maintain the local saliency of extremely small targets.

The lower branch is the directional context branch, which performs adaptive average pooling along the height and width directions to obtain row and column context:

$$\mathbf{X}_h = \mathrm{AvgPool}_{(H,1)}(\mathbf{X}) \in \mathbb{R}^{B\times C\times H\times 1}, \quad \mathbf{X}_w = \mathrm{AvgPool}_{(1,W)}(\mathbf{X}) \in \mathbb{R}^{B\times C\times 1\times W}.$$

Then, the row and column context is aggregated into a 2D context feature by broadcasting addition:

$$\mathbf{G} = \mathbf{X}_h \oplus \mathbf{X}_w \in \mathbb{R}^{B\times C\times H\times W}.$$

where $\oplus$ denotes broadcasting addition. This operation injects the long-range statistics of the row and column directions into each spatial location, effectively capturing the structural variation trends of the background. Next, a lightweight activation network generates spatial weights:

$$\mathbf{A} = \sigma(\psi(\mathbf{G})), \quad \mathbf{A} \in \mathbb{R}^{B\times C\times H\times W}.$$

where $\sigma(\cdot)$ represents the Sigmoid function, $\psi(\cdot)$ consists of two stripe convolutions, which further model the contextual dependencies in the horizontal and vertical directions. In practice, $\psi(\cdot)$ is implemented as a two-layer stripe-convolution stack:

$$\psi(\mathbf{G}) = \mathrm{Conv}_{1\times k_s}(\mathrm{Conv}_{k_s\times 1}(\mathbf{G})),$$

We adopt a sigmoid-based soft gating mechanism, which performs continuous reweighting rather than hard masking, thereby suppressing structured clutter while avoiding over-suppression of extremely weak targets. In our implementation, we set $k_s = 11$ (i.e., 1×11 and 11×1 stripe kernels) to capture sufficiently long-range directional statistics with marginal overhead. Moreover, to further reduce parameters, the stripe convolutions are implemented with channel reduction ratio $r = 2$ (i.e., $g_c = C/r$) and grouped convolutions with groups $= g_c$, which substantially lowers the cost compared with standard (ungrouped) convolutions while preserving the effectiveness of directional context modelling. Finally, the CSA output is the attention-gated local features:

$$\mathrm{CSA}(\mathbf{X}) = \mathbf{A} \odot \mathbf{L}.$$

where $\odot$ represents element-wise multiplication. Intuitively, A tends to have smaller weights at background textures and spike noise to suppress responses, while boosting weights in potentially small target regions, thereby enhancing the contrast between the target and background while preserving localisation details.

CSA-Block, through complementary modelling of directional context and local details, provides more robust encoded representations for infrared small targets. The long-range statistics generated by directional pooling are especially effective in suppressing background structured noise. DWConv retains local sensitivity to avoid smoothing of small targets caused by global interactions. The residual structure, along with BN and MLP, ensures stable training and enhances channel mixing capabilities.

3.3.2. Hierarchical Gated Fusion (HGF)

The proposed HGF is shown in Figure 3b. It selectively injects cross-layer information through high-level guidance, spatial alignment, channel projection, and residual fusion, effectively alleviating the noise injection and localisation degradation issues caused by simple addition/concatenation in infrared small UAV target detection. Unlike gate-based fusion, HGF first aligns high-level features to the low-level spatial resolution by bilinear interpolation, and then performs lightweight channel adaptation using a 1×1 convolution. This design maintains feature consistency while avoiding parameter expansion and excessive noise injection, thereby producing more stable encoded representations for subsequent query selection and decoder regression.

Given two feature paths from different layers $\mathbf{X}_l \in \mathbb{R}^{B \times C_l \times H \times W}$ (low-level/high-resolution) and $\mathbf{X}_h \in \mathbb{R}^{B \times C_h \times H' \times W'}$ (high-level/low-resolution), HGF performs bilinear interpolation to resize $\mathbf{X}_h$ to (H, W) and then applies a 1×1 convolution for channel projection from C_h to C_l:

$$\mathbf{V} = \phi(\mathcal{I}(\mathbf{X}_h; H, W))$$

where $\mathcal{I}(\cdot)$ denotes bilinear interpolation with `align_corners = False`, and $\phi(\cdot)$ is a 1×1 convolution mapping. Finally, HGF aggregates the aligned features via residual addition:

$$\mathbf{Y} = \mathbf{X}_l + \mathbf{V}, \quad \mathbf{Y} \in \mathbb{R}^{B \times C_l \times H \times W}$$

In the encoder of this paper, HGF operates hierarchically on the multi-scale path. At each stage, deeper features are resized and projected using the above operator and then injected into higher-resolution branches through residual fusion, enabling progressive guidance from semantics to details. This hierarchical top-down fusion improves the response consistency of small targets and mitigates cross-scale misalignment accumulation in cluttered infrared backgrounds. This hierarchical fusion mechanism enhances the selectivity of multi-scale aggregation, strengthening target responses while suppressing background interference during cross-layer propagation.

3.4. Magnitude-Aware Linear Attention Encoder (MALA-AIFI)

Infrared small target detection requires global context to suppress clutter and provide discriminative references, while maintaining controlled computational cost. When flattening a feature map into $L = H \times W$ tokens, standard self-attention incurs a quadratic complexity of $\mathcal{O}(L^2)$, which is inefficient for high-resolution feature encoding. To address this issue, we introduce a lightweight attention layer, termed MALA-AIFI, for efficient feature encoding. It adopts linear attention to enable global information interaction, while preserving fine localisation and detail responses for small targets through two-dimensional positional modelling and local enhancement.

Given the backbone output feature $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$, MALA-AIFI follows a two-level residual Transformer-style encoding structure:

$$\mathbf{X}_1 = \mathrm{LN}\big(\mathbf{X} + \mathrm{Drop}(\mathrm{MALA}(\mathbf{X}))\big), \mathbf{Y} = \mathrm{LN}\big(\mathbf{X}_1 + \mathrm{Drop}(\mathrm{FFN}(\mathbf{X}_1))\big), \tag{1}$$

where $\mathrm{MALA}(\cdot)$ denotes the linear attention module [33], and $\mathrm{FFN}(\cdot)$ performs channel-wise mixing using 1×1 convolutions:

$$\mathrm{FFN}(\mathbf{X}_1) = \mathrm{Conv}^{(2)}_{1\times1}\Big(\mathrm{Drop}\big(\sigma(\mathrm{Conv}^{(1)}_{1\times1}(\mathbf{X}_1))\big)\Big). \tag{2}$$

here, $\sigma(\cdot)$ denotes the GELU activation function by default.

Concretely, let $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ be the projected features of $\mathbf{X}$ and $\mathbf{Z}$ denote the linear-attention output. We compute a magnitude factor from the query activation as $\boldsymbol{\alpha} = \sigma(\mathrm{MLP}(\|\mathbf{Q}\|_2))$ and apply it to rectify the attention response:

$$\mathbf{Z} = \mathrm{LinearAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}), \qquad \mathrm{MALA}(\mathbf{X}) = \boldsymbol{\alpha} \odot \mathbf{Z}, \tag{3}$$

where $\odot$ denotes element-wise multiplication. This magnitude-aware scaling adaptively boosts weak target responses while preventing strong clutter activations from dominating the global interaction.

MALA-AIFI achieves a favorable balance between efficiency and representational capacity in infrared small-target feature encoding for UAV detection. By enabling global interaction with linear complexity, it models long-range dependencies even on high-resolution features, effectively suppressing background clutter and enhancing target discriminability. In addition, a magnitude-aware correction mechanism conditioned on feature activation strength is incorporated to compensate for the inherent insensitivity of linear attention to amplitude variations, allowing weak responses from extremely small targets to be preserved more reliably.

4. Experiments and Analysis

4.1. Datasets

To evaluate the detection efficiency of the proposed method, we conducted experiments on two publicly available datasets, SIRST-UAVB and IRSTD, to assess our DCM-DETR. SIRST-UAVB [59] is a real-world infrared small target benchmark dataset from Southwest University of Science and Technology, primarily targeting aerial small objects such as UAVs and birds. The dataset contains 3000 infrared images, covering various seasons and weather conditions over a year, with complex backgrounds. It emphasises challenges such as multi-pose, multi-scale, and occlusion in the spatial domain. The dataset has a high proportion of small targets, some of which are nearly invisible to the naked eye. The annotations were manually performed based on target trajectories and verified repeatedly. The dataset includes 2955 bounding boxes for UAVs and 1742 for birds. However, due to the difficulty in accurately delineating small bird targets, mask annotations are only provided for UAVs.

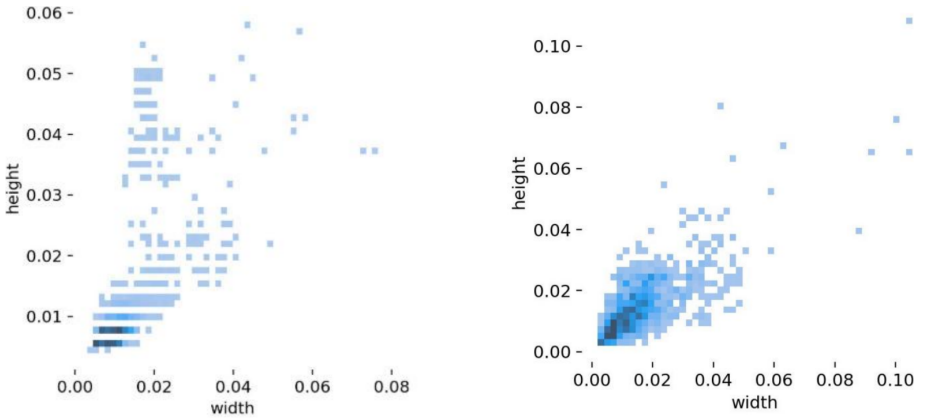
IRSTD-1K [60] is a real-world infrared small target detection benchmark from Xi'an University of Electronic Science and Technology. It contains approximately 1000 images captured by real infrared cameras and provides pixel-level manual annotations. The images have a uniform resolution of 512 × 512. The dataset covers various target types, including UAVs, animals, biological targets, ships, and vehicles. Due to the long imaging distances and small target sizes, the images are often accompanied by low contrast, noise and clutter interference. The background scenes are also diverse, including the sea, rivers, fields, mountains, urban areas, and clouds, providing a comprehensive evaluation of the robustness and generalisation ability of infrared small target detection methods. In experimental setups, the dataset is typically split into training, validation, and test sets in a 7:1:2.

IRSTD-1K [60] is a real-world infrared small target detection benchmark from Xi'an University of Electronic Science and Technology. It contains approximately 1000 images captured by real infrared cameras and provides pixel-level manual annotations. Although IRSTD-1K provides pixel-level masks, our detection experiments are conducted based on bounding boxes. Therefore, during preprocessing we convert each binary mask into an axis-aligned tight bounding box by computing the minimal enclosing rectangle of all foreground pixels: $x_{\min} = \min(x)$, $y_{\min} = \min(y)$, $x_{\max} = \max(x)$, $y_{\max} = \max(y)$, yielding the box

$[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ in pixel coordinates. If multiple targets exist in one image, we compute the bounding box for each connected component separately.

The images have a uniform resolution of 512×512. The dataset covers various target types, including UAVs, animals, biological targets, ships, and vehicles. Due to the long imaging distances and small target sizes, the images are often accompanied by low contrast, noise and clutter interference. The background scenes are also diverse, including the sea, rivers, fields, mountains, urban areas, and clouds, providing a comprehensive evaluation of the robustness and generalisation ability of infrared small target detection methods. In experimental setups, the dataset is typically split into training, validation, and test sets in a 7:1:2.

From the normalised target bounding box width and height distribution in Figure 4, it can be observed that the vast majority of target samples in both the SIRST-UAVB and IRSTD datasets are concentrated in the very small region near the origin, with only a few outliers representing medium-sized targets. This demonstrates a strong scale imbalance and long-tail distribution. Extremely small targets are prone to attenuation or even disappearance during feature downsampling, resulting in insufficient separability. Additionally, infrared targets are often weak contrast bright spots, subject to interference from complex backgrounds and noise, which leads to an increased number of false detections and causes minor localisation errors to result in significant IoU fluctuations, thus making detection and regression learning more difficult.



(a) SIRST-UVAB (b) IRSTD

Figure 4. Distribution of bounding box normalised sizes for the SIRST-UAV and IRSTD datasets. The color intensity represents the sample density, where darker blue indicates a higher frequency of bounding boxes.

4.2. Evaluation Metrics and Experimental Setup

4.2.1. Evaluation Metrics

In object detection evaluation, Precision (P), Recall (R), Average Precision (AP), and mean AP (mAP) are used as the primary accuracy metrics. Based on the matching relationship between predicted and ground truth boxes, the true positives (TP), false positives (FP), and false negatives (FN) are counted [61]. Precision measures the proportion of predicted samples that are true targets, defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

Recall measures the proportion of true targets that are detected, defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

With the variation in the confidence threshold, the P–R curve can be obtained, and AP is the area under this curve, used to comprehensively assess detection performance at different thresholds. The mAP is obtained by averaging the AP across all classes. In this paper, mAP_{50} refers to the mean average precision at an IoU threshold of 0.5, and $mAP_{50:95}$ represents the average precision over the IoU range from 0.5 to 0.95. The calculation of average precision is as follows:

$$\begin{aligned} \mathrm{AP} &= \int_0^1 \mathrm{Precision}(t)\,\mathrm{d}t \\ \mathrm{mAP} &= \frac{1}{N}\sum_{n=1}^{N} \mathrm{AP}_n \end{aligned} \tag{6}$$

Additionally, we also evaluate model complexity and inference efficiency to measure deployment feasibility. Parameter count (#P) represents the number of trainable parameters in a model, with the unit M denoting millions, meaning the model has a number of trainable parameters in the millions. Size refers to the space occupied by the model in storage; GFLOPs indicates the number of floating-point operations required for one forward pass, used to characterise computational complexity; FPS (Frames Per Second) is the end-to-end throughput under fixed hardware and inference settings, used to assess real-time performance. By considering #P, Size, GFLOPs, and FPS together, we can more comprehensively compare the trade-offs between accuracy, efficiency, and resource overhead, providing a reference for resource-constrained scenarios, such as UAV platforms.

4.2.2. Experimental Setup

The experimental environment is based on Red Hat 4.8.5-28 OS (Red Hat, Raleigh, NC, USA) with PyTorch 2.1 (PyTorch, Santa Clara, CA, USA), Python 3.11 (Python Software Foundation, Wilmington, DE, USA), and CUDA 12.1 (NVIDIA, Santa Clara, CA, USA), and all experiments are conducted on a cluster equipped with two 80 GB NVIDIA A800 GPUs (NVIDIA, Santa Clara, CA, USA). We train the model for 300 epochs with a batch size of 4, using 4 dataloader workers and an input resolution of 640 × 640. The dataset is split into training/validation/testing sets with a 7:1:2 ratio. For reproducibility, we fix the random seed to 0 and enable deterministic training. During inference, we use the default confidence thresholds of 0.001 for validation and 0.25 for prediction, and we do not use half-precision inference. All ablation experiments are performed under the same fixed seed, with three trials for each configuration, and we report the average results. To improve generalisation, we apply data augmentation including random cropping, flipping, and scaling. For comparison, we adopt RT-DETR with a ResNet-18 backbone. The remaining hyperparameters are summarised in Table 1.

Table 1. Training Hyperparameters.

Feature	Parameter
Optimiser	AdamW
Base learning rate	0.0001
Weight decay	0.0001
Global gradient clip norm	0.1
Linear warmup steps	2000
Minimum learning rate	0.00001

4.3. Ablation Study

On the SIRST-UAVB dataset, Table 2 reports the ablation results of the proposed DRFE, IEE, and MALA-AIFI modules in terms of accuracy and efficiency. The mAP_{50} values

represent the average of three independent runs, with a fixed random seed and consistent experimental conditions. The standard deviation across these runs is ±0.1, all ablation experiments follow this procedure.

The RT-DETR baseline achieves an mAP_{50} of 88.3% with 20.1 M and 58.3 GFLOPs. Replacing the baseline with DRFE improves mAP_{50} to 89.5% while substantially reducing model complexity to 13.5 M and 48.3 GFLOPs. Introducing IEE alone yields 89.8% mAP_{50} with 19.3 M and 49.0 GFLOPs, adding MALA-AIFI alone leads to a modest gain (89.1% mAP_{50}) but slightly increases computation to 58.9 GFLOPs with 20.2 M. When DRFE and IEE are combined, performance further rises to 90.4% mAP_{50} with the lowest parameter count (13.4 M) and 44.4 GFLOPs, confirming their complementarity. Finally, integrating all components achieves the best overall trade-off, reaching 90.6% mAP_{50} with 13.6 M and 45.0 GFLOPs, validating that these modules cooperate to produce stronger and more efficient representations for infrared small UAV detection.

Table 2. Ablation study on SIRST-UAVD.

Method	#P (M)	GFLOPs	mAP_{50}
Baseline (RT-DETR)	20.1	58.3	88.3
Baseline + DRFE	13.5	48.3	89.5
Baseline + IEE	19.3	49.0	89.8
Baseline + MALA-AIFI	20.2	58.9	89.1
Baseline + DRFE + IEE	13.4	44.4	90.4
Baseline + DRFE + MALA-AIFI	13.6	48.5	90.1
Baseline + DRFE + IEE + MALA-AIFI	13.6	45.0	90.6

On the IRSTD dataset, the ablation experiments showed the same trend as those on the SIRST-UAVB dataset, as shown in Table 3. After progressively introducing DRFE, IEE, and MALA-AIFI, the overall performance continued to improve, and the model became more lightweight. Our model finally achieved a detection accuracy of 86.4%, a 3.7% improvement over the baseline model, with the model size reduced by 32.8%.

Table 3. Ablation study results on the IRSTD dataset.

Method	#P (M)	GFLOPs	mAP_{50} (%)
Baseline	20.1	58.3	82.7
Baseline + DRFE	13.5	48.3	83.9
Baseline + DRFE + IEE	13.4	44.4	85.2
Baseline + DRFE + IEE + MALA-AIFI	13.5	44.7	86.4

To verify the effectiveness of the proposed C2f-APC module, we conduct an ablation study. As reported in Table 4, progressively extending C2f-APC from higher layers to additional stages yields a steady performance improvement while simultaneously reducing computational cost. The best trade-off is achieved by Scheme 4, where C2f-APC is deployed in all stages from Stage 2 to Stage 5, reaching 89.5% mAP_{50} with the lowest complexity of 13.5 M and 48.3 GFLOPs. This is mainly because C2f-APC generates multi-directional displaced feature views via asymmetric padding and fuses them using shared convolutions, thereby injecting directional context without additional downsampling. The shallower stages (Stage 2/3) emphasise preserving and enhancing fine-grained details and boundaries of small targets while suppressing local noise, whereas the deeper stages (Stage 4/5) provide stronger semantic discrimination and global background suppression. When enabled across all four stages, the directional enhancement becomes complementary across

scales, leading to more stable representations from low-level details to high-level semantics, and consequently, the best overall performance.

Table 4. Ablation study of deploying C2f-APC at different stages on IRSTD. ○ indicates that C2f-APC is applied at the corresponding stage (blank means not applied).

No.	Stage 5	Stage 4	Stage 3	Stage 2	#P (M)	GFLOPs	mAP_{50}
1	○				14.1	52.1	88.1
2	○	○			13.7	51.0	88.4
3	○	○	○		13.6	49.2	89.1
4	○	○	○	○	13.5	48.3	89.5

Table 5 summarises the ablation study on the IRSTD dataset for the IEE architecture, assessing the effects of CSA-Block, HGF, and RA. The baseline model (without any module) achieves 82.7% mAP_{50} with 20.1 M parameters and 58.3 GFLOPs. Enabling CSA-Block alone improves performance to 84.2% mAP_{50}, while incorporating HGF or RA individually yields 84.1% and 83.5% mAP_{50}, respectively, indicating that each component contributes positively to infrared small-target detection under the IEE design. Notably, activating all three modules jointly attains the best result of 84.8% mAP_{50}, corresponding to a +2.1% gain over the baseline, while reducing the parameter count to 19.3 M and the computational cost to 49.0 GFLOPs. These results validate the complementary benefits of CSA-Block, HGF, and RA, and demonstrate that the IEE design improves accuracy with improved computational efficiency.

Table 5. Ablation study of CSA-Block, HGF, and RA in the IEE architecture on the IRSTD dataset. ✓ indicates that the corresponding module is enabled, and × indicates that it is disabled.

CSA-Block	HGF	RA	#P (M)	GFLOPs	mAP_{50} (%)
×	×	×	20.1	58.3	82.7
✓	×	×	20.5	58.2	84.2
×	✓	×	19.9	57.3	84.1
×	×	✓	20.2	59.2	83.5
✓	✓	✓	19.3	49.0	84.8

To verify the performance of the proposed DRFE backbone, we compared the feature maps generated by the baseline model and the DCM-DETR backbone on two datasets, as shown in Figure 5. The first row represents SIRST-UAVB, and the second row represents IRSTD. Figure 5a shows the ground truth of the original image, and Figure 5b presents the feature maps generated by RT-DETR. The RT-DETR feature maps exhibit clear signs of channel deactivation and background dominance. Some channels have responses close to zero, while the high-response regions of the remaining channels are often concentrated at building edges, texture structures, or clutter areas. The point-like energy corresponding to small targets is easily averaged out or overwhelmed by background statistical consistency during downsampling and global interactions, leading to numerous discrete false responses and unstable activations.

In contrast, the feature maps generated by DCM-DETR, shown in Figure 5c, display a more continuous, clean, and consistent saliency distribution. The target neighborhoods form more concentrated and continuous high-response regions, and the false activations of background structural textures are effectively suppressed. This results in clearer target-background separability and more reliable discriminative representations under strong

clutter and weak contrast conditions, providing a more robust feature foundation for the subsequent detection head.

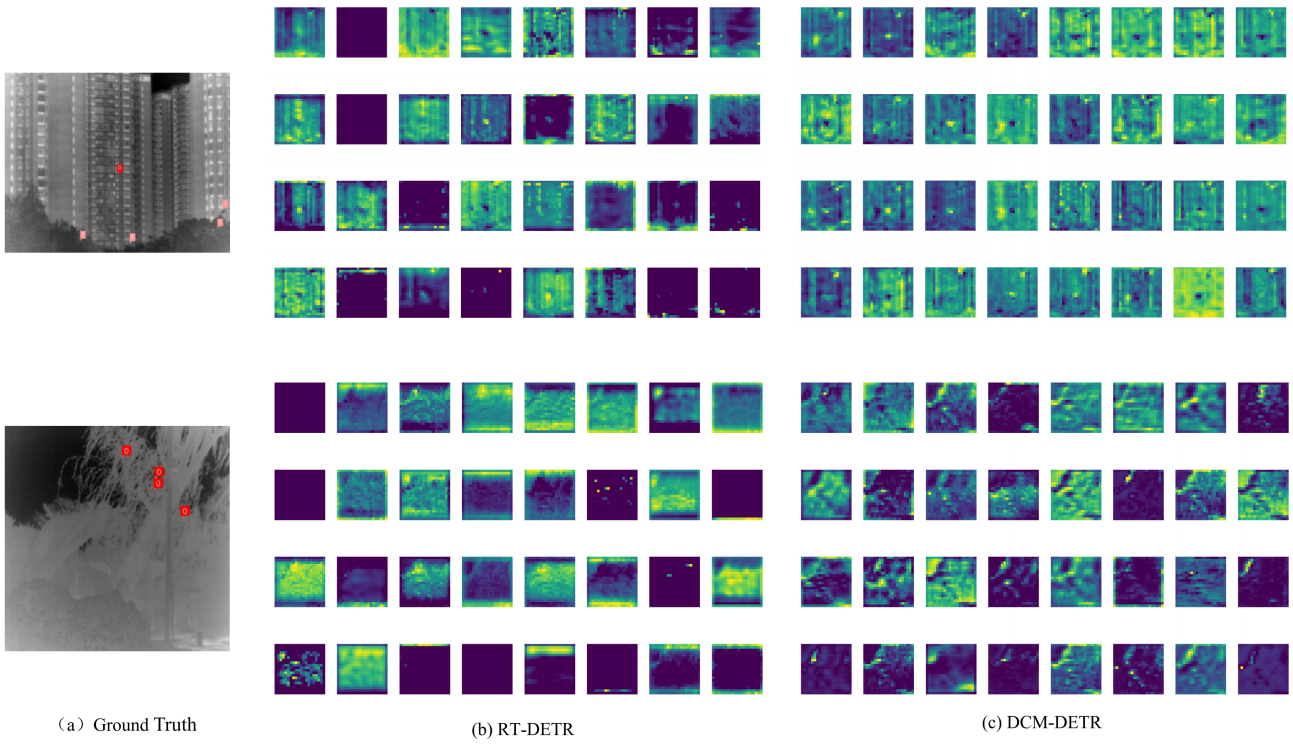


Figure 5. Feature map comparison on two datasets. The first row represents the SIRST-UAVB dataset, and the second row represents the IRSTD dataset. (**a**) Original images and Ground Truth of the targets. (**b**) Feature maps generated by the baseline model. (**c**) Feature maps generated by the proposed DCM-DETR.

4.4. Comparison Experiments

The proposed method was benchmarked against representative detectors on two public datasets. Table 6 reports the results on SIRST-UAVB. Overall, lightweight YOLO-family models struggle in this small-target setting. For instance, YOLOv8n achieves only 50.77% mAP_{50} and 24.33% $mAP_{50:95}$ over all classes. This performance drop is mainly due to the fact that point-like targets are easily attenuated by repeated downsampling, while local convolutional matching is prone to being distracted by cluttered high-frequency structures and thermal noise peaks, which in turn, increases false alarms. Note that YOLO baselines are NMS-based; thus, we follow the Ultralytics default NMS settings for fair and reproducible evaluation.

Among infrared-oriented baselines, ALCNet attains 84.2% mAP_{50} and 41.3% $mAP_{50:95}$, LE-IRSTD performs worse with 72.2% and 40.1%, while EFLNet is slightly stronger than RT-DETR on the overall split, reaching 87.31% mAP_{50} and 44.12% $mAP_{50:95}$.

In contrast, the end-to-end RT-DETR yields a substantial improvement, reaching 86.26% mAP_{50} and 43.39% $mAP_{50:95}$. This suggests that query-based set prediction, together with global context modelling, is more effective at separating weak target cues from complex backgrounds. Building upon RT-DETR, DCM-DETR further improves accuracy to 87.35% mAP_{50} and 44.24% $mAP_{50:95}$. The gain is particularly evident for the challenging Bird category, where mAP_{50} increases from 77.52% to 79.39% and $mAP_{50:95}$ rises from 35.50% to 36.15%. These results indicate that directional receptive-field enhancement and saliency recalibration reduce background-induced responses and improve the discriminability of small targets.

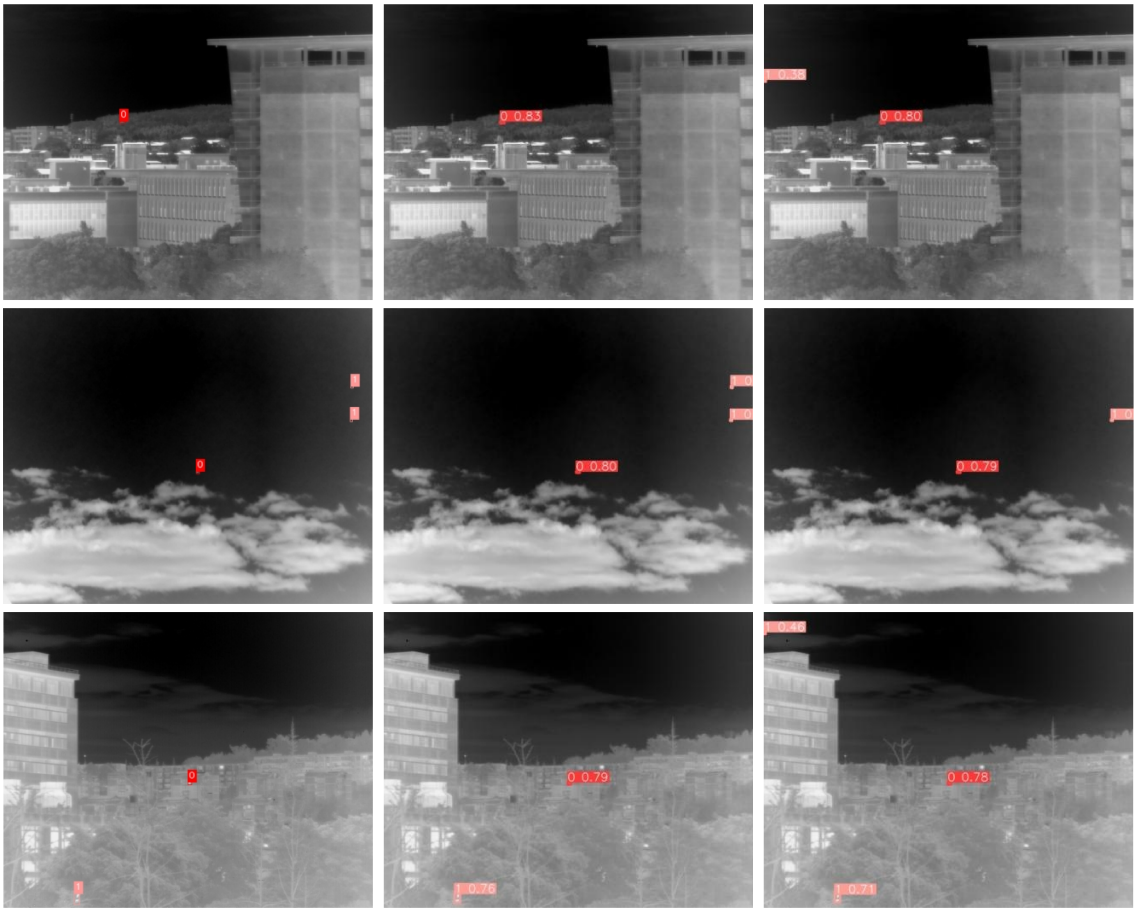
In addition, hierarchical gated fusion and residual alignment mitigate localisation drift introduced by cross-scale mismatch, leading to a more consistent improvement under stricter IoU criteria, as reflected by the steady increase in $mAP_{50:95}$. Notably, the accuracy gains come with lower complexity. Compared with RT-DETR, DCM-DETR reduces the parameter count and computation from 77.0 M and 56.9 GFLOPs to 51.9 M and 44.2 GFLOPs. Meanwhile, the model size and computational cost are both reduced relative to the baseline, further supporting the lightweight design objective. This demonstrates that more targeted feature encoding and selective fusion can simultaneously enhance detection performance and efficiency by avoiding redundant computation.

Table 6. Comparison experiments of different algorithms on the SIRST-UAVB dataset.

Method	Category	P	R	mAP_{50}	$mAP_{50:95}$	Size	GFLOPs
YOLOv8n	ALL	77.11	41.44	50.77	24.33	5.4	6.8
	UAV	84.85	51.18	63.85	31.75		
	Bird	69.36	31.70	37.70	16.91		
YOLOv10n [24]	ALL	0.7408	0.4185	0.5084	0.2406	5.5	6.5
	UAV	0.8105	0.5282	0.6407	0.3164		
	Bird	0.6712	0.3088	0.3762	0.1647		
YOLO11n	ALL	0.7667	0.3937	0.4910	0.2296	5.2	6.3
	UAV	0.8233	0.5045	0.6265	0.3066		
	Bird	0.7102	0.2829	0.3556	0.1527		
YOLOv13n [62]	ALL	76.25	49.58	58.06	27.62	5.2	6.1
	UAV	76.25	49.58	58.06	27.62		
	Bird	67.36	28.33	46.31	21.23		
ALCNet [51]	ALL	85.50	82.20	84.20	41.30	18.2	34.3
LE-IRSTD [63]	ALL	83.30	77.40	72.20	40.10	10.3	7.4
EFLNet [64]	ALL	87.60	82.30	87.31	44.12	37.3	102.1
RT-DETR [29]	ALL	91.40	84.55	86.26	43.39	77.0	56.9
	UAV	96.61	94.59	94.99	51.27		
	Bird	86.19	74.51	77.52	35.50		
DCM-DETR	ALL	94.18	84.54	87.35	44.24	51.9	44.2
	UAV	96.26	93.74	95.32	52.33		
	Bird	92.09	75.35	79.39	36.15		

Figure 6 presents a qualitative comparison of the SIRST-UAVB dataset. Subfigure (a) shows the ground truth, (b) shows the proposed method, and (c) shows RT-DETR. Three representative interference conditions are included. The first row contains strong edges and layered building textures. The second row contains cloud-induced bright bands with a large-scale high-gradient background. The third row shows an urban scene with mixed vegetation and building clutter. Across all cases, the proposed method yields more concentrated responses around the target, and the predicted bounding boxes align more closely with the ground truth. It also produces fewer background-triggered candidates, indicating stronger suppression of

clutter-induced activations. In comparison, RT-DETR is more likely to introduce additional low-confidence false positives near image boundaries, prominent structural contours, or bright bands, and its predictions exhibit occasional localisation drift.



（a）Ground Truth (b) DCM-DETR (c) RT-DETR

Figure 6. Qualitative comparison of detection results on the SIRST-UAVB dataset. From left to right: (**a**) ground-truth annotations, (**b**) our method, and (**c**) RT-DETR. Red and pink boxes indicate predicted targets with confidence scores.

Table 7 further validates these observations on the IRSTD dataset, where our method achieves a favourable balance between detection performance and inference efficiency. It obtains 84.85% precision and 84.56% recall, with mAP_{50} and $mAP_{50:95}$ of 83.42% and 39.48%, respectively. The mAP_{50} is the highest among all compared methods, exceeding Faster R-CNN by 22.12 percentage points and Cascade R-CNN by 7.22 points. It also outperforms lightweight YOLO variants by clear margins, including YOLOv8n, YOLOv10n, and YOLOv13n, and improves upon RT-DETR-r18 by 2.01%.

In terms of throughput, although lightweight YOLO models deliver higher frame rates, their accuracy remains substantially lower. Two-stage detectors can be competitive in either precision or recall, but they incur higher latency and larger model sizes. Notably, our method reduces computation to 44.2 GFLOPs and reaches 93.45 FPS, which is nearly twice the speed of RT-DETR-r18, while simultaneously improving accuracy; it is sufficient to meet the real-time requirements of edge-device deployment. Overall, these results confirm that the proposed feature enhancement and efficient encoding design improve small-target separability and stabilise recall without sacrificing the benefits of end-to-end detection, leading to a more favorable accuracy–efficiency trade-off.

Table 7. Comparison of different algorithms on the IRSTD dataset.

Method	P	R	mAP_{50}	$mAP_{50:95}$	#P (M)	FPS
Faster-RCNN [17]	75.0	51.7	61.3	27.8	165.6	59.7
Cascade-RCNN [65]	86.5	64.3	76.2	41.0	277.2	39.6
YOLOv8n	77.70	74.81	79.08	38.16	5.3	184.04
YOLOv10n [24]	78.65	74.81	79.08	38.16	5.4	245.34
YOLO11n	79.55	72.48	81.02	39.02	5.3	184.04
YOLOv13n [62]	78.15	71.14	75.13	36.37	5.2	160.20
RT-DETR-r18 [29]	83.95	82.55	81.41	39.31	77.0	47.43
DCM-DETR	84.85	84.56	83.42	39.48	51.9	93.45

To visually compare the performance of various algorithms, Figure 7 shows the performance of multiple detection methods on the SIRST-UAVB and IRSTD datasets. The bar chart allows for a clear comparison of the performance differences among the algorithms. DCM-DETR performs excellently on both datasets, especially in mAP and recall, outperforming traditional YOLO series methods and other baseline methods.

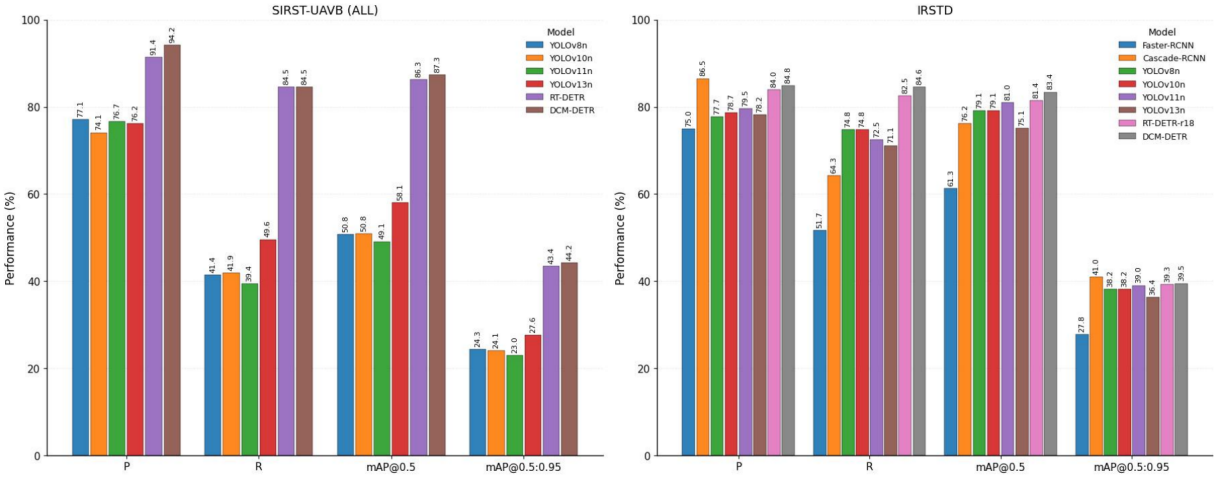


Figure 7. Performance comparison based on different detection methods on the SIRST-UAVB and IRSTD datasets.

As shown in Figure 8, DCM-DETR achieves a more favorable efficiency profile than RT-DETR on the IRSTD dataset. Specifically, DCM-DETR reduces the model size from 77.0 MB to 51.9 MB and decreases the computational cost from 56.9 to 44.2 GFLOPs, indicating a lighter model footprint and lower computation demand. Meanwhile, DCM-DETR substantially improves inference speed, increasing the throughput from 47.43 FPS to 93.45 FPS. Overall, the results demonstrate that DCM-DETR provides a better accuracy–efficiency trade-off and supports the lightweight and real-time design claim under the same evaluation setting.

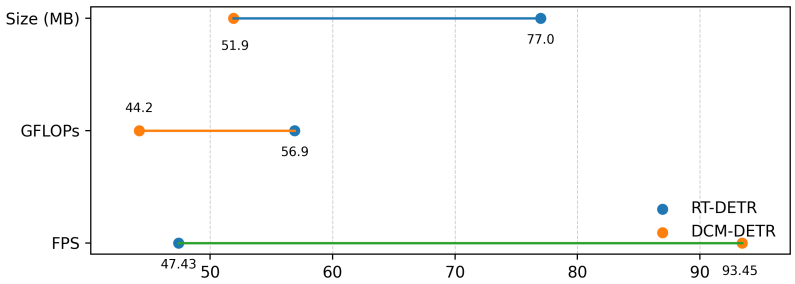


Figure 8. Efficiency comparison between Baseline and DCM-DETR on the IRSTD dataset in terms of model size, computational cost (GFLOPs), and inference speed (FPS).

Figure 9 presents response distributions for three representative scenes from the IRSTD dataset. The baseline produces saliency maps with dispersed and spatially shifted high-response regions in areas dominated by strong structures or non-uniform radiation, as highlighted by the dashed circles. Specifically, prominent non-target activations appear along tree-canopy boundaries, near image borders, and around water surfaces or shorelines. In these high-gradient regions, background peaks often dominate the weak target energy, leading to diluted target responses and degraded spatial consistency.

In comparison, DCM-DETR yields heatmaps with clearer target emphasis and stronger background suppression. Responses in the target neighborhood form more compact, higher-contrast local peaks, while background activations are smoother and lower in magnitude, with a marked reduction in spurious responses. These observations suggest that the proposed feature modelling better preserves the subtle local energy peaks associated with weak small targets, while restraining exaggerated responses to structural textures and intensity variations induced by global interactions. Consequently, it improves target–background separability and stabilises the spatial distribution of responses.

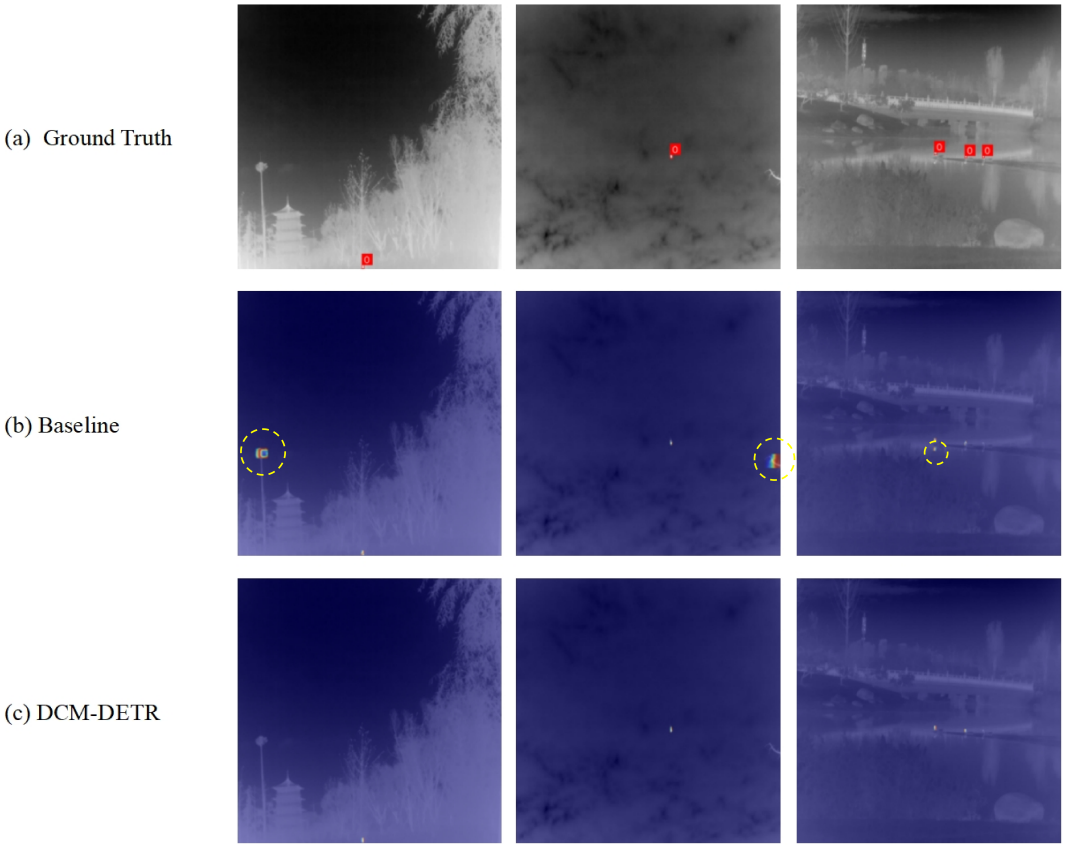


Figure 9. Visual comparison on challenging IRSTD scenes. Compared with the baseline, DCM-DETR detects small targets more reliably and produces fewer background-induced false alarms. Yellow dashed circles denote baseline failure regions.

Figure 10 presents a visual comparison of several typical complex background scenarios. (a) represents the ground truth, and (b)–(d) show the predictions from different methods. In the first row, with a strong gradient background at the water surface and tree line intersection, and in the second row, where structural clutter is caused by reflections of the bridge and water surface, DCM-DETR can more stably cover multiple targets, with predicted bounding boxes more consistent with the ground truth. In contrast, RT-DETR tends to miss or falsely detect some targets in the same scenario, while YOLO11n results in insufficient multi-target recall. In the third row, where there is a large smooth background

but with strong bright objects and edge interference locally, and in the fourth row, where high-reflective surfaces and vegetation boundaries form a strong contrast, DCM-DETR still accurately locates the targets, whereas comparison methods are more likely to experience box center drift or even miss targets.

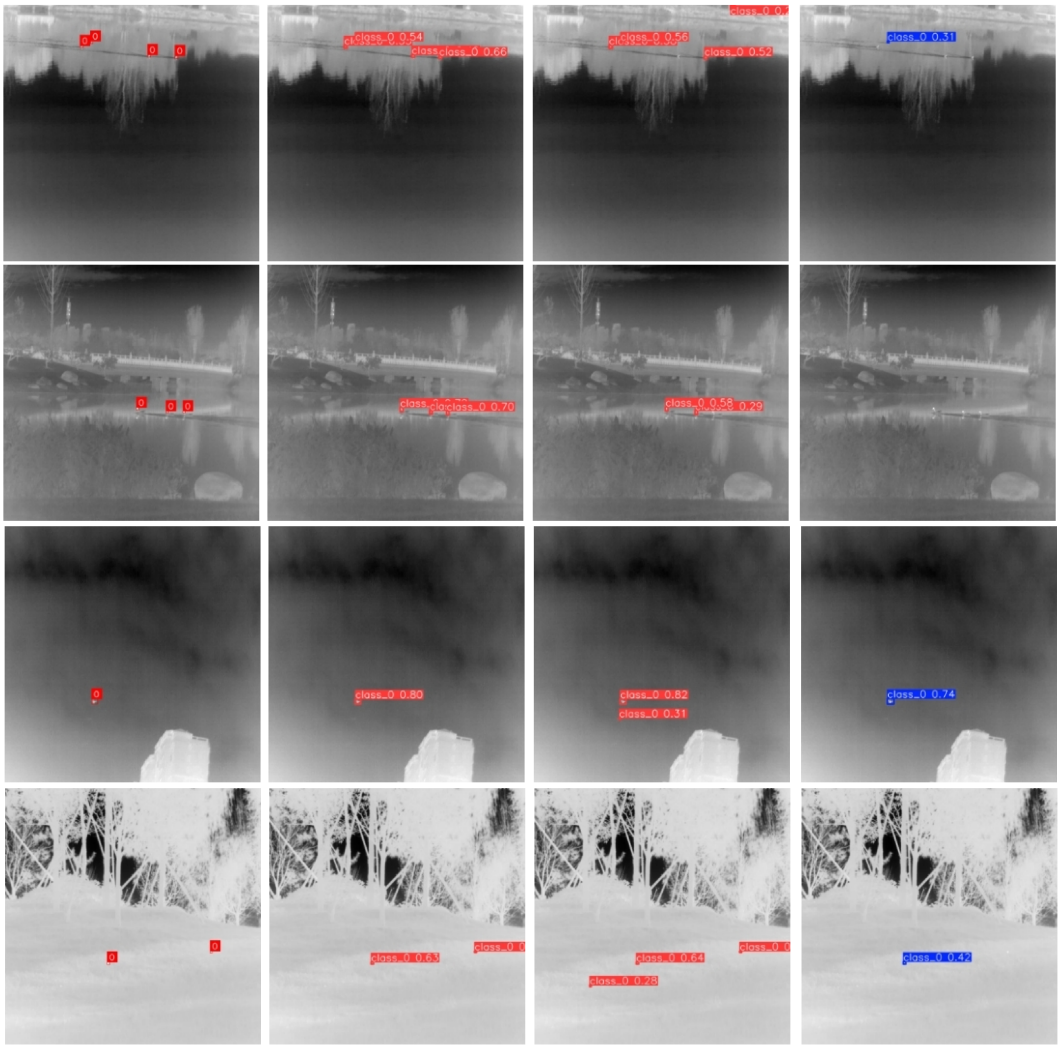


Figure 10. Qualitative comparison of infrared target detection on the IRSTD dataset. Columns correspond to (**a**) ground-truth annotations, (**b**) DCM-DETR, (**c**) RT-DETR, and (**d**) YOLO11n. Bounding boxes are shown with confidence scores.

The differences above primarily arise from the fact that the discriminative information of infrared small targets heavily relies on weak local energy peaks. Deep representations are diluted by downsampling or suffer from semantic-detail misalignment in cross-scale fusion, making target responses more easily overwhelmed by the gradient structures and statistical consistency of the background, leading to reduced recall and unstable localisation. DCM-DETR, by enhancing weak target representations and suppressing background-dominated responses, while also improving cross-scale alignment, ensures that the queries in the decoding stage more consistently focus on the real target neighborhoods, resulting in more stable detection performance in strong clutter and multi-target scenarios.

5. Conclusions

This paper addresses feature dilution, background interference, and localisation drift in infrared small UAV detection under weak contrast, heavy clutter, and severe scale

mismatch. We propose DCM-DETR, an end-to-end framework built on RT-DETR that strengthens weak target representations, suppresses background responses via global modelling, and improves cross-scale consistency. DRFE enlarges the effective receptive field and preserves faint target cues through directional displacement perspectives. In the encoder, the CSA-Block recalibrates saliency to reduce background-dominated attention aggregation, HGF/RA performs selective cross-layer fusion with scale-consistent alignment, and MALA-AIFI models global dependencies while correcting weak responses at lower complexity, thereby stabilising decoder queries on true target regions. Experiments on two public datasets show consistent gains in both accuracy and efficiency. On SIRST-UAVB, DCM-DETR improves mAP_{50} by 1.09% over RT-DETR, while reducing model size by 25.1 M and computation by 12.7 GFLOPs. On IRSTD, it further increases mAP_{50} by 2.01% over RT-DETR-r18 and runs at 93.45 FPS with 44.2 GFLOPs, delivering nearly a 2× speedup. Qualitative results also indicate fewer background-triggered false alarms and more stable localisation in cluttered scenes.

Nevertheless, the proposed method still has areas that warrant further research. First, under extreme conditions of low signal-to-noise ratio, long-distance imaging, or strong thermal reflection interference, target responses may still be mixed with high background highlights, causing confidence fluctuations and slight localisation shifts. This suggests that the discriminability of boundary samples could be further improved. Second, end-to-end detection might be influenced by data distribution discrepancies when transitioning across sensors, seasons, or scenes, and its generalisation robustness needs to be enhanced. Third, although complexity has been reduced, there is still potential for further lightweighting and hardware-friendliness, especially under stricter edge-end power consumption and latency constraints. Future work will focus on three aspects: introducing uncertainty modelling and hard sample learning to enhance stability under extreme weak target conditions; combining domain generalisation/domain adaptation strategies to improve cross-scene and cross-sensor transferability and further reducing computational overhead through structured pruning, quantisation, and efficient operator design to promote real-time deployment and long-term stable operation of the model in real-world applications such as low-altitude security, together with comprehensive on-device deployment and benchmarking on representative edge hardware under practical power and latency constraints.

Author Contributions: Conceptualisation, L.L. and J.S.; Methodology, L.L. and H.H.; Software, H.H.; Validation, J.S.; Visualisation, J.S.; Writing—original draft, L.L. and J.S.; Writing—review and editing, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency.

Data Availability Statement: The SIRST-UAVB is available at https://github.com/JN-Yang/PConv-SDloss-Data (accessed on 15 September 2025); the IRSTD-1K is available at https://github.com/RuiZhang97/ISNet (accessed on 20 October 2025).

Acknowledgments: The authors acknowledge the referees for carefully reading this paper and giving many helpful comments. The authors also express their gratitude to the reviewers for their insightful comments.

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

References

1. Zhang, C.; Kovacs, J.M. The application of small unmanned aerial systems for precision agriculture: A review. *Precis. Agric.* **2012**, *13*, 693–712. [CrossRef]
2. Gupta, L.; Jain, R.; Vaszkun, G. Survey of important issues in UAV communication networks. *IEEE Commun. Surv. Tutor.* **2015**, *18*, 1123–1152. [CrossRef]

3. Kumar, N.; Singh, P. Small and dim target detection in infrared imagery: A review, current techniques and future directions. *Neurocomputing* **2025**, *630*, 129640. [CrossRef]
4. Hamdi, A.; Alkouz, B.; Shahzaad, B.; Bouguettaya, A.; Neiat, A.; Salim, F.; Kim, D. Drone-as-a-Service: Research Challenges and Directions. *Proc. IEEE* **2025**, *113*, 0018–9219. [CrossRef]
5. Zhi, Y.; Fu, Z.; Sun, X.; Yu, J. Security and privacy issues of UAV: A survey. *Mob. Netw. Appl.* **2020**, *25*, 95–101. [CrossRef]
6. Wei, X.; Ma, J.; Sun, C. A survey on security of unmanned aerial vehicle systems: Attacks and countermeasures. *IEEE Internet Things J.* **2024**, *25*, 34826–34847. [CrossRef]
7. Rawat, S.S.; Verma, S.K.; Kumar, Y. Review on recent development in infrared small target detection algorithms. *Procedia Comput. Sci.* **2020**, *167*, 2496–2505. [CrossRef]
8. Zhao, M.; Li, W.; Li, L.; Hu, J.; Ma, P.; Tao, R. Single-frame infrared small-target detection: A survey. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 87–119. [CrossRef]
9. Han, J.; Moradi, S.; Faramarzi, I.; Liu, C.; Zhang, H.; Zhao, Q. A local contrast method for infrared small-target detection utilizing a tri-layer window. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1822–1826. [CrossRef]
10. Wang, Y.; Bai, S.; Liang, W.; Li, R.; Liu, X. Infrared small target detection based on weighted local contrast enhancement. *Infrared Phys. Technol.* **2025**, *150*, 105924. [CrossRef]
11. Zhao, M.; Li, L.; Li, W.; Tao, R.; Li, L.; Zhang, W. Infrared small-target detection based on multiple morphological profiles. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6077–6091. [CrossRef]
12. Gao, C.; Meng, D.; Yang, Y.; Wang, Y.; Zhou, X.; Hauptmann, A.G. Infrared patch-image model for small target detection in a single image. *IEEE Trans. Image Process.* **2013**, *22*, 4996–5009. [CrossRef]
13. Liu, Y.; Liu, X.; Hao, X.; Tang, W.; Zhang, S.; Lei, T. Single-frame infrared small target detection by high local variance, low-rank and sparse decomposition. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–17. [CrossRef]
14. Zhou, D.; Wang, X. Research on high robust infrared small target detection method in complex background. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
15. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 142–158. [CrossRef] [PubMed]
16. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
18. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
20. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
21. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
22. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
23. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. Yolov9: Learning what you want to learn using programmable gradient information. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–21.
24. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
25. Tian, Y.; Ye, Q.; Doermann, D. Yolov12: Attention-centric real-time object detectors. *arXiv* **2025**, arXiv:2502.12524.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
27. Dosovitskiy, A. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
28. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
29. 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, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
30. Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Syst. Appl.* **2021**, *172*, 114602. [CrossRef]

31. Liu, J.; Jing, D.; Zhang, H.; Dong, C. Srfad-net: Scale-robust feature aggregation and diffusion network for object detection in remote sensing images. *Electronics* **2024**, *13*, 2358. [CrossRef]
32. Jiang, Q.; Yu, H.; Jin, X.; Wang, P.; Lee, S.J.; Yao, S.; Jiang, H.; Lan, W.; Zhou, W. Attention-guided network for infrared unmanned aerial vehicle target detection. *Eng. Appl. Artif. Intell.* **2026**, *163*, 113105. [CrossRef]
33. Fan, Q.; Huang, H.; Ai, Y.; He, R. Rectifying magnitude neglect in linear attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 21505–21514.
34. Zhao, X.; Zhang, W.; Zhang, H.; Zheng, C.; Ma, J.; Zhang, Z. ITD-YOLOv8: An infrared target detection model based on YOLOv8 for unmanned aerial vehicles. *Drones* **2024**, *8*, 161. [CrossRef]
35. Zhao, L.; Zhang, Y.; Li, Y.; Zhong, H. Global and Local Context-Aware Detection for Infrared Small UAV Targets. *Drones* **2025**, *9*, 804. [CrossRef]
36. Sun, S.; Mo, B.; Xu, J.; Li, D.; Zhao, J.; Han, S. Multi-YOLOv8: An infrared moving small object detection model based on YOLOv8 for air vehicle. *Neurocomputing* **2024**, *588*, 127685. [CrossRef]
37. Wang, Y.; Wang, B.; Huo, L.; Fan, Y. GT-YOLO: Nearshore infrared ship detection based on infrared images. *J. Mar. Sci. Eng.* **2024**, *12*, 213. [CrossRef]
38. Zhan, W.; Zhang, C.; Guo, S.; Guo, J.; Shi, M. EGISD-YOLO: Edge guidance network for infrared ship target detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 10097–10107. [CrossRef]
39. Chen, Y.; Yuan, X.; Wang, J.; Wu, R.; Li, X.; Hou, Q.; Cheng, M.M. YOLO-MS: Rethinking multi-scale representation learning for real-time object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 4240–4252. [CrossRef]
40. Yuan, M.; Shi, X.; Wang, N.; Wang, Y.; Wei, X. Improving RGB-infrared object detection with cascade alignment-guided transformer. *Inf. Fusion* **2024**, *105*, 102246. [CrossRef]
41. Guo, J.; Gao, C.; Liu, F.; Meng, D. Dpdetr: Decoupled position detection transformer for infrared-visible object detection. *arXiv* **2024**, arXiv:2408.06123. [CrossRef]
42. Guo, J.; Gao, C.; Liu, F.; Meng, D.; Gao, X. Damsdet: Dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 464–481.
43. Yang, H.; Wang, J.; Bo, Y.; Wang, J. ISTD-DETR: A deep learning algorithm based on DETR and Super-resolution for infrared small target detection. *Neurocomputing* **2025**, *621*, 129289. [CrossRef]
44. Wang, S.; Jiang, H.; Li, Z.; Yang, J.; Ma, X.; Chen, J.; Tang, X. PHSI-RTDETR: A lightweight infrared small target detection algorithm based on UAV aerial photography. *Drones* **2024**, *8*, 240. [CrossRef]
45. Yue, X.; Liu, L.; Du, Y. IR-DETR: An efficient detection transformer with multi-layer feature fusion for infrared small targets. *Infrared Phys. Technol.* **2025**, *150*, 105926. [CrossRef]
46. Pi, Y.; Kong, L.; Yang, B.; Chang, R.; Pu, H.; Zhou, M.; Luo, J. IRTransUNet: Efficient transformer embedding UNet for infrared small target detection. *Infrared Phys. Technol.* **2025**, *151*, 106061. [CrossRef]
47. Xiao, Y.; Meng, F.; Wu, Q.; Xu, L.; He, M.; Li, H. Gm-detr: Generalized muiltispectral detection transformer with efficient fusion encoder for visible-infrared detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 5541–5549.
48. Zhu, Y.; Ma, Y.; Fan, F.; Huang, J.; Wu, K.; Wang, G. Toward accurate infrared small target detection via edge-aware gated transformer. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 8779–8793. [CrossRef]
49. Hu, J.; Zhang, T.; Zhao, M. Meac: A multi-scale edge-aware convolution module for robust infrared small-target detection. *Sensors* **2025**, *25*, 4442. [CrossRef]
50. Ma, X.; Guo, F.M.; Niu, W.; Lin, X.; Tang, J.; Ma, K.; Ren, B.; Wang, Y. Pconv: The missing but desirable sparsity in dnn weight pruning for real-time execution on mobile devices. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 5117–5124.
51. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Attentional local contrast networks for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9813–9824. [CrossRef]
52. Wu, F.; Zhang, T.; Li, L.; Huang, Y.; Peng, Z. RPCANet: Deep unfolding RPCA based infrared small target detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 4809–4818.
53. Yang, H.; Mu, T.; Dong, Z.; Zhang, Z.; Wang, B.; Ke, W.; Yang, Q.; He, Z. Pbt: Progressive background-aware transformer for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–13. [CrossRef]
54. Zhang, F.; Lin, S.; Xiao, X.; Wang, Y.; Zhao, Y. Global attention network with multiscale feature fusion for infrared small target detection. *Opt. Laser Technol.* **2024**, *168*, 110012. [CrossRef]
55. Xu, S.; Zheng, S.; Xu, W.; Xu, R.; Wang, C.; Zhang, J.; Teng, X.; Li, A.; Guo, L. Hcf-net: Hierarchical context fusion network for infrared small object detection. In Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME), Niagara Falls, ON, Canada, 15–19 July 2024; pp. 1–6.

56. Wu, H.; Huang, X.; He, C.; Xiao, H.; Luo, S. Infrared small target detection with Swin Transformer-based multi-scale atrous spatial pyramid pooling network. *IEEE Trans. Instrum. Meas.* **2024**, *74*, 5003914.
57. Tong, X.; Zuo, Z.; Su, S.; Wei, J.; Sun, X.; Wu, P.; Zhao, Z. ST-trans: Spatial-temporal transformer for infrared small target detection in sequential images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–19. [CrossRef]
58. Liu, J.; Wang, Y.; Cao, Y.; Guo, C.; Shi, P.; Li, P. Unified Spatial-Frequency modeling and alignment for Multi-Scale small object detection. *Symmetry* **2025**, *17*, 242. [CrossRef]
59. Yang, J.; Liu, S.; Wu, J.; Su, X.; Hai, N.; Huang, X. Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 9202–9210.
60. Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; Guo, J. ISNet: Shape matters for infrared small target detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 877–886.
61. Liu, J.; Zhao, J.; Cao, Y.; Wang, Y.; Dong, C.; Guo, C. Road manhole cover defect detection via multi-scale edge enhancement and feature aggregation pyramid. *Sci. Rep.* **2025**, *15*, 10346. [CrossRef]
62. 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.
63. Jing, Y.; Zheng, Y.; Zhao, J.; Cui, G.; Zhang, T. Make Both Ends Meet: A Synergistic Optimization Infrared Small Target Detection with Streamlined Computational Overhead. *arXiv* **2025**, arXiv:2504.21581. [CrossRef]
64. Yang, B.; Zhang, X.; Zhang, J.; Luo, J.; Zhou, M.; Pi, Y. EFLNet: Enhancing Feature Learning for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5906511. [CrossRef]
65. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.

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