remote sensing

MDPI

Article

DTRFR: A Unified Detector for Diverse Target Detection in High-Spatial-Resolution Spaceborne Infrared Video

Xiaoying Wu [1,2,3], Dandan Li [1,2], Xin Chen [1,2], Kai Hu [1,2,3] and Peng Rao [1,2,*]

[1] National Key Laboratory of Infrared Detection Technologies, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, 500 Yutian Road, Shanghai 200083, China; wuxiaoying23@mails.ucas.ac.cn (X.W.)
[2] Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: peng_rao@mail.sitp.ac.cn

Highlights

What are the main findings?

- A unified end-to-end framework (DTRFR) is developed for mixed-size infrared small-target detection in high-spatial-resolution spaceborne videos.
- Multi-scale feature extraction and adaptive temporal alignment are jointly employed to enhance robustness under size variation and dynamic backgrounds.

What are the implications of the main findings?

- The proposed framework enables reliable detection of both in-distribution and distribution-shift targets in realistic spaceborne infrared scenarios.
- This work provides a practical foundation for high-spatial-resolution spaceborne infrared video analysis and compatibility with existing multi-frame target detection requirements.

Abstract

Spaceborne infrared small-target detection plays a critical role in space-sky early warning, disaster rescue, and reconnaissance tracking, benefiting from all-time, all-weather, and wide-area monitoring capabilities. The deployment of high-spatial-resolution infrared payloads (ground sampling distance, GSD < 10 m) has introduced pronounced scale diversity among targets, leading to size-sensitive performance degradation in existing detectors and heightened risks of missed detections or false alarms in mixed-size scenarios. Furthermore, multi-frame infrared small-target detection methods often face challenges in maintaining consistent temporal coherence during feature propagation across sequences. To overcome these limitations in high-resolution spaceborne infrared videos, we propose DTRFR, an end-to-end unified detection framework built on an enhanced recurrent feature refinement architecture. This approach incorporates a realistic SITP-QLSD dataset derived from QLSAT-2 infrared backgrounds, featuring diverse scenes, multi-size small targets, and a dedicated generalization sub-test set with extremely small targets partially unseen in training; a multi-scale IRFeatureExtractor leveraging parallel convolutions and dilated receptive fields for improved cross-scale discrimination and clutter suppression; and an adaptive gating pyramid deformable alignment module to optimize sequence alignment and enhance temporal consistency, enabling robust performance across various clutter levels and dynamic backgrounds. Extensive evaluations on SITP-QLSD demonstrate that DTRFR attains competitive performance, achieving mIoU of 74.32% and Pd of 94.51% on the main set, with strong robustness on the generalization sub-test set (Pd = 92.37%). Compared to single-frame and multi-frame baselines, the proposed method achieves higher detection accuracy with significantly reduced false alarms, benefiting from multi-

Check for updates

Academic Editor: Gong Cheng

Received: 1 February 2026
Revised: 23 February 2026
Accepted: 2 March 2026
Published: 4 March 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.

scale feature extraction that enables robust detection of small targets of different sizes in infrared videos.

Keywords: infrared small-target detection; spaceborne infrared; multi-size target detection

1. Introduction

Spaceborne infrared small-target detection plays a pivotal role in critical applications such as space-sky early warning [1], disaster rescue [2,3], and reconnaissance tracking [4,5], owing to its unique all-time, all-weather, and wide-area monitoring capabilities [6]. This capability is particularly valuable for the timely discovery and tracking of high-value targets [7].

Historically, early spaceborne infrared payloads were limited by low resolution, in which weak small targets typically occupied fewer than 5 × 5 pixels and represented only 0.038% of a 256 × 256 image [8]. Recent rapid advancements in high-spatial-resolution infrared payloads (ground sampling distance, GSD < 10 m) have enabled simultaneous refinement of targets and backgrounds [9]. Examples include KOMPSAT-3A (5.5 m mid-wavelength infrared, MWIR), WorldView-3 (3.7 m short-wavelength infrared, SWIR) [10], HotSat-1 (3.5 m MWIR) [11], and Albedo Clarity-1 (2 m thermal infrared) . For aircraft targets with wingspans of 25–65 m, pixel coverage has expanded significantly. Infrared small targets are typically defined as those occupying less than 0.15% of the image area [12]. As a result, spaceborne infrared small-target detection has shifted from primarily point-like ultra-small targets to mixed-size small targets, introducing substantial challenges from geometric scale diversity.

The continuous improvement of GSD also leads to finer-grained and highly heterogeneous earth-facing backgrounds (e.g., land heat sources, ocean waves, cloud edges, and atmospheric disturbances), significantly increasing local false alarm risks due to clutter resembling targets. This evolution has created an urgent need for advanced detection methods tailored to the high-resolution era.

1.1. Single-Frame Infrared Small-Target Detection (SIRST)

Single-frame methods focus on spatial cues to separate targets from backgrounds in individual images. They are typically classified into four main categories [13]:

Filter-based decoupling: These approaches use spatial or frequency-domain filters to suppress backgrounds and highlight isolated targets. Examples include max/median filters [14], morphological top-hat transformations [15,16], wavelet transforms [17], and quaternion Fourier phase spectrum methods [18,19].

Human visual system (HVS)-inspired local contrast: Motivated by visual perception, these methods enhance targets via local contrast measures such as LCM [20], RLCM [21], ELCM [22], and LEF-LCD [23], which is used for background suppression, and local inverse entropy techniques [24].

Tensor decomposition: By exploiting structural priors, these techniques separate sparse targets from low-rank backgrounds. The infrared patch-image (IPI) model [25] and its variants (WIPI [26], NIPPS [27], NRAM [28]) are prominent examples, with extensions incorporating tensor representations for limited information scenarios [29].

With the development of deep learning and open-source datasets [30–34], numerous CNN-based single-frame detectors [30–32,35] have emerged, effectively utilizing local information. However, these methods often underutilize global context. Recent advances in sequence modeling, including Transformers [36] and Mamba [37], have enabled hy-

brid models such as MiM-ISTD [38] to integrate global and local features for improved performance.

Single-frame methods perform reasonably in air-based scenarios but struggle in complex earth-facing spaceborne environments, where temporal information is essential for higher accuracy.

1.2. Multi-Frame Infrared Small-Target Detection (MIRST)

Multi-frame methods leverage spatiotemporal relationships across video sequences to improve detection in dynamic scenes. Recent progress includes 3D convolutions, LSTM/RNN-based modeling, and attention mechanisms.

CNN-based approaches [39,40] learn motion features in a data-driven manner. Examples include SSTNet with Conv-LSTM for spatiotemporal tensor extraction [41], 3D convolution combined with Conv-LSTM [42], STDMANet for multi-scale spatiotemporal attention [43], signal-to-signal networks exploiting temporal differences [44], spatial–temporal Transformers for global self-attention [45], deformable attention aggregation and cross-scale attention with multi-label propagation [46]. However, infrared small targets lack sufficient details, limiting the effectiveness of traditional global modeling with Transformers. Hybrid methods like LMAFormer with optical flow-guided cross-attention [47] improve diversity handling but incur high computational cost.

For spaceborne infrared video MIRST, approaches such as motion-encoded temporal relations [48] and Recurrent Feature Refinement (RFR) [49] have shown promise. RFR accumulates temporal information via recurrent refinement to boost precision, but inevitably accumulates errors over sequences, particularly in high-resolution dynamic scenes with complex backgrounds and motion. More recently, DQAligner [50] introduces a dynamic query aligner with a dynamic receptive field pyramid deformable convolution, expanding the adaptive dynamic receptive field to better handle large-motion and multi-scale target displacements for precise feature alignment. Additionally, MFE-Net [51] enhances motion features across multi-frame sequences through dynamic background mapping and multi-frame differencing, improving detection robustness for dynamic infrared space targets under complex conditions.

In salient object detection (SOD), scale imbalance frequently causes biased performance [52]. Existing models, constrained by low-resolution assumptions, often overlook compatibility with mixed-size targets under high-resolution payloads, leading to scale-sensitive fluctuations, missed small targets, and increased false alarms. Meanwhile, MIRST methods face persistent challenges in balancing detection accuracy and error propagation during temporal information utilization.

To address mixed-size small-target detection (i.e., different-sized targets within the small-target regime) and compatibility with extremely small targets (including partly out-of-distribution cases) in high-resolution spaceborne infrared earth-facing scenarios, this paper proposes DTRFR, a unified end-to-end framework based on enhanced Recurrent Feature Refinement. The main contributions are:

1. Construction of the SITP-QLSD dataset using real QLSAT-2 infrared backgrounds, featuring diverse scenes, mixed-size targets (5×5–9×9), and a generalization sub-test set (3×3–5×5) with extremely small targets, filling the gap in evaluating size-difference impacts.
2. Design of a multi-scale IRFeatureExtractor using serial-to-parallel convolutions and dilated receptive fields to enhance cross-scale discriminability and clutter suppression.
3. Proposal of an adaptive Gating Pyramid Deformable Alignment mechanism to optimize multi-frame feature alignment and improve detection robustness in sequences with dynamic backgrounds.

Experiments demonstrate competitive performance on the main set and robust generalization on the sub-test set, validating the method's adaptability and extension potential.

2. Materials and Methods

2.1. Overall Network Architecture

Given the unique characteristics of infrared small targets in satellite videos—such as low signal-to-noise ratio (SNR), significant scale variations, and high susceptibility to background clutter—the proposed method enables efficient detection without relying on overly complex or large-scale models.

As illustrated in Figure 1, the proposed network is built upon the Recurrent Feature Refinement (RFR) architecture [49], with targeted modifications and enhancements tailored for multi-frame infrared small-target detection (MIRST).

The method processes a sequence of infrared frames in a recurrent manner. For each current frame (denoted as frame i), the IRFeatureExtractor first extracts spatially robust and small-target-optimized feature representations from the raw input frame. Meanwhile, features from the previous frame (frame $i-1$) are propagated forward through the adaptive gating pyramid deformable alignment (AGPDA) module, which performs adaptive deformable alignment to generate well-aligned propagated features that account for motion and background dynamics across frames. These aligned propagated features are then further refined by the temporal–spatial–frequency modulation (TSFM) module (following the formulation in [49] with minor task-specific adaptations), which enhances temporal consistency and suppresses irrelevant spatiotemporal noise. Finally, the current-frame features from the IRFeatureExtractor are fused with the refined propagated features, and the combined spatiotemporal features are fed into the detection head. The detection head, implemented as a standard ResUNet architecture in this work, processes the fused feature maps to produce precise target predictions, outputting either segmentation masks or bounding boxes for the small infrared targets.

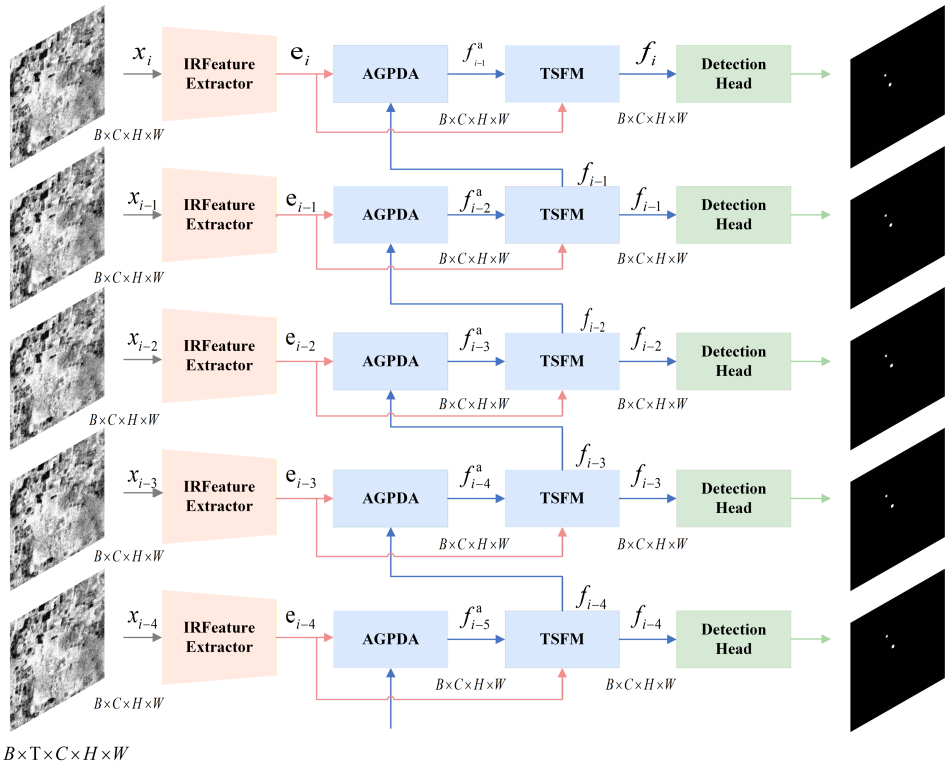


Figure 1. Overall network structure of the proposed DTRFR framework.

The following subsections provide detailed descriptions of the two novel modules introduced in this work: the IRFeatureExtractor (see Figure 2) and the adaptive gating pyramid deformable alignment (AGPDA) module (see Figure 3).

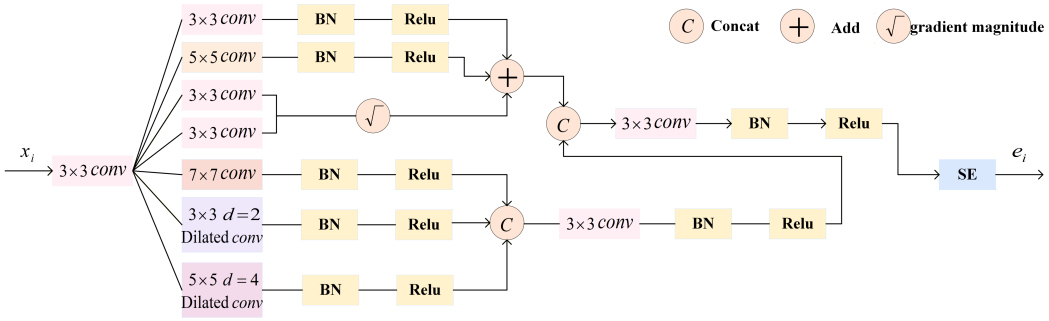


Figure 2. IRFeatureExtractor.

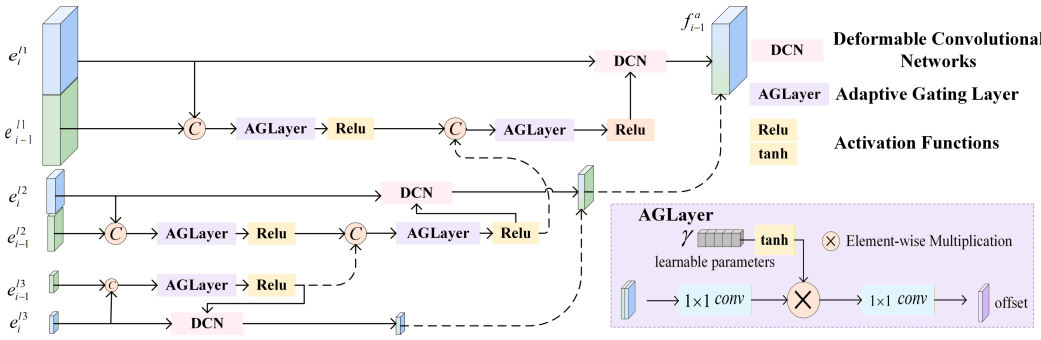


Figure 3. Adaptive gating pyramid deformable alignment.

2.1.1. IRFeatureExtractor

This subsection details the architecture and design principles of the proposed IRFeatureExtractor module, a key innovation of this work. Shown in Figure 2, the module aims to significantly enhance infrared small-target detection performance through a serial-to-parallel multi-scale feature extraction mechanism.

The core design involves first extracting shallow features in series, followed by parallel multi-scale convolutions with dilated receptive fields. This structure strengthens the discriminability of features for infrared small targets across different scales while effectively suppressing similar background clutter. It improves detection robustness under dynamic backgrounds and high-frequency noise, and meets the compatibility requirements for diverse small targets in high-spatial-resolution scenarios. The serial-to-parallel multi-scale convolutions and dilated receptive fields enable the module to learn multi-scale features in a data-driven manner by capturing contextual information across sizes, effectively mitigating this imbalance and enabling robust representation of both small and extremely small targets.

As an auxiliary component, the module incorporates an edge feature extraction branch with spatially directional, initialized learnable convolutions. However, it contributes only subtle edge cues, avoiding excessive noise amplification or boundary blurring in infrared images where targets often lack sharp edges.

The overall innovation stems from a fusion and optimization of existing techniques. Multi-scale global feature extraction draws inspiration from the atrous spatial pyramid pooling (ASPP) in DeepLab [53] and the multi-kernel parallel design in GoogLeNet's Inception modules [54], enabling efficient capture of multi-scale context. Edge enhancement is motivated by prior work embedding Sobel operators into CNNs for trainable gradient

extraction [55], but is treated as a secondary aid, with the global multi-scale mechanism remaining dominant to ensure robust and targeted feature representation.

The module proceeds as follows:

First, shallow features are extracted via an initial convolution:

$$x' = \sigma(\text{Conv}_{3\times3}(x)), \tag{1}$$

where $\text{Conv}_{k\times k}$ denotes a 2D convolution with kernel size $k \times k$, stride 1, and padding to preserve spatial dimensions; σ is the ReLU activation function; and the output is $x' \in \mathbb{R}^{B\times C_{\text{mid}}\times H\times W}$.

Second, serial-to-parallel multi-scale convolutions extract deep global and local features. The multi-scale global branch uses parallel paths to capture scale variations in infrared small targets (from point-like to mildly extended shapes) and handle dynamic backgrounds (e.g., moving clouds or ocean waves) and high-frequency noise (e.g., sensor artifacts). The branch consists of a large-kernel convolution and dilated convolutions with varying rates.

The large-kernel path aggregates contextual information:

$$f_1 = \sigma(\text{BN}(\text{Conv}_{7\times7}(x'))), \tag{2}$$

where BN denotes batch normalization to stabilize training and mitigate internal covariate shift.

The dilated paths use dilation rates $d = 2$ and $d = 4$ (with kernel size $k = 3$), yielding effective receptive fields of $k + (k-1)(d-1)$ (5×5 and 9×9 for $k = 3$):

$$f_2 = \sigma(\text{BN}(\text{Conv}_{3\times3,d=2}(x'))), \quad f_3 = \sigma(\text{BN}(\text{Conv}_{3\times3,d=4}(x'))). \tag{3}$$

These parallel paths enable scale-aware feature discrimination: smaller kernels/dilations capture fine details for point targets, while larger ones model extended context to suppress background clutter. The fusion layer concatenates features along the channel dimension and reduces dimensionality via a 1×1 convolution:

$$f_{\text{global}} = \sigma(\text{BN}(\text{Conv}_{1\times1}([f_1; f_2; f_3]))), \tag{4}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation. Theoretically, this multi-scale approach optimizes the intersection-over-union ratio between target and background features in the representation space. In noisy environments, dilated convolutions act as low-pass filters through sparse sampling, reducing sensitivity to local perturbations. In dynamic backgrounds, parallel paths allow adaptive scale weighting, enhancing overall robustness.

As an auxiliary branch, local edge features provide subtle refinement cues, particularly in edge-like clutter scenarios. However, edge reliance is deliberately limited, as excessive dependence can amplify high-frequency noise in infrared images where targets often lack sharp boundaries. Local features are extracted via parallel standard and Sobel convolutions:

$$e_1 = \sigma(\text{BN}(\text{Conv}_{3\times3}(x'))), \quad e_2 = \sigma(\text{BN}(\text{Conv}_{5\times5}(x'))), \tag{5}$$

with Sobel gradient detection kernels initialized as

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad K_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \tag{6}$$

replicated across channels and made trainable:

$$e_x = \text{Conv}_{K_x}(x'), \quad e_y = \text{Conv}_{K_y}(x'), \quad e_{\text{mag}} = \sqrt{e_x^2 + e_y^2}. \tag{7}$$

The local edge feature is then as follows:

$$e_{\text{local}} = e_1 + e_2 + e_{\text{mag}}. \tag{8}$$

Due to potential target gradients being vulnerable to noise, this branch serves only as a minor supplement.

Finally, global and local features are concatenated and fused to form deep representations:

$$f_{\text{fused}} = \sigma(\text{BN}(\text{Conv}_{3\times3}([f_{\text{global}}; e_{\text{local}}]))). \tag{9}$$

To enable channel-wise interaction, squeeze-and-excitation (SE) attention [56] is applied:

$$y = \text{AvgPool}(f_{\text{fused}}), \quad s = \sigma(\text{FC}_2(\sigma(\text{FC}_1(y)))), \quad e = f_{\text{fused}} \odot s, \tag{10}$$

where $\odot$ denotes element-wise multiplication, and FC layers compress channels by a reduction ratio of 16.

Notably, this approach effectively mitigates scale imbalance in multi-target object detection tasks by learning multi-scale features that capture contextual information at different scales, enabling robust detection of small targets of different sizes in high-resolution scenarios.

2.1.2. Adaptive Gating Pyramid Deformable Alignment (AGPDA)

Shown in Figure 3, the Adaptive Gating Pyramid Deformable Alignment (AGPDA) module takes as input the current-frame feature $\mathrm{e}_t \in \mathbb{R}^{B\times C\times H\times W}$ and the previous-frame feature $\mathrm{e}_{t-1} \in \mathbb{R}^{B\times C\times H\times W}$, where B is the batch size, C is the number of channels, and $H \times W$ denotes the spatial dimensions. The module generates multi-scale feature representations through pyramid downsampling, followed by learnable gating offsets and deformable convolution to align the previous-frame features to the current frame, producing the aligned feature $\mathrm{f}_{t-1}^{a} \in \mathbb{R}^{B\times C\times H\times W}$.

To capture multi-scale information for targets and backgrounds, the module constructs a three-level feature pyramid via convolutional downsampling. For both the current and previous frames, each pyramid level is generated through a two-stage convolutional process consisting of stride-2 downsampling followed by stride-1 refinement:

$$\tilde{\mathrm{e}}_{t,l2} = \sigma(\text{Conv}_{3\times3,s=2}(\mathrm{e}_{t,l1})), \quad \mathrm{e}_{t,l2} = \sigma(\text{Conv}_{3\times3,s=1}(\tilde{\mathrm{e}}_{t,l2})), \tag{11}$$

$$\tilde{\mathrm{e}}_{t,l3} = \sigma(\text{Conv}_{3\times3,s=2}(\mathrm{e}_{t,l2})), \quad \mathrm{e}_{t,l3} = \sigma(\text{Conv}_{3\times3,s=1}(\tilde{\mathrm{e}}_{t,l3})). \tag{12}$$

with analogous operations applied to e_{t-1} to generate the three-level features $\{\mathrm{e}_{t,l1}, \mathrm{e}_{t,l2}, \mathrm{e}_{t,l3}\}$ and $\{\mathrm{e}_{t-1,l1}, \mathrm{e}_{t-1,l2}, \mathrm{e}_{t-1,l3}\}$, where σ is the ReLU activation function ensuring nonlinear expressiveness. This pyramid structure integrates cross-scale information, enhancing robustness for small targets in backgrounds with complex motion dynamics.

The adaptive gating layer, shown in the bottom-right corner of Figure 3, is the core innovation of the module. It generates learnable offset guidance for feature alignment using two 1×1 convolutions (to minimize parameters) combined with a learnable parameter $\gamma \in \mathbb{R}^{C\times1\times1}$:

$$\mathrm{x}_1 = \text{Conv}_{1\times1}(\mathrm{x}), \quad \mathrm{x}_2 = \mathrm{x}_1 \odot \tanh(\gamma), \quad \mathrm{x}_{\text{out}} = \text{Conv}_{1\times1}(\mathrm{x}_2), \tag{13}$$

where $\odot$ denotes element-wise multiplication, and tanh is a non-linear activation function whose output is bounded in the range $[-1, 1]$. This bounded range allows $\tanh(\gamma)$ to function as an effective gating activation: values close to $+1$ enable strong feature amplification and pass-through, values near 0 suppress irrelevant or noisy channels, and values near -1 permit sign inversion when advantageous. By leveraging the non-linear and saturating properties of tanh, the adaptive gating layer effectively suppresses redundant and irrelevant information during feature alignment, thereby providing more precise and robust offset guidance for deformable convolution with extremely low parameter overhead (only C parameters for γ). Compared to conventional attention mechanisms (e.g., SE), this gating layer has an extremely low parameter count (only $C \times 1 \times 1$ offset terms) yet effectively adapts feature importance, suppressing background noise and motion artifacts.

To achieve precise alignment of previous-frame features to the current frame, the module employs deformable convolution, with offsets generated by the adaptive gating layer:

$$\text{offset}_i = \text{AdaptiveGatingLayer}([\text{e}_{t-1,\text{l}i}; \text{e}_{t,\text{l}i}]), \tag{14}$$

$$\text{feat}_i = \text{DeformConv}(\text{e}_{t-1,\text{l}i}, \text{offset}_i), \tag{15}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation, and $\text{e}_{t-1,\text{l}i}$ and $\text{e}_{t,\text{l}i}$ are the i-th pyramid level features. Deformable convolution dynamically adjusts the sampling locations of the convolution kernel via the learned offsets, accommodating non-rigid target motion and background variations. The gating-guided offsets further enhance alignment precision and temporal consistency, improving feature reliability and overall detection robustness across sequences.

At each pyramid level (from high to low resolution), the module processes offsets and features sequentially:

At the highest level l3, initial offsets are generated using current and previous frame features.

At intermediate levels l2 and l1, current-level features are fused with upsampled features from the previous level:

$$\text{offset}_i = \sigma(\text{AdaptiveGatingLayer}([\text{offset}_i; \text{Upsample}(\text{offset}_{i+1})])), \tag{16}$$

$$\text{feat}_i = \text{AdaptiveGatingLayer}([\text{feat}_i; \text{Upsample}(\text{feat}_{i+1})]), \tag{17}$$

where Upsample denotes bilinear interpolation with a scale factor of 2 to match spatial dimensions. This cross-scale fusion ensures that high-resolution features benefit from long-range contextual information from lower-resolution levels, enhancing adaptability to complex motion backgrounds.

2.1.3. Loss Function

To address the severe foreground–background class imbalance inherent in infrared small-target detection, this work adopts a multi-frame averaged SIoU loss function [49].

Let the predicted probability map be $P \in [0,1]^{H \times W}$ and the corresponding binary ground-truth mask be $G \in \{0,1\}^{H \times W}$. The intersection over union (IoU) is defined as follows:

$$\text{IoU}(P, G) = \frac{\sum_i P_i \cdot G_i + 1}{\sum_i P_i + \sum_i G_i - \sum_i P_i \cdot G_i + 1}, \tag{18}$$

where i indexes pixels, and the constant 1 is a smoothing term to prevent division by zero and improve training stability.

Based on this definition, the SIoU loss for a single frame is as follows:

$$\mathcal{L}_{\text{SIoU}} = 1 - \text{IoU}(P, G). \tag{19}$$

For multi-frame predictions, the loss is averaged across frames to ensure balanced contributions from different time steps and to mitigate temporal inconsistency in sequences:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{\text{SIoU}}^{(n)}, \tag{20}$$

where N denotes the number of temporal frames.

This multi-frame averaging strategy not only alleviates the extreme imbalance between sparse targets and dense backgrounds but also promotes temporal consistency in the optimization process, making it particularly suitable for multi-frame infrared small-target detection tasks.

3. Results

3.1. Dataset Preparation

To comprehensively evaluate the proposed DTRFR method for small-target detection in high-spatial-resolution spaceborne infrared videos, we constructed the SITP-QLSD dataset and performed comparative experiments with the public SITP-QLEF dataset [6]. Detailed parameters of all datasets are summarized in Table 1.

Table 1. Detailed parameters of the datasets used in this study.

Dataset	Target Size	Target Pixel	Background Std	Seq.	Mode	Frames	T-Num	SNR	GSD	Resolution	Band
1	5×5–9×9	42.1 ± 12	69.20	88	Push-broom	45,425	80,483	2–4	14 m	256×256	MWIR
2	3×3–5×5	15.6 ± 8	–	12	Push-broom	6395	15,956	2–4	14 m	256×256	MWIR
3	5×5	25	26.52	67	Staring	11,029	23,796	4	14 m	256×256	MWIR

We constructed the SITP-QLSD dataset using real high-resolution infrared video data from the Qilu-2 satellite (acquired in 2024 by the core payload developed by the Shanghai Institute of Technical Physics, with a GSD of 14 m in the mid-wavelength infrared). As shown in Figure 4, the dataset covers diverse dynamic terrestrial scenes including mountains, lakes, urban areas, clouds, rivers, farmlands, harbors, and oceans with a background standard deviation of up to 69.20, reflecting the high granularity and complexity under high-resolution imaging.

Small targets of different sizes are synthetically injected using a Gaussian kernel simulation $G(t_{\text{size}}, \sigma)$, where t_{size} ranges from 3 to 9 pixels and $\sigma = 0.15$, balancing typical high-resolution payload requirements (e.g., aircraft wingspans of 25–65 m) and compatibility with existing small-target detection scenarios. The infrared camera is simulated in push-broom mode. A small, controlled relative scene shift of 0–2 pixels per frame is introduced to mimic platform motion and increase background temporal complexity.

To simulate realistic infrared small-target detection video sequences, targets are initially placed at random positions $(x_{\text{initial}}, y_{\text{initial}})$ and move along the x and y directions with velocities $v_x = \frac{x_T - x_{\text{initial}}}{T}$ and $v_y = \frac{y_T - y_{\text{initial}}}{T}$, where (x_T, y_T) is the position at frame T and velocities range from -2 to 2 pixels/frame. Upon complete exit from the field of view, new targets enter from the edge at (x_{t1}, y_{t1}), with updated velocities $v_x = \frac{x_T - x_{t1}}{T - t_1}$ and $v_y = \frac{y_T - y_{t1}}{T - t_1}$. The direction is randomized relative to the image rotation center based on entry orientation.

Target energy injection is strictly controlled by a preset target SNR (TSNR, randomly generated from 2–4). Given the local background mean $I_b(t)$ and standard deviation $\sigma_b(t)$ at frame t, the normalized target template energy is

$$E_{\text{norm}} = \sum_{x,y} G_{\text{target}}(x, y), \tag{21}$$

the peak amplitude A is computed as

$$A = \text{TSNR} \cdot \sigma_b(t), \tag{22}$$

and the synthetic image is

$$I_{\text{syn}}(x, y; t) = I_b(t) + A \cdot \frac{G_{\text{target}}(x - x_c, y - y_c)}{E_{\text{norm}}}, \tag{23}$$

where (x_c, y_c) is the sub-pixel center coordinate of the target (supporting bilinear interpolation placement).

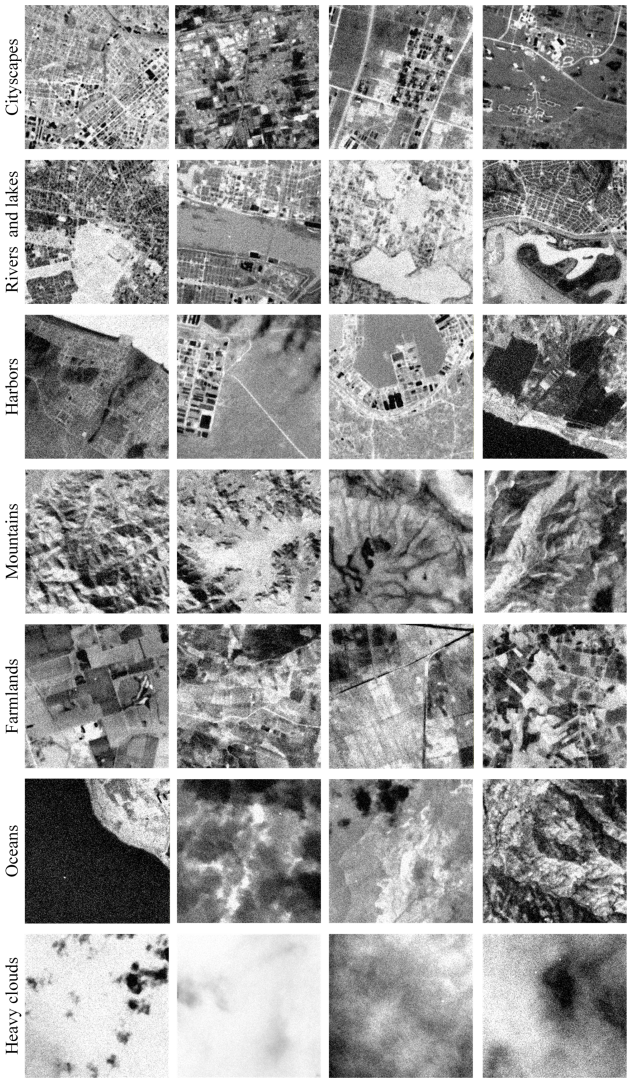


Figure 4. Examples of diverse terrestrial scenes in the SITP-QLSD dataset.

The SITP-QLSD dataset comprises two complementary parts: Dataset 1 (main test set) and Dataset 2 (generalization sub-test set). Dataset 1 features targets sized 5×5–9×9 pixels, reflecting typical distributions under current high-resolution payloads;

the training and test sets are split approximately 8:2 by sequence (71 for training and 17 for testing). Dataset 2 is exclusively used as an independent test set, with approximately 66.7% of the test sequences from Dataset 1 randomly selected and augmented with partly zero-shot extremely small targets (3×3–5×5 pixels, average size about 33.3% of Dataset 1) to evaluate robustness under scale distribution shifts.

Dataset 3 consists of the raw infrared sequences from SITP-QLEF [6], acquired in staring imaging mode. It exhibits relatively uniform target sizes (mean $\approx$ 25 pixels), lower background complexity (standard deviation 26.52), and the same GSD of 14 m. To simulate realistic platform jitter in staring scenarios, mild motion blur equivalent to approximately 0.004° platform vibration is incorporated. Dataset 3 serves as a supplementary benchmark to validate the proposed method's performance and robustness across different imaging modes (push-broom scanning vs. staring).

The target size distributions across the datasets are compared in Figure 5, highlighting the size diversity in SITP-QLSD compared to the more uniform sizes in Dataset 3.

In all experiments, the model is trained and evaluated independently on Dataset 1 and Dataset 3, while Dataset 2 is strictly excluded from training and reserved solely for independent generalization testing. This protocol ensures fair, unbiased, and realistic assessment of cross-size and cross-scene robustness.

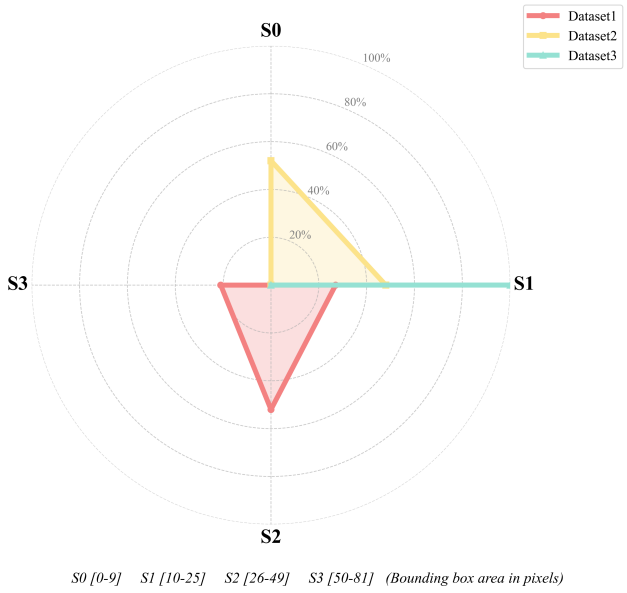


Figure 5. Target size distribution comparison across datasets (radar chart view).

3.2. Evaluation Metrics

To comprehensively evaluate the proposed method for multi-size infrared small-target detection, we adopt a dual-level evaluation framework consisting of target-level and pixel-level metrics. This combined approach provides a balanced assessment of practical target detection reliability and fine-grained segmentation quality. Together, these complementary perspectives offer a thorough assessment of both detection reliability and segmentation precision. Specifically, target-level evaluation assesses whether each individual small target is correctly detected and localized as a whole object: a prediction is counted as a true positive if its bounding box, center point, or segmentation mask satisfies a predefined overlap or distance criterion with the ground-truth target. In contrast, pixel-level evaluation treats the task as dense per-pixel prediction and measures the agreement between the predicted probability map/segmentation map and the ground-truth mask at every pixel position.

3.2.1. Target-Level Metrics

- Detection Probability (Pd): The fraction of ground-truth targets that are successfully detected.

$$\mathrm{Pd} = \frac{N_{\mathrm{true}}}{N_{\mathrm{gt}}}, \tag{24}$$

  where N_{true} is the number of correctly matched ground-truth targets and N_{gt} is the total number of ground-truth targets.
- False Alarm Rate (FA): The average number of falsely detected instances per image (FA per image) [47], which is the most widely adopted form in infrared small-target detection literature for its direct engineering interpretability.

$$\mathrm{Fa} = \frac{N_{\mathrm{false}}}{N_{\mathrm{all}}}, \tag{25}$$

  where N_{false} denotes the number of background pixels incorrectly classified as target pixels, N_{all} represents the total number of pixels in the image.
- Receiver Operating Characteristic (ROC) Curve: The trade-off curve obtained by plotting Pd against Fa while varying the detection confidence threshold. The area under this curve serves as a threshold-independent indicator of the model's ability to achieve high detection rates while maintaining low false alarm levels.

All target-level metrics follow the shooting-rules criterion proposed in [48], which is specifically designed for small infrared targets with point-level or low-resolution annotations. A ground-truth target is considered correctly detected if any predicted pixel falls within a 3×3 region centered at the ground-truth centroid. Predicted pixels lying outside the 9×9 exclusion zone around all ground-truth centroids are counted as false alarms. This pixel-by-pixel judgment rule prevents over-counting and ensures realistic one-to-one matching.

Target-level metrics are prioritized in engineering contexts because they emphasize target uniqueness, suppress duplicate alarms, and better reflect deployment reliability under high-clutter or distribution-shifted conditions.

3.2.2. Pixel-Level Metrics

- Mean Intersection over Union (mIoU): Averages the per-sample IoU across all test images, effectively mitigating the dominance of larger targets.

$$\mathrm{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i + \mathrm{FN}_i}, \tag{26}$$

  where N is the total number of samples, and TP_i, FP_i, FN_i are the true positive, false positive, and false negative pixels in the i-th sample.

The pixel-level metric offers a fine-grained assessment of segmentation precision and background suppression capability. It is less sensitive to matching ambiguities and annotation inconsistencies, making it complementary to target-level evaluation.

All statistics are accumulated globally over the test set. This dual-level reporting ensures a complete, complementary, and application-oriented evaluation consistent with established practices in the infrared small-target detection community.

3.3. Quantitative Results

In this section, we conduct comparative experiments between the proposed DTRFR and representative single-frame infrared small-target detection (SIRST) methods, including ACM [57], ALCNet [58], DNANet [59], ISTUD-UNet [60], ResUNet [61], UIU-

Net [62], and MSHNet [63]. RFR-based frameworks [49] (i.e., RFR+ACM, RFR+ALCNet, RFR+DNANet, and RFR+ResUNet) are selected as the primary multi-frame baselines because they are specifically designed for spaceborne infrared video and explicitly address sequence temporal propagation, which is the core challenge considered in this work.

For single-frame methods, all hyperparameters follow the original papers except for the learning rate and number of training epochs. Multi-frame methods uniformly adopt a sequence length of 10. All models are trained from scratch for 20 epochs using an input image resolution of 128 × 128 on a single NVIDIA RTX 4090D GPU. The learning rate is initialized at 0.0005 for Dataset 1 (Dataset 2 is excluded from training) and 0.01 for Dataset 3 (following the original paper setting), with a decay factor of 0.5 applied every 5 epochs. During inference, a fixed threshold of 0.5 is used to generate binary detection masks.

Quantitative results are reported separately on Dataset 1 (main test set), Dataset 2 (generalization sub-test set), the full SITP-QLSD (Dataset 1 + Dataset 2), and Dataset 3, as detailed in Tables 2–4. The best-performing result in each column is enclosed in a solid box, and the second-best result is enclosed in a dashed box.

Table 2. Test results on Dataset 1. The best-performing result is enclosed in a solid box, and the second-best result is enclosed in a dashed box.

Frame	Methods	Pd (%)	FA ($\times 10^{-5}$)	mIoU (%)	Time (ms)
Single Frame	ACM	67.26	84.38	32.54	1.39
	ALCNet	67.18	72.86	34.46	1.34
	DNANet	61.65	20.24	44.23	17.54
	ISTUD-UNet	70.90	38.40	44.40	4.15
	ResUNet	63.14	25.37	43.42	0.10
	UIU-Net	88.97	4356.07	2.38	7.00
	MSHNet	61.31	27.16	42.22	7.92
Multi Frame	RFR+ACM	69.52	40.06	42.53	2.21
	RFR+ALCNet	74.24	82.17	35.39	2.16
	RFR+DNANet	63.08	15.71	46.36	18.71
	RFR+ResUNet	68.59	30.26	45.20	2.06
	Ours	94.51	5.88	74.32	2.68

On Dataset 1 (Table 2), which reflects typical target size distributions under current high-resolution payloads (primarily 5 × 5–9 × 9 pixels), the proposed method achieves an mIoU of 74.32%, representing an absolute improvement of 27.96 percentage points (a relative improvement of 60.3%) over the strongest multi-frame baseline, RFR+DNANet (46.36%). The Pd reaches 94.51%, surpassing all single-frame methods (highest: UIU-Net at 88.97%, an improvement of 5.54 percentage points) and outperforming the best multi-frame method RFR+ALCNet (74.24%) by 20.27 percentage points. The FA is only 5.88×10^{-5}, the lowest among the 12 compared methods, representing a 62.6% reduction compared to RFR+DNANet (15.71×10^{-5}), nearly three orders of magnitude lower than the high-Pd single-frame method UIU-Net (4356.07×10^{-5}), and approximately 92.8% lower than RFR+ALCNet (82.17×10^{-5}). These results demonstrate comprehensive superiority in detection accuracy, false alarm suppression, and real-time performance under typical high-resolution earth-facing scenarios.

In terms of inference efficiency, the proposed method achieves a low latency of 2.68 ms/frame (approximately 373 FPS), ranking among the fastest high-Pd and high-mIoU methods. It is significantly faster than the best multi-frame segmentation baseline RFR+DNANet (18.71 ms, ∼7× faster) and outperforms the single-frame high-Pd method DNANet (17.54 ms). Notably, while achieving the highest mIoU and Pd, the FA remains at the 10^{-5} level with superior inference speed, realizing a global optimum across accuracy, false alarms, and real-time performance.

On Dataset 2 (Table 3), serving as an independent generalization sub-test set with target sizes significantly reduced to 3×3–5×5 pixels (area approximately 1/4–1/9 of Dataset 1, average size about 33.3% of Dataset 1) and including zero-shot extremely small targets partially unseen in training, the method demonstrates strong robustness under severe scale distribution shifts. Although the inherent blob expansion of extremely small targets leads to a noticeable drop in mIoU and a slight increase in FA compared to Dataset 1, DTRFR still accurately detects weak targets while effectively suppressing background clutter: mIoU reaches 23.80%, Pd achieves 92.47%, and FA is only 62.82×10^{-5}. The Pd outperforms the strongest single-frame method UIU-Net (90.53%) by 1.94 percentage points and the best multi-frame method RFR+ALCNet (74.32%) by 18.15 percentage points; the FA is reduced by over two orders of magnitude compared to UIU-Net (4296.88×10^{-5}) and by 7.9% compared to RFR+ALCNet (68.22×10^{-5}). These results indicate that the model effectively captures robust representations of small targets across different sizes within the small-target regime.

The proposed method exhibits consistently strong target-level performance (Pd $>$ 92% across both datasets with very low false alarm rates), demonstrating robust detection and localization capability for small targets of various sizes—even under severe scale distribution shifts and for zero-shot extremely small targets. Concurrently, the substantial pixel-level improvement (e.g., +27.96 percentage points in mIoU on Dataset 1) indicates markedly superior delineation of target shape and boundaries compared to existing approaches. Although pixel-level metrics naturally decline on the more challenging Dataset 2 due to extreme target scale and inherent ambiguity, the maintained high Pd, well-controlled FA, and still-competitive mIoU suggest that the model preserves strong semantic awareness of target presence while effectively mitigating over-segmentation and background clutter. Overall, this complementary behavior across target-level reliability and pixel-level precision underscores the effectiveness of the proposed dual-temporal representation learning strategy in achieving a balanced trade-off between detection robustness and fine-grained shape recovery for multi-size infrared small targets.

Across the complete SITP-QLSD dataset (Dataset 1 + Dataset 2, Table 4), the proposed method achieves a Pd of 93.51%, mIoU of 47.19%, and FA of 29.45×10^{-5}. The Pd surpasses the best single-frame method UIU-Net (89.74%) by 3.77 percentage points and the best multi-frame method RFR+ALCNet (74.28%) by 19.23 percentage points; mIoU improves by 9.66 percentage points (a relative improvement of 25.7%) over the strongest multi-frame RFR+ResUNet (37.53%) and by 45.29 percentage points over the highest single-frame UIU-Net (1.90%). These outcomes highlight the algorithm's exceptional robustness to approximately one order of magnitude variation in target size, validating its capability for precise detection of mixed-size small targets in complex high-resolution spaceborne infrared earth-facing scenarios, and providing dual assurance of high performance and engineering reliability for small-target early warning systems.

On Dataset 3 (Table 4), the proposed method achieves an mIoU of 36.72%, representing an improvement of 1.45 percentage points over the strongest multi-frame baseline RFR+DNANet (35.27%) and 6.29 percentage points over the highest single-frame method ALCNet (30.43%). The Pd reaches 95.14%, surpassing all compared methods (second-best single-frame: MSHNet at 95.07%, +0.07 percentage points; best multi-frame: RFR+DNANet at 94.88%, +0.26 percentage points). The FA is only 3.27×10^{-5}, ranking among the lowest for multi-frame methods. Although slightly higher than RFR+ALCNet (1.24×10^{-5}), given its Pd of only 93.90%, the proposed method achieves superior false alarm control under high Pd conditions.

Table 3. Test results on Dataset 2. The best-performing result is enclosed in a solid box, and the second-best result is enclosed in a dashed box.

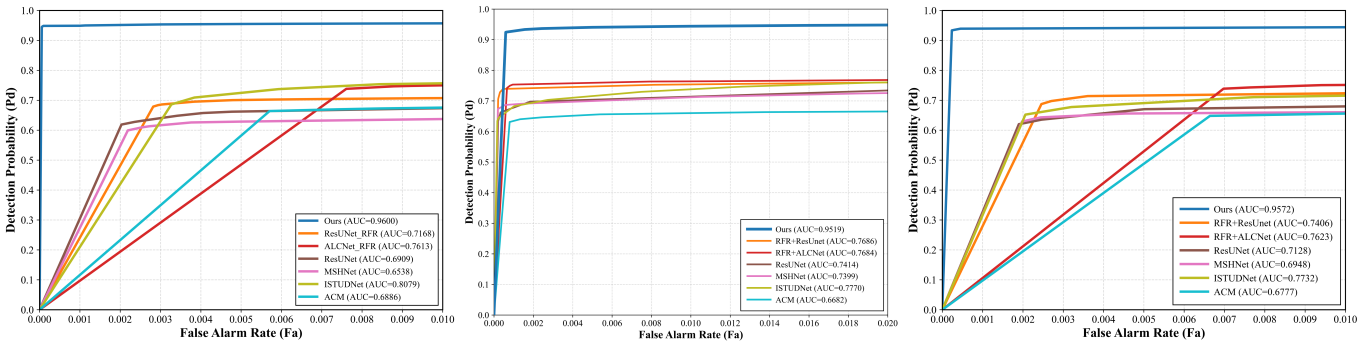
Frame	Methods	Pd (%)	FA ($\times 10^{-5}$)	mIoU (%)
Single Frame	ACM	63.99	132.27	15.41
	ALCNet	65.88	68.68	20.24
	DNANet	59.13	7.28	26.76
	ISTUD-UNet	64.44	24.27	22.97
	ResUNet	63.92	23.58	25.67
	UIU-Net	90.53	4296.88	4.29
	MSHNet	67.35	20.48	26.25
Multi Frame	RFR+ACM	72.34	35.67	24.10
	RFR+ALCNet	74.32	68.22	21.32
	RFR+DNANet	68.75	30.39	23.73
	RFR+ResUNet	70.95	22.60	22.63
	Ours	92.47	62.82	23.80

Notably, Dataset 3 features fixed target sizes with minimal scale diversity, in contrast to the pronounced mixed-size challenges in the main SITP-QLSD set. The consistent and strong performance gains observed here further highlight the effectiveness of the multi-scale IRFeatureExtractor in capturing robust representations even in near-distribution scenarios lacking significant scale variation. In particular, the adaptive gating pyramid deformable alignment module excels at enforcing precise temporal alignment and strong coherence across frames, effectively addressing platform-induced motion variations in staring mode. This results in the highest detection probability (Pd) paired with satisfactory false alarm levels, confirming the framework's outstanding stability, generalization capability, and seamless adaptability to diverse imaging modalities, including low-motion staring configurations.

Table 4. Test results on SITP-QLSD and Dataset 3. The best-performing result is enclosed in a solid box, and the second-best result is enclosed in a dashed box.

Frame	Methods	SITP-QLSD			Dataset 3		
		Pd (%)	FA ($\times 10^{-5}$)	mIoU (%)	Pd (%)	FA ($\times 10^{-5}$)	mIoU (%)
Single Frame	ACM	65.65	104.38	24.67	93.90	1.91	25.10
	ALCNet	66.54	71.13	28.67	94.50	2.60	30.43
	DNANet	60.42	14.88	37.67	94.85	2.36	29.42
	ISTUD-UNet	67.73	32.02	36.66	94.37	7.23	28.75
	ResUNet	63.52	24.63	36.27	93.56	1.28	30.84
	UIU-Net	89.74	4331.57	1.90	83.12	859.31	3.80
	MSHNet	64.29	25.71	35.71	95.07	2.43	28.57
Multiple Frame	RFR+ACM	70.91	38.24	34.85	93.41	3.73	30.10
	RFR+ALCNet	74.28	76.40	29.76	93.90	1.24	33.43
	RFR+DNANet	65.87	21.79	36.29	94.88	2.32	35.27
	RFR+ResUNet	69.74	27.09	37.53	94.60	2.14	31.79
	Ours	93.51	29.45	47.19	95.14	3.27	36.72

We plotted the ROC curves of various algorithms at different false alarm (FA) rates to comprehensively compare their detection performance. As shown in Figure 6, the proposed method's ROC curves on Dataset 1, Dataset 2, and the complete SITP-QLSD dataset lie at the top in nearly all FA intervals, consistently achieving the highest detection probability (Pd). This demonstrates that the method effectively suppresses complex background clutter while maintaining exceptionally high target detection rates, thereby achieving an optimal balance between Pd and FA.



(a) (b) (c)

Figure 6. ROC curves of various methods on SITP-QLSD datasets. (**a**) Dataset 1. (**b**) Dataset 2. (**c**) SITP-QLSD (Dataset 1 + Dataset 2).

In high-resolution spaceborne infrared earth-facing scenarios, the proposed method exhibits significant and consistent performance advantages across diverse conditions: typical medium-sized small targets (Dataset 1), extreme scale shifts to very small targets (Dataset 2), and the full mixed-size distribution. These results further validate the algorithm's strong robustness and generalization capability in real-world infrared small-target detection tasks, benefiting from its ability to capture multi-scale features that allow robust detection of small targets across different sizes, combined with the adaptive gating pyramid deformable alignment mechanism that compensates for motion and jitter effects.

3.4. Visualization Results

To further illustrate the performance differences among various methods across diverse and challenging scenarios, qualitative visualization results on the SITP-QLSD dataset are presented in Figure 7. The selected backgrounds consist of real high-resolution spaceborne infrared scenes acquired from the Qilu-2 satellite, encompassing a wide variety of complex terrestrial and marine environments, including rivers, lakes, urban areas, farmlands, harbors, mountains, airports, clouds, and open ocean surfaces. These heterogeneous and highly textured backgrounds generate strong clutter with spatial patterns that frequently resemble small targets, posing significant challenges to reliable infrared small-target detection, particularly across various target sizes.

Single-frame methods exhibit varying degrees of difficulty in suppressing background clutter while maintaining complete target detection. UIU-Net achieves relatively high detection rates across most target sizes and scenes but generates a substantially large number of false alarms (green circles), particularly in textured and structured backgrounds such as cloud edges, urban heat sources, and ocean wave patterns, which severely restricts its practical utility. MSHNet shows more balanced but still inconsistent performance, with both missed detections (blue circles) and false alarms appearing across different target sizes and clutter conditions.

Multi-frame baselines (RFR+ACM and RFR+ALCNet) benefit from temporal information but still encounter noticeable limitations in complex scenes. RFR+ACM tends to produce fewer false alarms in some cases but suffers from frequent missed detections (blue circles) across various target sizes. RFR+ALCNet detects more targets overall than RFR+ACM but exhibits a higher incidence of false alarms (green circles), especially in high-clutter regions. Overall, these methods display reduced robustness to size variations and background diversity compared to the proposed approach, consistent with their quantitative performance (Tables 2–4).

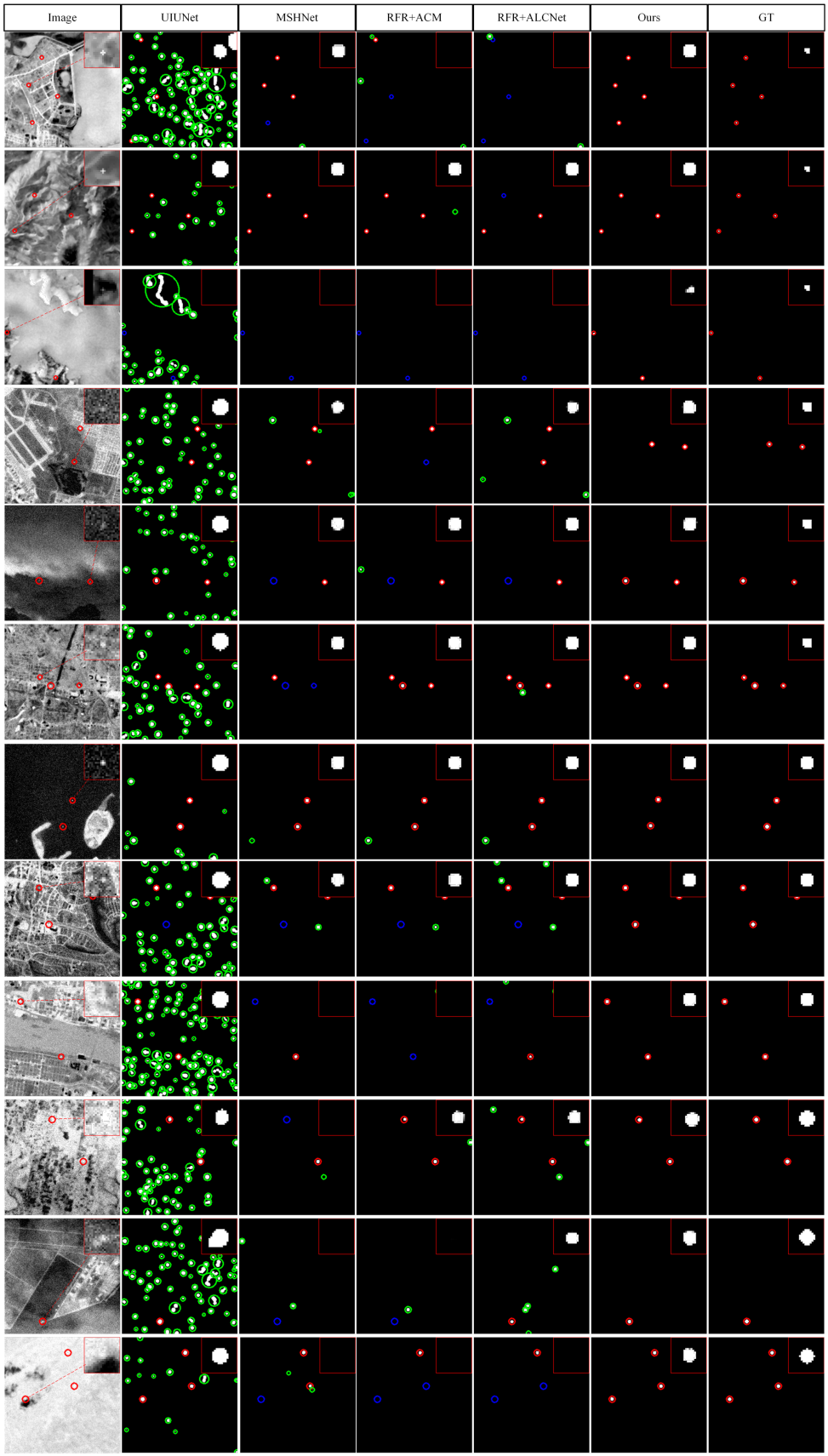


Figure 7. Visualization results of different algorithms on the SITP-QLSD dataset. Sub-figures are ordered from top to bottom by increasing zoomed genuine target size: 3 × 3, 5 × 5, 7 × 7, and 9 × 9 pixels (three per size). For better visualization, genuine targets are highlighted in red boxes in the upper-right corner. Red circles indicate accurately detected targets, blue circles represent missed detections, and green circles denote false alarms.

In contrast, the proposed method consistently delivers stable and accurate detection of multi-scale small targets across nearly all tested high-clutter scenes and size categories. Benefiting from the enhanced multi-scale context modeling in the IRFeatureExtractor and the adaptive gating offset-guided pyramid deformable alignment module, the method effectively captures the intrinsic spatiotemporal continuity of targets while strongly suppressing background interference and optimizing alignment effects over sequences. As a result, DTRFR markedly reduces both missed detections (blue circles) and false alarms (green circles) compared to the baselines, achieving high detection completeness together with excellent false alarm suppression in nearly all visualized cases.

These qualitative observations strongly corroborate the quantitative superiority of DTRFR reported earlier (e.g., highest Pd = 94.51% and lowest FA = 5.88 $\times$ 10^{-5} on Dataset 1, robust Pd = 92.47% on the challenging generalization sub-test set with extremely small targets). The visualization results further confirm the method's excellent robustness and generalization capability in realistic, high-resolution spaceborne infrared scenarios characterized by pronounced size diversity and complex earth-facing clutter.

3.5. Ablation Study

To thoroughly evaluate the contribution of each module to the overall detection performance, ablation experiments were conducted on Dataset 1 (featuring scale-diverse targets in complex spaceborne infrared scenes) and Dataset 3 (primarily single-sized targets). The results are summarized in Table 5, where module A refers to the enhanced IRFeatureExtractor and module B denotes the Adaptive Gating Pyramid Deformable Alignment. FLOPs and parameters are computed based on an input image sequence with a resolution of 10 $\times$ 256 $\times$ 256. All other metrics and experimental settings follow the hyperparameter configurations described earlier. The best performance in each column is enclosed in a solid box.

Table 5. Ablation study results. FLOPs and parameters are computed based on an input image sequence with a resolution of 10 $\times$ 256 $\times$ 256. The best performance in each column is enclosed in a solid box (highest mIoU and Pd, lowest FA). Here, $\times$ indicate that the corresponding module was removed, while ✓ indicate that the corresponding module was not removed in the ablation experiments.

A	B	FLOPs	Params	Dataset 1			Dataset 3		
		(G)	(M)	mIoU (%)	Pd (%)	FA ($\times 10^{-5}$)	mIoU (%)	Pd (%)	FA ($\times 10^{-5}$)
×	×	72.28	1.01	45.20	68.59	30.26	31.79	94.60	2.14
✓	×	96.07	1.04	71.93	95.14	11.87	35.51	94.22	1.41
×	✓	62.85	0.98	46.21	67.81	26.49	33.38	95.12	4.96
✓	✓	86.64	1.02	74.32	94.51	5.88	36.72	95.14	3.27

Activating only module A (✓ ×) yields excellent performance on the scale-diverse Dataset 1, achieving an mIoU of 71.93% and Pd of 95.14%, with 96.07 G FLOPs and 1.04 M parameters. Benefiting from serial-to-parallel multi-scale feature extraction and dilated receptive fields, module A significantly enhances global context modeling and captures cross-size semantic associations, enabling precise discrimination between targets and clutter in complex backgrounds. On Dataset 1 (single-sized targets), mIoU increases to 35.51%, FA is substantially reduced to 1.41 $\times$ 10^{-5}, and Pd remains high at 94.22% (a minor decrease of 0.92%). The complementary interactions among parallel convolutional branches effectively prevent feature submergence by large kernels, delivering robust background suppression and accuracy improvements even in uniform-scale scenarios, thus highlighting the module's strong multi-scale adaptability and clutter rejection capability.

In contrast, enabling only module B (× ✓) achieves greater overall robustness at lower computational cost (62.85 G FLOPs, a reduction of 9.22 G; 0.98 M parameters, down by 0.06 M). On Dataset 1, mIoU rises to 46.21% (gain of 1.00%), Pd slightly decreases by 0.78% to 67.81%, but FA is notably reduced. On Dataset 3, mIoU reaches 33.38% (gain of 2.02%), Pd improves to 95.12% (gain of 0.52%), with FA at 4.96×10^{-5}. The gating offset-guided pyramid deformable alignment mechanism effectively models inter-frame target motion and optimizes temporal feature alignment, delivering strong robustness across push-broom-mode systems with significant platform or background dynamics. In staring-mode scenarios (Dataset 3) characterized by minimal motion, module B further strengthens auxiliary temporal consistency, resulting in notably higher detection probability (Pd) and mIoU compared to module A alone, thereby contributing to enhanced overall performance across diverse imaging conditions.

The full model (✓ ✓), integrating both modules, delivers the optimal balance of accuracy and robustness: mIoU of 74.32% and Pd of 94.51% on Dataset 1, and mIoU of 36.72%, Pd of 95.14%, and FA as low as 3.27×10^{-5} on Dataset 3. The relatively modest gains on Dataset 3 arise primarily from its predominantly single-scale target distribution, which limits the full exploitation of module A's multi-scale features. Nevertheless, the complete model consistently outperforms all ablated variants across both datasets. Module A provides robust multi-scale features, while module B addresses temporal coherence and helps mitigate noise in dynamic settings. Their complementary synergy achieves superior overall detection precision, robustness, and efficiency, validating the effectiveness of the proposed dual-module architecture for high-resolution spaceborne infrared small target detection.

To distinguish the contribution of non-linear gating from plain linear channel weighting in the Adaptive gating layer, we perform a targeted ablation by removing the `tanh` non-linearity in Table 6. The best performance in each column is enclosed in a solid box. Specifically, we compare the following two configurations:

- Full model (only weight factor): The `tanh` operation is removed, reducing the gating layer to a simple linear channel-wise weighting $x_2 = x_1 \odot \gamma$.
- Full model: The complete model using $x_2 = x_1 \odot \tanh(\gamma)$, which incorporates the proposed bounded non-linear gating mechanism.

Table 6. Ablation on the tanh non-linearity in the adaptive gating layer (Dataset 1). Removing tanh reduces the mechanism to channel weighting. The best performance in each column is enclosed in a solid box. (highest mIoU and Pd, lowest FA).

Method	Pd (%)	FA ($\times 10^{-5}$)	mIoU (%)
Full model (only weight factor γ)	94.02	11.14	70.61
Full model	94.51	5.88	74.32

Pd increases by 0.49% (from 94.02% to 94.51%), FA decreases substantially by approximately 89.5% (from 11.14 to 5.88×10^{-5}), and mIoU rises by 3.71% (from 70.61% to 74.32%). These declines demonstrate that the bounded non-linear gating mechanism of $\texttt{tanh}(\gamma)$ is essential for enabling effective channel-wise gating. In contrast, relying solely on a weight factor alone lacks boundedness and non-linearity control, leading to poor performance. This ablation confirms that the gating effect primarily stems from the non-linear transformation $\texttt{tanh}(\gamma)$, rather than from the learnable parameter γ functioning merely as an unconstrained scaling factor.

4. Discussion

Although the proposed method achieves state-of-the-art performance on the SITP-QLSD dataset (mIoU of 47.19%, Pd of 93.51%, FA of 29.45×10^{-5}), several aspects remain open for improvement, reflecting inherent trade-offs in high-resolution spaceborne infrared small target detection.

1. Idealized target modeling

The Gaussian PSF used for target simulation captures the core diffusive physics but overlooks real payload details (e.g., non-uniform radiation, edge blooming, atmospheric effects). This may cause minor sim-to-real mismatches. However, the learned multi-scale features enable robust detection of targets of different sizes, maintaining Pd = 92.47% on the generalization sub-test set despite distribution shift.

2. Modest gains on near-distribution small targets

In Dataset 3 (single-sized targets aligned with the training distribution), multi-scale mechanisms yield limited improvement (mIoU +4.92%, Pd +0.54% over the baselines). This is because the model prioritizes scale-agnostic features rather than size-specific features. On the generalization sub-test set, high mIoU values across methods often stem from the predicted diffusion blobs being slightly larger than ground-truth point annotations, overestimating overlap while preserving high detection completeness.

The core advantage of the proposed method lies in learning multi-scale features, ensuring reliable detection across different target sizes even under severe distribution shifts.

To further advance the proposed method, future work could integrate realistic PSF measurements and sim-to-real adaptation techniques to bridge simulation-reality gaps, add a lightweight coarse-to-fine refinement head to correct blob inflation and enhance boundary precision for tiny targets, and introduce scale-adaptive weighting or dedicated small-target branches to achieve finer performance in near-distribution scenarios without compromising cross-size robustness. Finally, we plan to systematically investigate the potential of preliminary denoising as a preprocessing step to further improve performance under low SNR conditions, and evaluate its quantitative impact on both the proposed method and existing approaches in noisy real-world sequences.

5. Conclusions

In summary, this paper addresses the urgent need for diverse small-target detection in high-spatial-resolution spaceborne infrared earth-facing scenarios by proposing DTRFR, an end-to-end unified detection framework that integrates multi-scale compatibility with sequence robustness. By adopting an approach that emphasizes multi-scale feature learning, the proposed method effectively alleviates key limitations of existing detectors—such as size sensitivity and suboptimal alignment in sequences—while achieving strong performance on the realistic SITP-QLSD dataset.

The main contributions and findings are as follows:

1. A realistic SITP-QLSD dataset is constructed from QLSAT-2 infrared backgrounds, featuring diverse scenes, mixed-size small targets, and a dedicated generalization sub-test set with extremely small targets, providing a reliable benchmark for evaluating size-diverse and generalization detection in complex spaceborne scenarios.
2. The multi-scale IRFeatureExtractor module, leveraging serial-to-parallel convolutions and dilated receptive fields, effectively enhances cross-scale feature representation and clutter suppression, enabling accurate target discrimination across different sizes.
3. The adaptive gating pyramid deformable alignment mechanism optimizes multi-frame feature alignment through adaptive gating modulation, enhancing temporal coherence and delivering superior overall robustness.

4. Extensive experiments demonstrate superior detection accuracy and false alarm suppression compared to single-frame and multi-frame baselines, with mIoU of 74.32% and Pd of 94.51% on the main set, and robust Pd of 92.37% on the generalization sub-test set. The core strength lies in learning multi-scale features, which enables reliable detection of extremely small targets under severe distribution shifts.

These results validate the effectiveness of the proposed framework and its design in achieving reliable, size-robust detection for high-resolution spaceborne infrared small target applications.

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

Funding: This research was funded by the China Shanghai Municipal 2022 "Science and Technology Innovation Action Plan" Outstanding Academic/Technical Leader Program Projects: 22XD1404100; National Natural Science Foundation of China: 62175251; Talent Plan of Shanghai Branch, Chinese Academy of Sciences: CASSHB-QNPD-2023-007; the CAS Project for Young Scientists in Basic Research: YSBR-113; Shanghai Leading Talent Program of Eastern Talent Plan: QNKJ2024003.

Data Availability Statement: The SITP-QLSD dataset is available at https://github.com/silverphoebus7-cell/SITP-QLSD (accessed on 23 February 2026).

Acknowledgments: The authors thank the National Key Laboratory of Infrared Detection Technologies and Shanghai Institute of Technical Physics for providing computational resources and satellite data access. During the preparation of this manuscript, the authors used AI-assisted language editing tools for language editing, text refinement, and assistance. The authors have reviewed and edited all outputs and take full responsibility for the content of this publication.

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

References

1. Sun, Y.; Yang, J.; An, W. Infrared Dim and Small Target Detection via Multiple Subspace Learning and Spatial-Temporal Patch-Tensor Model. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 3737–3752. [CrossRef]
2. 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.* **2021**, *59*, 6077–6091. [CrossRef]
3. Wu, P.; Huang, H.; Qian, H.; Su, S.; Sun, B.; Zuo, Z. SRCANet: Stacked Residual Coordinate Attention Network for Infrared Ship Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5003614. [CrossRef]
4. Cui, Y.; Lei, T.; Chen, G.; Zhang, Y.; Zhang, G.; Hao, X. Infrared Small Target Detection via Modified Fast Saliency and Weighted Guided Image Filtering. *Sensors* **2025**, *25*, 4405. [CrossRef]
5. Driggers, R.; Pollak, E.; Grimming, R.; Velazquez, E.; Short, R.; Holst, G.; Furxhi, O. Detection of Small Targets in the Infrared: An Infrared Search and Track Tutorial. *Appl. Opt.* **2021**, *60*, 4762–4777. [CrossRef]
6. Guo, L.; Rao, P.; Gao, C.; Su, Y.; Li, F.; Chen, X. Adaptive Differential Event Detection for Space-Based Infrared Aerial Targets. *Remote Sens.* **2025**, *17*, 845. [CrossRef]
7. He, H.; Wan, M.; Xu, Y.; Kong, X.; Liu, Z.; Chen, Q.; Gu, G. WTAPNet: Wavelet Transform-Based Augmented Perception Network for Infrared Small-Target Detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5037217. [CrossRef]
8. Wang, Y.; Cao, L.; Su, K.; Dai, D.; Li, N.; Wu, D. Infrared Moving Small Target Detection Based on Space–Time Combination in Complex Scenes. *Remote Sens.* **2023**, *15*, 5380. [CrossRef]
9. Parry, I.; Hawker, G.; Gomez-Jenkins, M.; Goncalves, M.; Ang, E.; Barkhuysen, R.; Desborough, P.; Donaghy, J.; Dovhalenko, T.; Gonzalez, S.; et al. Innovative Technologies for Very-High-Resolution MWIR and LWIR Earth Observations. In *Small Satellites Systems and Services Symposium (4S 2024)*; SPIE: Bellingham, WA, USA, 2025; Volume 13546, pp. 632–642. [CrossRef]
10. Fevgas, G.; Lagkas, T.; Argyriou, V.; Sarigiannidis, P. New vegetation stress assessment approach via WorldView-3 imagery, validated with UAV thermal imaging. *Int. J. Remote Sens.* **2025**, *46*, 4764–4780. [CrossRef]

11. Lin, M.; Jin, M.; Li, J.; Bai, Y. GEOSatDB: Global Civil Earth Observation Satellite Semantic Database. *Big Earth Data* **2024**, *8*, 522–539. [CrossRef]
12. Chapple, P.B.; Bertilone, D.C.; Caprari, R.S.; Angeli, S.; Newsam, G.N. Target Detection in Infrared and SAR Terrain Images Using a Non-Gaussian Stochastic Model. In *Targets and Backgrounds: Characterization and Representation V*; SPIE: Bellingham, WA, USA, 1999; Volume 3699, pp. 122–132. [CrossRef]
13. 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]
14. Arce, G.; McLoughlin, M. Theoretical Analysis of the Max/Median Filter. *IEEE Trans. Acoust. Speech Signal Process.* **1987**, *35*, 60–69. [CrossRef]
15. Chen, T.; Wu, Q.H.; Rahmani-Torkaman, R.; Hughes, J. A pseudo top-hat mathematical morphological approach to edge detection in dark regions. *Pattern Recognit.* **2002**, *35*, 199–210. [CrossRef]
16. Bai, X.; Zhou, F. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognit.* **2010**, *43*, 2145–2156. [CrossRef]
17. Strickland, R.N.; Hahn, H.I. Wavelet transform methods for object detection and recovery. *IEEE Trans. Image Process.* **1997**, *6*, 724–735. [CrossRef]
18. Qi, S.; Ma, J.; Li, H.; Zhang, S.; Tian, J. Infrared small target enhancement via phase spectrum of quaternion Fourier transform. *Infrared Phys. Technol.* **2014**, *62*, 50–58. [CrossRef]
19. Ren, K.; Song, C.; Miao, X.; Wan, M.; Xiao, J.; Gu, G.; Chen, Q. Infrared small target detection based on non-subsampled shearlet transform and phase spectrum of quaternion Fourier transform. *Opt. Quantum Electron.* **2020**, *52*, 168. [CrossRef]
20. Xu, Y.; Shao, A.; Kong, X.; Wu, J.; Chen, Q.; Gu, G.; Wan, M. Infrared small target detection based on sub-maximum filtering and local intensity weighted gradient measure. *IEEE Sens. J.* **2024**, *24*, 22236–22248. [CrossRef]
21. Han, J.; Liang, K.; Zhou, B.; Zhu, X.; Zhao, J.; Zhao, L. Infrared small target detection utilizing the multiscale relative local contrast measure. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 612–616. [CrossRef]
22. Guan, X.; Peng, Z.; Huang, S.; Chen, Y. Gaussian scale-space enhanced local contrast measure for small infrared target detection. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 327–331. [CrossRef]
23. Chen, L.; Rao, P.; Chen, X. Infrared dim target detection method based on local feature contrast and energy concentration degree. *Optik* **2021**, *248*, 167651. [CrossRef]
24. Chen, Z.; Luo, S.; Xie, T.; Liu, J.; Wang, G.; Lei, G. A novel infrared small target detection method based on BEMD and local inverse entropy. *Infrared Phys. Technol.* **2014**, *66*, 114–124. [CrossRef]
25. 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] [PubMed]
26. Dai, Y.; Wu, Y.; Song, Y. Infrared small target and background separation via column-wise weighted robust principal component analysis. *Infrared Phys. Technol.* **2016**, *77*, 421–430. [CrossRef]
27. Dai, Y.; Wu, Y.; Song, Y.; Guo, J. Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values. *Infrared Phys. Technol.* **2017**, *81*, 182–194. [CrossRef]
28. Zhang, L.; Peng, L.; Zhang, T.; Cao, S.; Peng, Z. Infrared small target detection via non-convex rank approximation minimization joint l2,1 norm. *Remote Sens.* **2018**, *10*, 1821. [CrossRef]
29. Dai, Y.; Wu, Y. Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3752–3767. [CrossRef]
30. 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 (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 867–876. [CrossRef]
31. Wu, T.; Li, B.; Luo, Y.; Wang, Y.; Xiao, C.; Liu, T.; Yang, J.; An, W.; Guo, Y. MTU-Net: Multilevel TransUNet for space-based infrared tiny ship detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5601015. [CrossRef]
32. Jiang, N.; Wang, K.; Peng, X.; Yu, X.; Wang, Q.; Xing, J.; Li, G.; Guo, G.; Ye, Q.; Jiao, J.; et al. Anti-UAV: A large-scale benchmark for vision-based UAV tracking. *IEEE Trans. Multimed.* **2023**, *25*, 486–500. [CrossRef]
33. Sun, H.; Bai, J.; Yang, F.; Bai, X. Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset IRDST. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5000513. [CrossRef]
34. Wang, H.; Zhou, L.; Wang, L. Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 29 October–2 November 2019; pp. 8508–8517. [CrossRef]
35. Zhao, B.; Wang, C.; Fu, Q.; Han, Z. A novel pattern for infrared small target detection with generative adversarial network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4481–4492. [CrossRef]

36. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 4 May 2021. Available online: https://openreview.net/forum?id=YicbFdNTTy (accessed on 23 February 2026).
37. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2024**, arXiv:2312.00752. [CrossRef]
38. Chen, T.; Ye, Z.; Tan, Z.; Gong, T.; Wu, Y.; Chu, Q.; Liu, B.; Yu, N.; Ye, J. MiM-ISTD: Mamba-in-Mamba for efficient infrared small target detection. *arXiv* **2024**, arXiv:2403.02148. [CrossRef]
39. Wang, Y.; Xu, Z.; Wang, X.; Shen, C.; Cheng, B.; Shen, H.; Xia, H. End-to-end video instance segmentation with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 8741–8750.
40. Chen, S.; Ji, L.; Zhu, J.; Ye, M.; Yao, X. SSTNet: Sliced spatio-temporal network with cross-slice ConvLSTM for moving infrared dim-small target detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5000912. [CrossRef]
41. Li, J.; Liu, P.; Huang, X.; Cui, W.; Zhang, T. Learning motion constraint-based spatio-temporal networks for infrared dim target detections. *Appl. Sci.* **2022**, *12*, 11519. [CrossRef]
42. Yan, P.; Hou, R.; Duan, X.; Yue, C.; Wang, X.; Cao, X. STDMANet: Spatio-temporal differential multiscale attention network for small moving infrared target detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5602516. [CrossRef]
43. Wang, P.; Niu, W.; Gao, W.; Guo, Y.; Peng, X. Dim moving point target detection in cloud clutter scenes based on temporal profile learning. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6006905. [CrossRef]
44. 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*, 5001819. [CrossRef]
45. Zhang, Z.; Shao, F.; Dai, Z.; Zhu, S. Towards robust video instance segmentation with temporal-aware transformer. *arXiv* **2023**, arXiv:2301.09416. [CrossRef]
46. Karim, R.; Zhao, H.; Wildes, R.P.; Siam, M. MED-VT++: Unifying multimodal learning with a multiscale encoder-decoder video transformer. *arXiv* **2024**, arXiv:2304.05930. [CrossRef]
47. Huang, Y.; Zhi, X.; Hu, J.; Yu, L.; Han, Q.; Chen, W.; Zhang, W. LMAFormer: Local motion aware transformer for small moving infrared target detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5008117. [CrossRef]
48. Li, R.; An, W.; Xiao, C.; Li, B.; Wang, Y.; Li, M.; Guo, Y. Direction-coded temporal U-shape module for multiframe infrared small target detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 555–568. [CrossRef] [PubMed]
49. Ying, X.; Liu, L.; Lin, Z.; Shi, Y.; Wang, Y.; Li, R.; Cao, X.; Li, B.; Zhou, S.; An, W. Infrared small target detection in satellite videos: A new dataset and a novel recurrent feature refinement framework. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5002818. [CrossRef]
50. Deng, C.; Guo, Y.; Xu, X.; Zhao, Z.; Xia, Y.; An, R.; Li, J.; Plaza, A. Learning Global Dynamic Query for Large-Motion Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.***2026**, *64*, 5002016. [CrossRef]
51. Li, F.; Rao, P.; Sun, W.; Su, Y.; Chen, X. A New Motion Feature-Enhanced Multiframe Spatial–Temporal Infrared Target Detection Network. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5006819. [CrossRef]
52. Li, F.; Xu, Q.; Bao, S.; Yang, Z.; Cong, R.; Cao, X.; Huang, Q. Size-invariance matters: Rethinking metrics and losses for imbalanced multi-object salient object detection. *arXiv* **2024**, arXiv:2405.09782. [CrossRef]
53. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
54. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
55. Sun, H.; Hao, X.; Wang, J.; Pan, B.; Pei, P.; Tai, B.; Zhao, Y.; Feng, S. Flame edge detection method based on a convolutional neural network. *ACS Omega* **2022**, *7*, 26680–26686. [CrossRef]
56. Hu, J.; Shen, L.; Sun, G.; Albanie, S. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
57. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Virtual, 5–9 January 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 1069–1078. [CrossRef]
58. 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]
59. Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; Guo, Y. Dense nested attention network for infrared small target detection. *IEEE Trans. Image Process.* **2023**, *32*, 972–986. [CrossRef] [PubMed]
60. Hou, Q.; Zhang, L.; Tan, F.; Xi, Y.; Zheng, H.; Li, N. Istdu-net: Infrared small-target detection u-net. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 7506205. [CrossRef]

61. Xiao, X.; Lian, S.; Luo, Z.; Li, S. Weighted res-unet for high-quality retina vessel segmentation. In Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education (ITME), Hangzhou, China, 19–21 October 2018; pp. 327–331. [CrossRef]
62. 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. [CrossRef] [PubMed]
63. Liu, Q.; Liu, R.; Zheng, B.; Wang, H.; Fu, Y. Infrared small target detection with scale and location sensitivity. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 17489–17498. [CrossRef]

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