*remote sensing*

MDPI

*Article*

# Spatio-Temporal Residual Attention Network for Satellite-Based Infrared Small Target Detection

**Yan Chang, Decao Ma *, Qisong Yang, Shaopeng Li and Daqiao Zhang**

PLA Rocket Force University of Engineering, Xi'an 710025, China; cyan313@outlook.com (Y.C.); qisong.yang.93@outlook.com (Q.Y.); spli16@mails.tsinghua.edu.cn (S.L.); zhangdq2012@pku.org.cn (D.Z.)
* Correspondence: madecaoedu@163.com

## Highlights

**What are the main findings?**

- A spatio-temporal detection framework is proposed for infrared small target detection in satellite video, which combines inter-frame residuals with spatial and temporal feature learning.
- The proposed method achieves superior detection accuracy and robustness compared with state-of-the-art approaches, particularly for tiny and dim targets in complex backgrounds.

**What is the implication of the main finding?**

- The framework provides an effective solution for detecting small moving aerial targets from satellite infrared video, supporting reliable long-range monitoring.
- This study demonstrates the potential of integrating temporal consistency and multi-scale spatial features to advance real-world remote sensing applications.

## Abstract

With the development of infrared remote sensing technology and the deployment of satellite constellations, infrared video from orbital platforms is playing an increasingly important role in airborne target surveillance. However, due to the limitations of remote sensing imaging, the aerial targets in such videos are often small in scale, low in contrast, and slow in movement, making them difficult to detect in complex backgrounds. In this paper, we propose a novel detection network that integrates inter-frame residual guidance with spatio-temporal feature enhancement to address the challenge of small object detection in infrared satellite video. This method first extracts residual features to highlight motion-sensitive regions, then uses a dual-branch structure to encode spatial semantics and temporal evolution, and then fuses them deeply through a multi-scale feature enhancement module. Extensive experiments show that this method outperforms mainstream methods in terms on various infrared small target video datasets, and has good robustness under low-signal-to-noise-ratio conditions.

**Keywords:** infrared video; satellite remote sensing; small object detection; inter-frame residual; spatio-temporal feature fusion



Academic Editor: Shuying Li

Received: 20 August 2025
Revised: 9 October 2025
Accepted: 10 October 2025
Published: 16 October 2025

**Citation:** Chang, Y.; Ma, D.; Yang, Q.; Li, S.; Zhang, D. Spatio-Temporal Residual Attention Network for Satellite-Based Infrared Small Target Detection. *Remote Sens.* **2025**, *17*, 3457. https://doi.org/10.3390/rs17203457



## 1. Introduction

With the increasing demand for global situation awareness and continuous wide-area surveillance, space-based infrared imaging systems have become a key technology for detecting and tracking airborne objects [1–3]. Compared with ground-based sensors,

infrared cameras mounted on satellites provide a wider field of view and offer long-term continuous monitoring. In particular, infrared satellite video can capture a large area of dynamic scenes, making it a promising solution for detecting maneuvering air targets (such as aircraft and other fast-moving platforms) [4]. However, due to the long viewing distance, these moving objects usually occupy only a few pixels in the image sequence, and often show low contrast compared with the clutter background, such as cloud, terrain, or atmospheric noise. These challenges have brought great difficulties to the traditional target detection algorithm, which requires a powerful solution tailored to the characteristics of infrared video data collected from space [5,6].

In recent years, there has been a growing interest in infrared small target detection [7,8]. The preliminary research named RLCM [9] is a multi-scale detection algorithm that uses the relative local contrast measure. ISNet [10] devises a Taylor finite difference (TFD)-inspired edge block and a two-orientation attention aggregation block to detect the precise shape information of infrared targets. Liu et al. [11] focused on boosting detection performance with a more effective loss but a simpler model structure by proposing a novel scale and location-sensitive loss to handle the limitations of existing losses. IAANet [12] introduces a coarse-to-fine interior attention-aware network for infrared small target detection. IRSAM [13] improves the encoder–decoder architecture to represent infrared small objects better.

Despite significant progress in infrared small target detection, existing methods still face serious limitations when applied to satellite-based video data. Many traditional methods rely heavily on single-frame spatial features, which are often not enough to distinguish very small or low contrast objects from complex backgrounds and noise [14]. In addition, the methods using time information tend to process consecutive frames independently or simply apply optical flow estimation, which may not be able to capture the subtle motion patterns of slow-moving or maneuvering targets. In addition, the multi-frame fusion technology usually has the problems of information redundancy or insufficient alignment, which leads to a decline in detection accuracy. Due to remote observation, severe atmospheric distortion, and low signal-to-noise ratio, these challenges are exacerbated in satellite infrared images, highlighting the need for more effective spatio-temporal feature extraction and fusion strategies customized for this unique application scenario [15].

To address the aforementioned challenges, this paper proposes a framework for infrared small target detection in satellite video sequences, as illustrated in Figure 1. The method integrates the inter-frame residual extraction with the dual-branch spatio-temporal feature fusion network, which effectively enhances the subtle motion cues of airborne small targets and captures the rich spatial background. A multi-scale feature enhancement module is designed to fuse the spatial and temporal information at different resolutions, and then a customized detection head is introduced for precise localization and classification. Extensive experiments conducted on public infrared video datasets show that the proposed method significantly improves detection accuracy and robustness in complex environments. The main contributions of this work are summarized as follows:

- We introduce an inter-frame residual module to explicitly highlight motion-related features and enhance the sensitivity of the network to subtle target motion in remote infrared satellite images.
- We design a dual-branch structure to encode spatial semantics and temporal evolution separately, which achieves more effective spatio-temporal feature fusion and reduces information redundancy.
- A multi-scale fusion strategy combined with a custom detection head is proposed to improve the detection performance of small low-contrast targets in complex backgrounds.

The remainder of this article is organized as follows: Section 2 briefly presents the related background on infrared small target detection and video-based target tracking. In Section 3, we formalize the research problem of infrared target detection, and then the details of our framework are described. Section 4 shows extensive experiments that validate the effectiveness of the proposed method. Finally, we conclude with a discussion of our framework and summarize future work in Section 5.

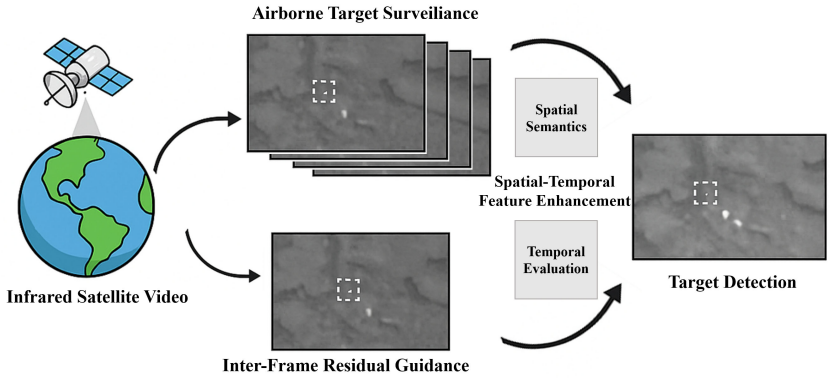



**Figure 1.** An illustration of the infrared small target detection problem with spatio-temporal information.

## 2. Related Work

### *2.1. Infrared Small Target Detection*

Infrared small target detection has important applications in military reconnaissance, search and rescue, environmental monitoring, and other fields. Different from traditional target detection tasks, infrared small targets have the characteristics of limited pixel occupation, low contrast in complex background, and weak signal-to-noise ratio [16,17]. These unique characteristics make the detection of small infrared targets particularly challenging. Over the years, many algorithms have been developed to solve these problems, from the traditional image processing technology, which focuses on contrast enhancement and clutter suppression, to the method based on deep learning, which aims to directly learn the discriminant features from the data. Although some progress has been made, the effective detection of small targets in infrared images is still an active and difficult research field.

Infrared small target detection has been applied to many scenes and has provided excellent performance. Hou et al. [18] calculated the likelihood map at first, where the pixel value represents the probability that the pixel belongs to a small target or background, and then applied a threshold to the likelihood map to extract real targets. Tong et al. [19] attempted to integrate the edge details and global contextual information of the target to improve IRSTD tasks. This method consists of a spatial pyramid pooling module and a dual-attention module, which focus on the global contextual information and the regions of interest, respectively. It increases the information exchange between feature maps using multi-scale feature fusion as well. Dai et al. [20] proposed a new label assignment scheme called all-scale pseudobox, which decouples the ground truth target size from the spatial assignment by using scale-adaptive pseudoboxes and also relaxes the scale constraints by treating all target boxes at all scales as positive samples. Li et al. [21] proposed a specialized network for hyperspectral point object detection, which uses a self-excited subpixel-scale attention module and achieves subpixel-scale deformable sampling while enabling self-excited amplification of object features.

Although existing infrared small target detection methods have made some progress, there are still some challenges to be solved. Existing methods mainly focus on single-frame spatial information, and cannot effectively use time dynamics. When the target shows subtle motion or is embedded in a chaotic background, it will lead to missed detection or false

alarm. The difficulty of distinguishing small targets from background clutter and sensitivity to noise further affect the detection accuracy. The assumption of target characteristics also limits their adaptability to various complex scenes. Therefore, there is an urgent need for a robust framework that can integrate spatio-temporal cues to enhance the detection of small and low-contrast objects in challenging infrared environments.

### *2.2. Video-Based Object Tracking*

Object tracking in consecutive frames is a key task in computer vision and remote sensing, which aims to locate a moving object between consecutive frames continuously [22–24]. In infrared surveillance, video sequences provide valuable temporal information that can be used to improve detection and tracking performance, especially for small and low-contrast targets. Unlike single-image detection, video-based tracking uses motion cues, temporal consistency, and dynamic context to distinguish objects from clutter and noise. This time dimension is important for detecting slow-moving or maneuvering objects that may not be important in a single frame. Therefore, effective video-based tracking methods are essential for applications such as airborne target monitoring.

Video-based object tracking can now be modeled as a supervised machine-learning problem due to the availability of publicly accessible datasets, such as [25–27]. Wan et al. [28] approached the MOT problem from a different perspective by directly obtaining the embedded spatio-temporal information of trajectories from raw video data. Chen et al. [29] proposed a historical-model-based tracker intended for satellite videos to improve the performance of the object tracking algorithm. Othmani et al. [30] presented a vehicle detection and tracking method for traffic video analysis based on deep learning technology. Ibrahim et al. [31] introduced deep online real-time tracking on thermal video-based online multi-object tracking in occlusion and thermal crossover scenes. Zhao et al. [32] proposed an adaptive diffusion timestep selection mechanism guided by visual complexity. Liu et al. [33] proposed an event camera calibration method utilizing a collimator with flickering star-based patterns, which first linearly solves camera parameters using the sphere motion model of the collimator, followed by nonlinear optimization to refine these parameters with high precision. Huang et al. [34] proposed a fusion localization method based on ridge estimation, combining the advantages of rich scene information from sequential imagery with the high precision of laser ranging to enhance localization accuracy.

Although considerable progress has been made in video-based target tracking methods, there are still some challenges in practical applications, especially for small infrared targets in satellite images [35–37]. Many methods rely on accurate motion estimation techniques, such as optical flow, which are unreliable in low-resolution, noisy, or cluttered infrared videos. Existing spatio-temporal feature fusion strategies may have redundancy or misalignment, resulting in decreased tracking accuracy and increased false alarm rate. Addressing these limitations requires the development of robust spatio-temporal representations and adaptive fusion mechanisms for the unique characteristics of IR satellite video sequences.

## 3. Proposed Algorithm

### *3.1. Problem Definition*

The objective of this work is to detect and localize small airborne targets in a sequence of satellite-based infrared video frames. Formally, given an input video sequence $\mathbf{I} = \{I_t\}_{t=1}^{T}$, where $I_t \in \mathbb{R}^{H \times W}$ represents the infrared image frame captured at time $t$, the goal is to identify a set of target bounding boxes $\mathcal{B}_t = \{b_t^i\}_{i=1}^{N_t}$ in each frame. Here, $N_t$ denotes the number of targets present at time $t$, and each bounding box $b_t^i = (x, y, w, h)$ is defined by its center coordinates $(x, y)$, width $w$, and height $h$.

The main challenge lies in the fact that these targets are tiny and usually occupy only a few pixels within a frame. They have a low contrast with the typically complex and noisy background of satellite infrared images. In addition, due to atmospheric interference and sensor noise, the appearance of the target may change in illumination or viewing angles over time. Long-distance observation can also lead to low spatial resolution and a weak signal-to-noise ratio, making precise detection more complex.

To address these challenges, our method utilizes the inherent temporal continuity of video sequences to model spatial and temporal features, thereby enhancing the representation of small moving targets. It requires the effective extraction of clues related to motion and the robust fusion of multi-frame information, which will be elaborated in detail in subsequent sections.

### *3.2. Overall Framework*

The proposed framework robustly detects small airborne targets from satellite-based infrared videos by integrating spatial and temporal cues in a unified architecture, as shown in Figure 2. Given a video sequence $\mathbf{I} = \{I_t\}_{t=1}^{T}$, the system processes the data in four stages: inter-frame motion enhancement, spatial feature extraction, temporal modeling, and detection based on multi-scale fusion.

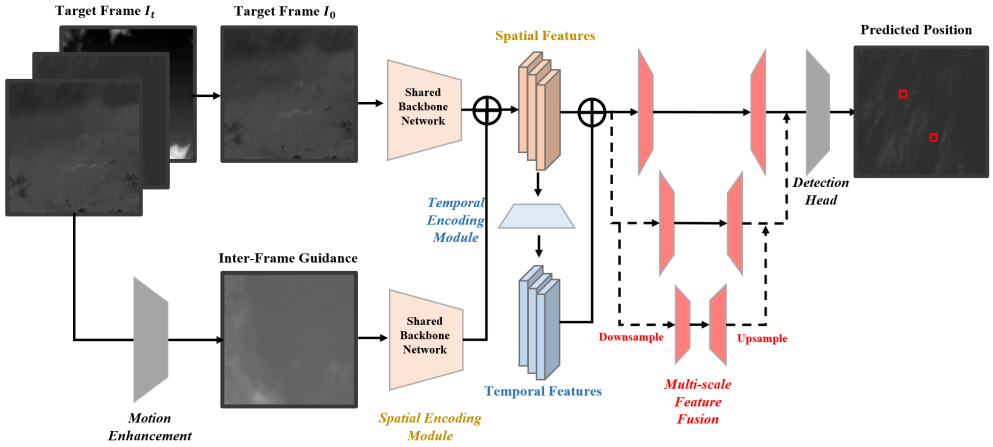



**Figure 2.** An overview of the spatio-temporal infrared small target detection framework.

The inter-frame motion information between adjacent frames is calculated to highlight the moving area while suppressing the static background. These motion cues, which serve as supplementary signals to the original frame, contribute to distinguishing targets that might not be distinguishable in a single-frame context. This residual motion cue provides dynamic information critical for distinguishing targets that may not be prominent in static spatial features.

The original frames and residual mappings are processed through the shared backbone network to extract spatial features. Furthermore, to simulate the consistency of the target and its cross-temporal movement, the time coding module aggregates multiple frame features, learns the movement trend and temporal correlation, enhances the signal of the real target, and simultaneously filters out noise and false alarms.

spatio-temporal features are then fused in a multi-scale manner to preserve details and coarse information. The fused features are passed to a customized detection head that performs precise localization and classification of targets. The detection outputs are then refined to produce the final detection results. The pseudo code of the entire procedure is given in Algorithm 1. This overall architecture effectively addresses the limitations of single-frame detection in complex infrared satellite video environments.

**Algorithm 1** Spatio-Temporal Infrared Small Target Detection Framework

**Input:** Infrared video sequence $\mathbf{I} = \{I_t\}_{t=1}^{T}$
**Output:** Detected target bounding boxes $\hat{\mathcal{B}} = \{\hat{\mathcal{B}}_t\}_{t=1}^{T}$
Initialize empty detection results: $\hat{\mathcal{B}} \leftarrow \varnothing$
**for** $t = 2$ to T do **do**
  Compute inter-frame motion enhancement: $R_t \leftarrow \mathcal{H}_r|I_t - I_{t-1}|$
  Extract spatial features from $I_t$ and $R_t$:
    $F^{I}_{spatial} \leftarrow Backbone(I_t)$
    $F^{R}_{spatial} \leftarrow Backbone(R_t)$
  Fuse spatial features:
    $F_{spatial} \leftarrow Fuse(F^{I}_{spatial}, F^{R}_{spatial})$
  Aggregate temporal features over window $[t-k, t]$:
    $F_{temporal} \leftarrow Temporal\ Encoder(F^{t-k}_{spatial}, ..., F^{t}_{spatial})$
  Perform multi-scale feature fusion:
    $F_{fused} \leftarrow MultiScaleFusion(F_{spatial}, F_{temporal})$
  Detect targets from fused features:
    $\hat{\mathcal{B}}_t \leftarrow DetectionHead(F_{fused})$
  Append $\hat{\mathcal{B}}_t$ to results: $\hat{\mathcal{B}} = \hat{\mathcal{B}} \cap \hat{\mathcal{B}}_t$
**end for**

### *3.3. Inter-Frame Motion Enhancement*

In satellite-based infrared imaging, small aerial targets typically exhibit weak contrast against complex and cluttered backgrounds. These targets may only have a few pixels and thus be indistinguishable in single-frame observations. To highlight potential target regions and suppress static background noise, we introduce an inter-frame motion enhancement mechanism. Given the video sequence $\mathbf{I} = \{I_t\}_{t=1}^{T}$, calculate the absolute residuals between consecutive frames to emphasize motion dynamics, as follows:

$$R_t = |I_t - I_{t-1}|, \quad t = 2, \ldots, T. \tag{1}$$

This residual image $R_t$ captures the change of pixel level over time, which can help to separate small moving objects from a stationary background. However, due to sensor fluctuations or atmospheric interference, the original residual may still have noise. Therefore, we adopt further enhancement techniques, including residual refinement and fusion strategy. In order to reduce the false alarm of residual noise while preserving the target signal, a lightweight convolutional filter $\mathcal{H}_r$ is used for each residual frame, as follows:

$$\tilde{R}_t = \mathcal{H}_r(R_t). \tag{2}$$

The thinning module is composed of two convolution layers, with ReLU activation and batch normalization. It helps to smooth the residual response and emphasizes local motion regions with uniform spatio-temporal gradients. In addition, instead of treating residuals as a separate stream, we refine each refinement residual $\tilde{R}_t$ along the channel dimension connected to the original frame $I_t$, as follows:

$$\bar{I}_t = \text{Concat}(I_t, \tilde{R}_t). \tag{3}$$

Then enter the combination $\bar{I}_t$ sent to the spatial feature extraction trunk for joint feature learning. The residual maps generated by this process help the network focus on dynamic objects and suppress static background noise. Both the original features and residual maps are processed through shared convolutional layers for spatial feature extraction. This fusion enables the network to integrate static and dynamic clues in the

early stage of the pipeline so as to improve the resolution of small targets that are almost invisible in the original infrared spectrum. This inter-frame motion enhancement strategy introduces the minimum computational overhead and provides a strong sensing bias for motion sensitive detection, especially in low SNR environments.

### *3.4. Spatial and Temporal Feature Encoding*

In small target detection from satellite infrared video, the visual cues available in any single frame are often too weak for accurate recognition. Therefore, we design a dual-branch architecture that encodes spatial and temporal features separately to enhance the robustness and continuity of target representation.

In particular, each input frame $I_t$ and its corresponding residual enhanced version $\bar{I}_t$ through a shared convolution backbone $\mathcal{F}_s$ to extract spatial features, as follows:

$$F_t = \mathcal{F}_s(\bar{I}_t). \tag{4}$$

Backbone $\mathcal{F}_s$ uses a ResNet-18 backbone to extract features from the input infrared images. When working with a resolution of $512 \times 512$, the feature maps from different layers, such as C3, C4, and C5, have sizes of $64 \times 64$, $32 \times 32$, and $16 \times 16$, respectively, with increasing depth. If the input resolution is reduced to $256 \times 256$, these feature maps correspondingly become smaller, for example, $32 \times 32$, $16 \times 16$, and $8 \times 8$. Each feature $F_t \in \mathbb{R}^{C \times H \times W}$ captures textures, edges, and local patterns that may indicate the presence of small targets. In order to further enhance the discrimination ability, we also extract the multi-scale representation (see Section 3.5 for details) to ensure fine-grained and context-aware spatial awareness.

Although spatial features provide static appearance clues, temporal consistency is crucial to verify the existence of real moving targets. We use the time encoder $\mathcal{F}_t$ to process the spatial feature sequence spanning the time window $\{F_{t-k}, \ldots, F_t\}$. Some temporal modeling techniques, such as temporal transformers, attention modules, and 3D CNNs, have indeed demonstrated significant success in capturing complex temporal dependencies and enhancing feature representations. In our work, we selected ConvLSTM for the temporal branch because it offers a balance of efficiency, effectiveness, and interpretability, especially suited for infrared small target detection scenarios. ConvLSTM integrates convolutional operations within the recurrent framework, which preserves spatial information while modeling temporal dependencies. This trait is particularly beneficial given the small scale and subtle motion characteristics of the targets we aim to detect, where maintaining spatial resolution and local details is crucial. Moreover, ConvLSTM is computationally less demanding compared with 3D CNNs and transformer-based models, making it more suitable for real-time or resource-constrained applications—an important consideration in satellite-based infrared systems. The ConvLSTM module that implements $\mathcal{F}_t$ captures the motion consistency and time dependence through the gated storage unit, as follows:

$$H_t, C_t = \text{ConvLSTM}(F_t, H_{t-1}, C_{t-1}), \tag{5}$$

where $H_t$ is the hidden state and $C_t$ is the unit state. This formula can realize remote memory tracking while maintaining the spatial structure. This method generates time-enhanced feature $F_t^{\text{temp}}$ and encodes motion mode and inter-frame correlation, which is essential for suppressing false positive (e.g., flickering noise) and confirming the real target trajectory. By jointly modeling spatial appearance and temporal dynamics, our system establishes a robust multidimensional representation of the target, which is ready for the fusion and detection of subsequent modules.

In this process, the attention functionality emerges through the integration of inter-frame motion enhancement and the spatio-temporal feature fusion modules. The inter-frame motion enhancement acts as a form of motion attention by highlighting regions with subtle movement between consecutive frames, effectively directing the network's focus towards dynamic regions that are more likely to contain targets. This residual computation suppresses static background clutter, thereby increasing the signal-to-noise ratio for moving small objects. Additionally, the spatio-temporal fusion modules serve as attention mechanisms by adaptively weighting the importance of local spatial features and global contextual cues across multiple scales. For example, our multi-scale fusion strategy effectively combines coarse semantic information with fine-grained details, enabling the network to dynamically attend to relevant features at different resolutions. This is achieved through learned fusion weights and feature concatenation, which implicitly guide the network to focus more on target-related cues and less on background noise or clutter. Thus, although we do not employ explicit attention modules such as attention gates or self-attention layers, the combination of inter-frame motion enhancement and multi-scale feature fusion functions as an implicit attention mechanism. It guides the network to prioritize salient regions pertinent to small moving infrared targets, especially under challenging conditions like low visibility and background clutter.

### *3.5. Multi-Scale Feature Fusion*

Both extracted spatial features $F_t$ and temporally enhanced features $F_t^{\text{temp}}$ need to be fused to perform accurate small target detection. Due to the tiny size and low contrast of aerial targets in infrared satellite videos, multi-scale fusion is essential to capture both fine local details and broad contextual information.

We design a cross-scale feature fusion module $\mathcal{F}_{\text{fuse}}$ to combine spatial and temporal cues at multiple levels. This involves downsampling higher-resolution features to a common size and concatenating them along the channel dimension, then using a $1 \times 1$ convolution to fuse these features into a single rich feature map. Specifically, a spatial pyramid of feature maps from $F_t$ is employed using a simple multi-scale encoder, which consists of downsampling through strided convolutions, as follows:

$$F_t^{(l)} = \text{Downsample}^{(l)}(F_t), \quad l = 1, \ldots, L. \tag{6}$$

Similarly, a temporal pyramid is generated from $F_t^{\text{temp}}$, as follows:

$$F_t^{\text{temp},(l)} = \text{Downsample}^{(l)}(F_t^{\text{temp}}). \tag{7}$$

At each scale level $l$, the corresponding spatial and temporal features are fused via channel-wise concatenation and processed with a fusion block (Conv + BN + ReLU), as follows:

$$F_t^{\text{fused},(l)} = \text{FusionBlock}\left(\text{Concat}\left(F_t^{(l)}, F_t^{\text{temp},(l)}\right)\right). \tag{8}$$

These fused features are then upsampled to a common resolution, as follows:

$$F_t^{\text{agg}} = \sum_{l=1}^{L} \text{Upsample}(F_t^{\text{fused},(l)}). \tag{9}$$

The aggregation allows the network to combine local, mid-range, and global information, enhancing its ability to detect small targets appearing at any scale. It directly combines multi-scale features at a common high resolution to facilitate the integration of both local and global information efficiently. This approach is grounded in the observation

that high-resolution feature fusion can effectively preserve fine details essential for small target detection without the need for iterative, hierarchical decoding stages. Compared with U-Net architecture, our method achieves performance levels close to those of the hierarchical approach, with the added benefit of reduced computational overhead and inference complexity. The direct upsampling allows the network to leverage multi-scale information simultaneously, thus maintaining spatial accuracy and feature richness, which are critical factors in infrared small target detection where target details are subtle and easily lost. This strategy simplifies the decoding pipeline while ensuring that rich semantic and spatial cues are adequately fused, resulting in detection performance comparable to that of more complex hierarchical decoders.

As for the detection head, it comprises two components: a heatmap prediction branch and a size regression branch. The head consists of two 3 × 3 convolutional layers with 256 filters, followed by separate output layers. The heatmap branch uses a 1 × 1 convolution with a sigmoid activation to produce the probability heatmap $P_t \in [0,1]^{H\times W}$ and generate per-pixel probabilities of target centers. The size regression branch employs a 1 × 1 convolution with linear outputs to predict the target width and height, as follows:

$$P_t = \sigma\Big(\mathcal{F}_{\text{det}}(F_t^{\text{agg}})\Big). \tag{10}$$

The heatmap $P_t$ indicates the likelihood of target presence at each pixel. During training, we supervise the output using binary cross-entropy loss or focal loss, depending on the level of class imbalance between target and background pixels.

A hybrid loss function that balances localization accuracy, class imbalance, and spatial sharpness is designed to train the network effectively. Let $\hat{P}_t \in [0,1]^{H\times W}$ denote the predicted probability map from the detection head at time $t$, and let $P_t^{gt} \in \{0,1\}^{H\times W}$ be the binary ground truth map where target pixels are labeled as 1. The overall loss function is composed of the following three terms:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{bce}} + \lambda_2\mathcal{L}_{\text{dice}} + \lambda_3\mathcal{L}_{\text{tv}}. \tag{11}$$

The binary cross-entropy (BCE) loss penalizes pixel-wise classification errors and ensures correct probability estimation, as follows:

$$\mathcal{L}_{\text{bce}} = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\Big[P_t^{gt}(i,j)\log\hat{P}_t(i,j) + (1-P_t^{gt}(i,j))\log(1-\hat{P}_t(i,j))\Big]. \tag{12}$$

The dice loss helps alleviate the class imbalance problem caused by the extreme sparsity of target pixels in most frames, as follows:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum_{i,j}\hat{P}_t(i,j)\cdot P_t^{gt}(i,j)}{\sum_{i,j}\hat{P}_t(i,j)^2 + \sum_{i,j}P_t^{gt}(i,j)^2 + \epsilon}. \tag{13}$$

The total variation regularization suppresses noise in the output heatmap and encourages spatial consistency, which is especially useful in cluttered backgrounds, as follows:

$$\mathcal{L}_{\text{tv}} = \sum_{i,j}\Big((\hat{P}_t(i+1,j) - \hat{P}_t(i,j))^2 + (\hat{P}_t(i,j+1) - \hat{P}_t(i,j))^2\Big). \tag{14}$$

During the derivation, we perform a simple connected component analysis or non-maximum suppression (NMS) on the binarized heatmap to extract the exact target coordinates. This step eliminates isolated noise peaks and ensures spatial consistency of the detection output.

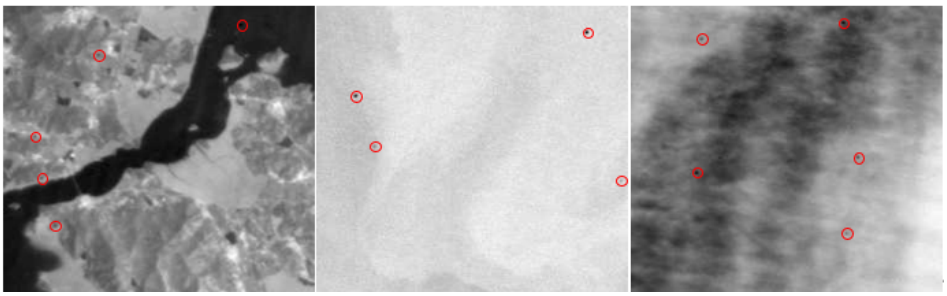
## 4. Experiments

### *4.1. Experimental Setup*

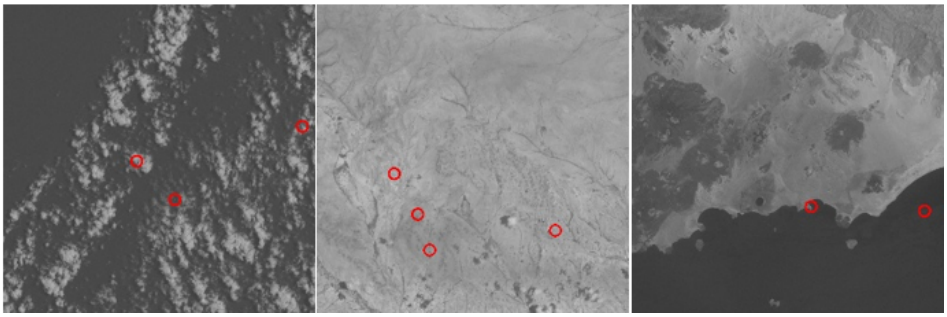
To validate the effectiveness of our proposed method for detecting small infrared targets in satellite video, we conduct experiments on two spaceborne infrared datasets.

The IRAir dataset [38] is a thermal infrared 256 × 256 SDG satellite image dataset for weak target detection, using civil aviation aircraft as the simulation object. The dataset contains 2000 sequence images, and each sequence contains 50 single band simulation images with the same background. This dataset analyzes the pixel characteristics of real air targets in different environments, including daytime imaging, night imaging, sea background, land background, cloud interference, etc. The values of different sequences range from 1 to 10 frames per second. The spatial resolution of the image is 30 m, and the corresponding target speed range is 7.4–8.3 pixel/s.

The IRSatVideo-LEO dataset [39] is a semi-simulated Landsat satellite image dataset with synthesized satellite motion, target appearance, trajectory, and intensity, which includes 200 sequences and 91,366 1024 × 1024 frames with mask annotations. It aims at localizing a scarcity of candidate target pixels from image sequences captured by low earth-orbiting (LEO) satellites of 400–2000 km. To ensure the generalization of the dataset, they randomly sample locations across each continent and ocean on earth, and the cloud cover ratios range from 0 to 61.25%. Illustrations of these two datasets are provided in Figure 3.



(**a**) IRAir dataset



(**b**) IRSatVideo-LEO dataset

**Figure 3.** Illustrations of different Infrared dim-small target datasets.

We randomly divide the dataset into training and test sets with a ratio of 80% and 20%. Data enhancement techniques, including random clipping, horizontal flipping, Gaussian noise injection, and intensity normalization, are applied in the training process to improve generalization. We use sliding window sampling with a fixed length of 5 frames per segment to maintain the time structure.

All experiments were conducted on NVIDIA Tesla V100 GPU (NVIDIA Corporation, Santa Clara, CA, USA). The Adam optimizer is used to train the model with 100 epochs. The initial learning rate is $1 \times 10^{-4}$, and the attenuation is 0.5 times for every 20 epoch.

We use a batch size of 8 and stop early based on validation loss to prevent overfitting. Weight initialization follows the Xavier unified scheme. The hyper-parameters $\lambda_1, \lambda_2, \lambda_3$ are set according to the validation performances, which are $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 0.1$ in the experiments.

### *4.2. Evaluation Metrics*

Several state-of-the-art methods are introduced and serve as comparison algorithms. SpecDETR [21] uses a multi-layer transformer encoder with self-excited subpixel-scale attention modules to directly extract deep spatial–spectral joint features from hyperspectral cubes, which eliminates dependence on pre-trained backbone networks commonly required by vision-based object detectors. RISTDnet [18] constructs a feature extraction framework combining handcrafted feature methods and convolutional neural networks, and establishes a mapping network between feature maps and the likelihood of small targets in the image. MSAFFNet [19] performs infrared small target detection based on an encoder–decoder framework, which also constructs multi-scale labels to focus on the details of the target contour and internal features based on edge information and an internal feature aggregation module. OSCAR [20] (the one-stage cascade refinement network) uses the high-level head as a soft proposal for the low-level refinement head, which is able to process the same target in a cascade coarse-to-fine manner.

All baseline models used for comparison were retrained and evaluated under the same experimental conditions to ensure a fair and objective comparison. Specifically, we reimplemented each baseline model using identical datasets, preprocessing protocols, data augmentation strategies, and training schedules. All results reported in our experiments are obtained from these reimplementations, trained from scratch or fine-tuned as appropriate, rather than directly borrowed from the original literature. Any results that we compare against from previous studies are explicitly reobtained under our standardized setting. Furthermore, to comprehensively evaluate the performance of various methods, we adopt a set of widely used quantitative metrics, focusing on both pixel-level and object-level accuracy. Given the extreme sparsity and small size of targets, it is crucial to use metrics that are sensitive to class imbalance and capable of reflecting true detection capability.

At the object level, we define a detection as correct (true positive) if the predicted target region overlaps with a ground truth region with an Intersection over Union (IoU) greater than a predefined threshold (commonly set to 0.5). These metrics are computed per frame and then averaged over the entire test set.

Precision represents the proportion of correctly detected targets among all predicted targets, as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{15}$$

Recall indicates the proportion of correctly detected targets among all ground truth targets, as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{16}$$

F1-score means the harmonic mean of precision and recall, as follows:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{17}$$

For multi-object detection tasks, we evaluate the precision–recall curve by varying the confidence threshold of predictions. The average precision (AP) is calculated as the area under the PR curve.

The false alarm rate helps quantify how many spurious detections the algorithm produces per frame, which is critical for real-world deployment, as follows:

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}. \tag{18}$$

Localization error (LE) measures the localization error between the center of a detected target and the corresponding ground truth center, as follows:

$$\mathrm{LE} = \frac{1}{N} \sum_{i=1}^{N} \|\hat{c}_i - c_i\|_2, \tag{19}$$

where $\hat{c}_i$ and $c_i$ denote the predicted and ground truth centers, respectively, and $N$ is the number of matched targets. In summary, these metrics provide a robust evaluation framework that captures both detection accuracy and robustness under challenging infrared satellite conditions.

### *4.3. Results*

#### 4.3.1. Comprehensive Comparison

Under the same experimental protocol, the proposed method was comprehensively evaluated against several state-of-the-art infrared small target detection algorithms on two datasets, as shown in Table 1. Our approach consistently achieves superior performance across all metrics. On IRAir, the method obtains the highest precision, recall, and F1-score, representing respective improvements over the best-performing baseline OSCAR. The AP reaches 80.45%, which is approximately a 12% relative gain over OSCAR, indicating that the proposed design maintains high discriminability even across varying confidence thresholds. Meanwhile, FAR is reduced to 14.23, a 10.96% relative reduction, and LE is lowered to 4.12 px, both of which are critical in practical operational contexts where excessive false positives and localization errors can significantly impact downstream decision making. These improvements reflect the method's enhanced capacity for suppressing background interference while preserving sensitivity to small low-contrast moving targets.

**Table 1.** Comparison of detection performance with state-of-the-art methods.

| | Method | Precision↑ | Recall↑ | F1-Score↑ | AP↑ | FAR↓ | LE(px)↓ |
|---|---|---|---|---|---|---|---|
| IRAir | SpecDETR | 72.53 | 67.99 | 78.06 | 68.48 | 17.22 | 6.60 |
| | RISTDnet | 65.78 | 61.89 | 67.78 | 63.56 | 28.56 | 10.90 |
| | MSAFFNet | 69.25 | 66.43 | 75.93 | 67.17 | 22.31 | 8.86 |
| | OSCAR | 75.86 | 77.02 | 78.55 | 71.86 | 15.98 | 5.53 |
| | Ours | **82.12** | **78.34** | **80.23** | **80.45** | **14.23** | **4.12** |
| IRSatVideo | SpecDETR | 78.15 | 71.47 | 75.12 | 72.71 | 18.24 | 15.94 |
| | RISTDnet | 69.25 | 66.43 | 75.93 | 67.17 | 22.31 | 26.86 |
| | MSAFFNetx | 72.79 | 68.66 | 73.14 | 66.76 | 26.75 | 23.30 |
| | OSCAR | 83.14 | 77.56 | 76.28 | 73.60 | 13.98 | 13.36 |
| | Ours | **84.12** | **79.34** | **83.23** | **79.45** | **13.34** | **12.12** |

The performance advantage persists on IRSatVideo, where the detector achieves the highest precision, recall, and AP, and still maintains the lowest FAR and LE among all SOTA methods. In addition, the recall improvement is more significant than that of IRAir, which indicates that the ability to capture time dependence is becoming increasingly important for longer sequences with more complex background dynamics. In this case, static spatial cues are often insufficient, and the fusion of inter-frame motion enhancement characteristics and time coding plays a decisive role in distinguishing the real target from the fluctuating background mode.

Table 2 shows the evaluation results under different input resolutions to further understand the robustness of the method. At 512 × 512 resolution, the model achieved the best effect, with an F1-score of 80.53% and an AP score of 77.25%. When reduced to

256 × 256, the F1-score decreased to 76.06% and the AP decreased to 73.21%, indicating that the multi-scale fusion strategy can effectively retain fine-grained target information even when the spatial details are reduced. The trade-off between accuracy and computational efficiency is obvious in reasoning speed: 21.4 FPS on 512 × 512 and 11.4 FPS on 256 × 256, providing flexibility for deployment in resource-constrained environments. It is worth noting that the competing methods show more obvious accuracy degradation in the case of low resolution, which means that our inter-frame motion enhancement representation is less dependent on the original pixel density. For computational complexity, our model requires approximately 14.1 GFLOPs per inference at an input resolution of 512 × 512. When downscaled to 256 × 256, the FLOPs decrease to around 4.9 GFLOPs due to the quadratic scaling with input size. These indicate that the network is computationally intensive but still within a feasible range for high-performance GPU platforms. In resource-constrained onboard systems, achieving real-time processing would necessitate optimization strategies such as model compression, quantization, or the adoption of lower-precision computations.

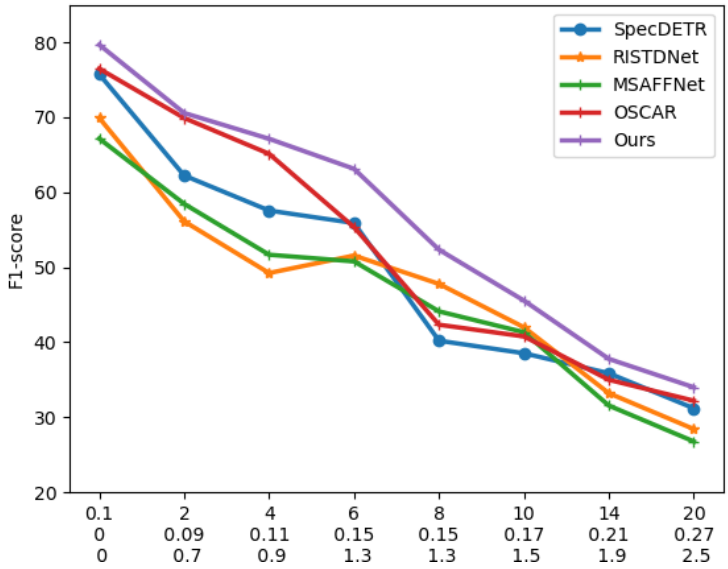
Qualitative analyses support these quantitative trends as well, as illustrated in Figure 4. In low-noise conditions, most methods could identify salient targets, but our approach generates more precise bounding boxes and minimizes background detections. Under severe noise, however, competitors, including OSCAR, tend to misclassify high-intensity clutter and moving background textures as targets. By contrast, the proposed framework maintains stability, which can be attributed to two core design choices. On the one hand, the inter-frame motion enhancement computation suppressed background components consistent over time, leaving motion-specific signal patterns. On the other hand, the multi-scale spatio-temporal fusion retained fine local details while integrating global semantic cues, enabling discrimination between genuine moving objects and noise.

#### 4.3.2. Ablation Study

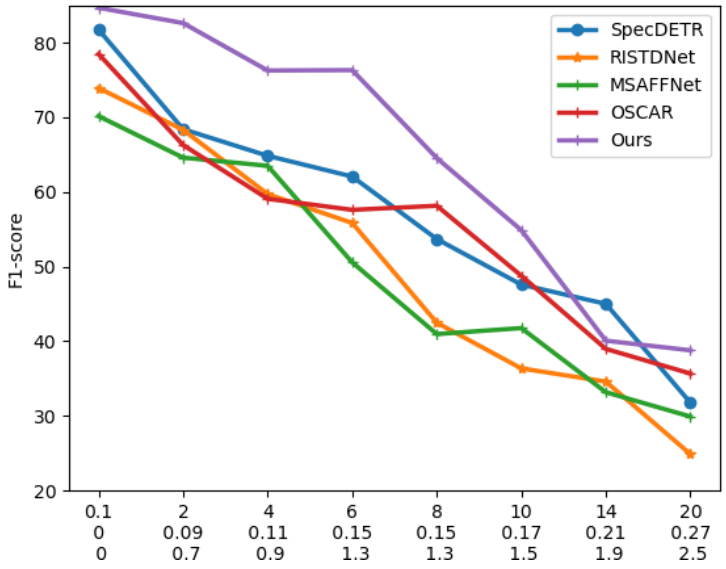
In Table 3, ablation experiments further highlight the contribution of each architectural component. Using only the inter-frame motion enhancement yields an F1-score of 65.12%, while a spatial-branch-only configuration achieves 62.89%, underscoring that motion or appearance cues alone are inadequate for challenging infrared backgrounds. Combining residual and spatial streams increases the F1-score to 74.90%, illustrating their complementarity in emphasizing moving targets while modeling structural details. The inclusion of temporal encoding into a residual-only model led to a substantial recall boost (62.15% to 79.56%), validating that temporal context helps maintain target trajectories and reject transient noise. The full configuration, integrating residual, spatial, and temporal branches, produces the best results on all metrics, especially AP (80.45%) and FAR (14.23%), confirming the necessity of all three elements.

**Table 2.** Effect of resolutions on accuracy and inference time.

| | Method | Precision↑ | Recall↑ | F1-Score↑ | AP↑ | FAR↓ | FPS↑ |
|---|---|---|---|---|---|---|---|
| 512 × 512 | SpecDETR | 73.11 | 67.73 | 73.30 | 66.87 | 20.58 | 58.3 |
| | RISTDnet | 67.32 | 64.26 | 63.91 | 61.76 | 34.06 | **81.0** |
| | MSAFFNet | 70.37 | 64.92 | 65.64 | 64.12 | 30.59 | 72.9 |
| | OSCAR | 76.47 | 71.01 | 77.86 | 72.75 | 16.30 | 43.4 |
| | Ours | **83.02** | **77.58** | **80.53** | **77.25** | **15.57** | 31.4 |
| 256 × 256 | SpecDETR | 67.85 | 72.23 | 65.54 | 64.15 | 28.92 | 34.7 |
| | RISTDnet | 65.80 | 61.94 | 59.25 | 57.54 | 39.49 | **45.2** |
| | MSAFFNet | 66.34 | 62.53 | 64.35 | 61.31 | 28.55 | 41.6 |
| | OSCAR | 72.51 | 74.57 | 72.32 | 72.47 | 31.18 | 27.4 |
| | Ours | **80.17** | **75.60** | **76.06** | **73.21** | **21.36** | 21.4 |




(**a**) IRAir dataset




(**b**) IRSatVideo dataset

**Figure 4.** Results of different networks in scenes under different noise intensity conditions.

**Table 3.** Performance contribution of each module on the IRAir dataset.

| Residual | Spatial | Temporal | Precision↑ | Recall↑ | F1-Score↑ | AP↑ | FAR↓ | LE(px)↓ |
|---|---|---|---|---|---|---|---|---|
| ✓ | | | 68.23 | 62.15 | 65.12 | 63.45 | 28.34 | 12.45 |
| | ✓ | | 65.67 | 60.34 | 62.89 | 60.12 | 30.12 | 13.67 |
| ✓ | ✓ | | 76.34 | 72.67 | 74.90 | 74.23 | 17.78 | 4.23 |
| ✓ | | ✓ | 81.12 | 79.56 | 78.84 | 79.32 | 15.65 | 5.01 |
| ✓ | ✓ | ✓ | 82.12 | 78.34 | 80.23 | 80.45 | 14.23 | 4.12 |

#### 4.3.3. Cross-Dataset Comparison

To further assess the robustness and design rationality of the proposed framework, complementary studies that go beyond the standard single-dataset evaluation are conducted. The first focuses on cross-dataset transferability, an important property for satellite-borne infrared small target detection where operational data often differ markedly from the training set in terms of background texture, sensor noise characteristics, and point-spread function. In this setting, the detector was trained on IRAir and evaluated directly on

IRSatVideo, and vice versa, without any domain-specific adaptation to measure zero-shot generalization. In addition, few-shot adaptation scenarios were considered by fine-tuning the model on a small fraction (1%, 5%, and 10%) of the target-domain training set, allowing us to plot performance–data size curves. Across both zero-shot and few-shot cases, the proposed method achieved higher AP and recall and maintained lower FAR than state-of-the-art baselines such as OSCAR and MSAFFNet, indicating a reduced sensitivity to domain shift and a strong ability to leverage even minimal adaptation data. Feature-space visualizations via t-SNE revealed consistent clustering of target embeddings across datasets for our method, while competing methods exhibited domain-specific separation, further corroborating the robustness of the motion-enhanced temporal representation.

Figure 5 illustrates the detection performance of different combinations of training test datasets, which are measured according to accuracy, recall rate, F1-score, average accuracy (AP), false-positive rate (FAR), and positioning error (LE). This method is always superior to the most advanced methods on IRAir and IRSatVideo datasets. It is worth noting that, compared with the best competitor, it achieves a higher F1-score (+1.68% on IRAir and +4.91% on IRSatVideo), while reducing FAR and LE, indicating that the accuracy and robustness have been improved. When the distribution of training and testing is different, the performance improvement is particularly obvious in cross-domain scenarios, which shows that the method has strong generalization ability.

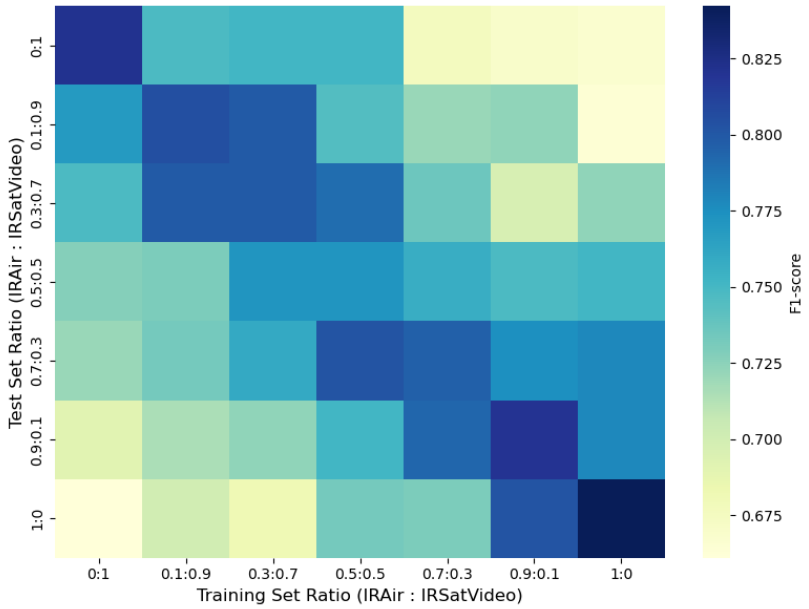
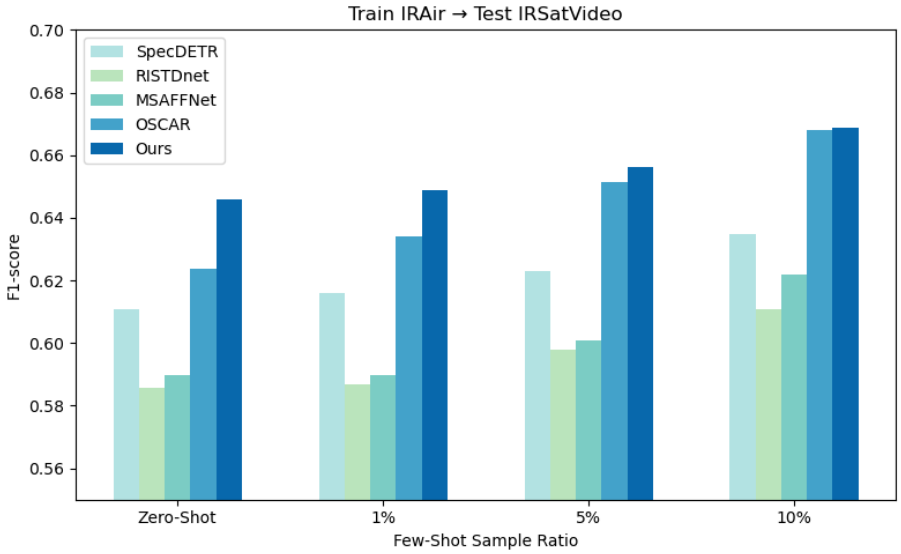
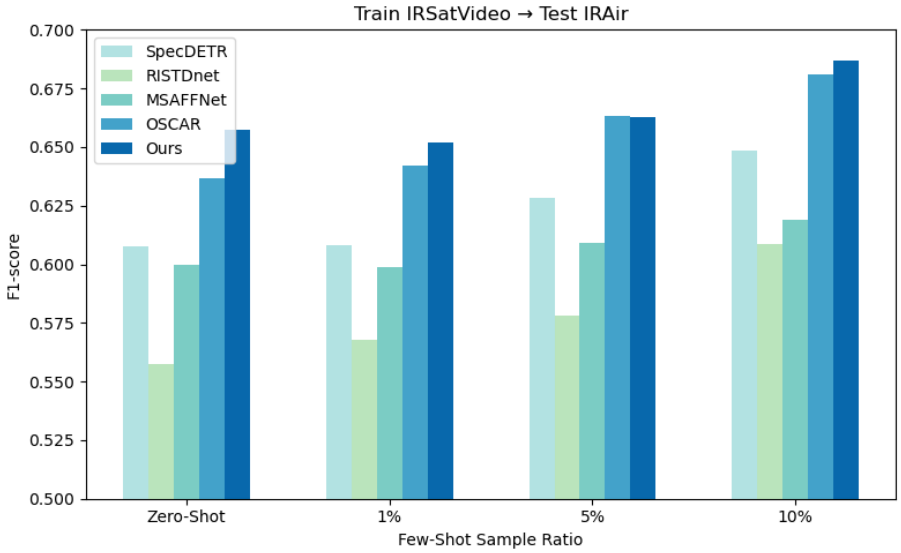



**Figure 5.** Impact of cross-dataset training and testing ratios on F1-score.

Figure 6 presents the results of few-shot adaptation experiments, where a limited portion of target-domain samples is incorporated into training. The bar chart reveals that, even with a small fraction (e.g., 10%) of target-domain samples, the proposed method achieves substantial performance improvements over zero-shot cross-domain detection. The gains are more significant compared with baseline methods, suggesting that the proposed spatio-temporal fusion and inter-frame motion enhancement modules effectively leverage scarce domain-specific information. As the proportion of target-domain samples increases, the performance gap narrows, but our method maintains a consistent advantage across all few-shot settings.




(**a**) IRAir dataset




(**b**) IRSatVideo dataset

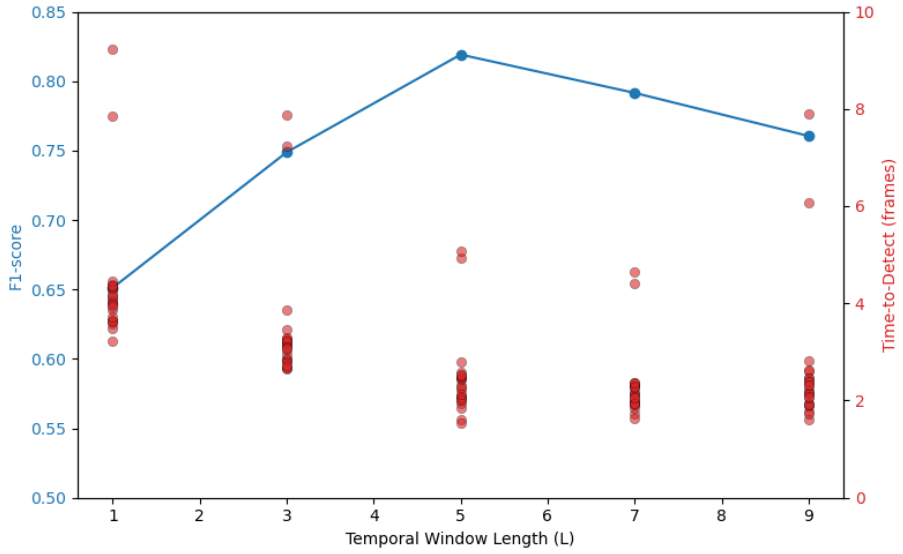
**Figure 6.** Cross-dataset few-shot adaptation performance comparison.

#### 4.3.4. Sensitivity Analysis

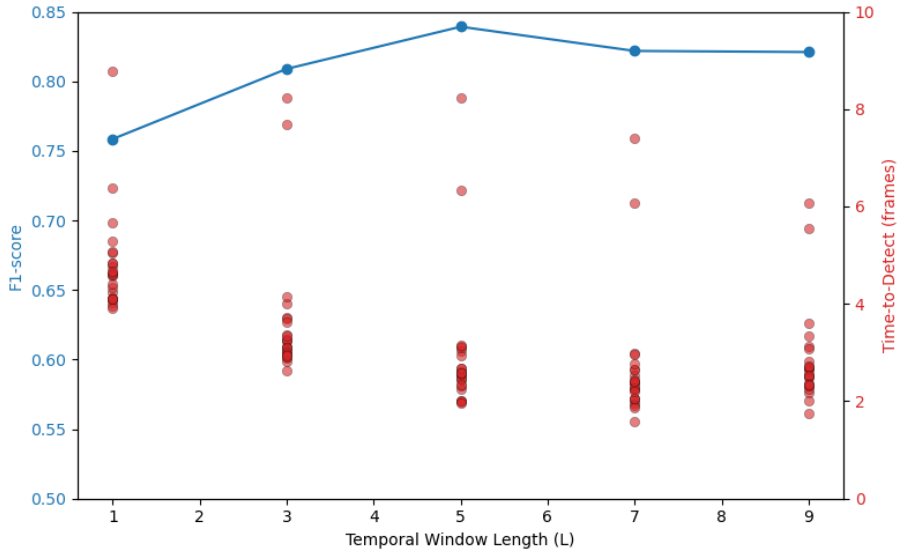
The other study explores the sensitivity of detection performance to the length of the time input window, aiming to determine the best balance between time context utilization and computational efficiency. We change the number of consecutive frames $L$ provided to the network from 1 (single-frame baseline) to 9, keeping all other training protocols unchanged. The results show that the transfer from $L = 1$ to $L = 5$ had a significant improvement in memory and AP, while FAR was reduced and the detection time was shortened, which proved the value of rich time clues. The best compromise is observed between $L = 5$ and $L = 7$, and the performance reaches the peak while maintaining an acceptable reasoning speed. After the window is increased to $L = 7$, there is marginal or no further improvement, and slight degradation is introduced in some sequences, which may be due to the accumulation of uncorrelated time noise, the reduction in resolution per frame in the fused representation, and the significant reduction in throughput per second. These findings not only verify the inclusion of temporal modeling in the architecture, but also

provide practical guidance for resource-constrained deployments. In these deployments, choosing an appropriate temporal depth can retain most of the accuracy advantages without affecting the real-time processing requirements.

Figure 7 analyzes the influence of time window length $L$ on detection accuracy and delay. The left Y-axis (line plot) shows that F1-score increases as $L$ increases from 1 to the optimal range (usually $L = 5$ or $L = 7$), benefiting from a richer time context. However, too long windows will not produce further accuracy improvement, and may even cause slight degradation due to time redundancy. The right Y-axis (scatter cloud) describes the detection time (TTD) distribution of each target. Although a longer window will slightly increase the average TTD, the proposed method maintains a balanced trade-off and achieves a high F1-score with the minimum delay growth. The long tail of TTD distribution represents a challenging low-contrast target, which is still shorter than the competitive method, indicating that the response ability to difficult cases has been improved.




(**a**) IRAir dataset




(**b**) IRSatVideo dataset

**Figure 7.** Effect of temporal window length on detection accuracy and latency.

The experimental results in Table 4 confirm that our model is relatively insensitive to small changes in hyperparameters, with only minor fluctuations in detection performance (F1-score). This indicates strong robustness and adaptability of our framework under different configurations. Such stability is critical for practical deployment, as it suggests that the method does not require precise hyperparameter tuning to achieve high performance.

**Table 4.** Performance with hyperparameter variations on the IRAir dataset.

| Hyperparameters | Values/F1-Score | | |
|---|---|---|---|
| Loss weight of residuals | 1.0 (original) | 0.5 | 2.0 |
| | 80.23 | 80.86 | 80.10 |
| ConvLSTM kernal size | $3 \times 3$ (original) | $5 \times 5$ | $7 \times 7$ |
| | 80.23 | 79.02 | 78.75 |
| Number of ConvLSTM layers | 1 (original) | 2 | 3 |
| | 80.23 | 79.65 | 80.73 |
| Learning rate | $1 \times 10^{-4}$ (original) | $2 \times 10^{-4}$ | $5 \times 10^{-4}$ |
| | 80.23 | 80.59 | 80.30 |

In general, the experimental evidence shows that the integration of inter-frame motion cues, dual-branch spatio-temporal coding, and multi-scale fusion achieves a balanced improvement in detection accuracy, noise robustness, and positioning accuracy. The framework always provides high accuracy and recall, achieves a large amount of AP gain, reduces FAR, and reliably executes in different resolutions. These characteristics not only verify the contribution of the method, but also emphasize the practical value of the system in the satellite-based infrared small target detection, in which the operational constraints require high reliability and adaptability to various sensing and environmental conditions.

## 5. Conclusions

This paper proposes a novel end-to-end framework for detecting small aerial targets in satellite-based infrared video sequences. We introduce an inter-frame motion-enhanced pipeline that highlights inter-frame dynamics, a spatio-temporal feature extraction backbone, and a multi-scale fusion strategy to integrate coarse semantic and fine-grained cues effectively. Extensive experiments on various datasets demonstrate that the method achieves superior performance compared with existing state-of-the-art approaches. The ablation study further verifies the contribution of each module, and shows that residual input, temporal attention coding, and multi-scale fusion play an important role in achieving robust detection performance under low visibility and background clutter.

In the future, we plan to explore the following directions: Research on more effective extraction of backbone network, e.g., attention-based transformer module. Combined with other sensing methods, such as visible spectrum or radar, it can further improve detection robustness under occlusion or adverse conditions. The framework provides a good foundation for various applications of remote monitoring and an early warning system, and lays a foundation for the further development of satellite-based dynamic target perception.

**Author Contributions:** Conceptualization, Q.Y.; Methodology, Y.C. and D.Z.; Software, S.L. and D.Z.; Validation, S.L. and D.Z.; Formal analysis, S.L.; Investigation, Y.C. and D.Z.; Resources, D.M. and D.Z.; Data curation, S.L. and D.Z.; Writing – original draft, Y.C.; Writing – review & editing, Y.C. and S.L.; Visualization, Y.C. and D.Z.; Supervision, Y.C., D.M. and Q.Y.; Project administration, D.M. and Q.Y.; Funding acquisition, D.M. and Q.Y. All authors have read and agreed to the published version of the manuscript.

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

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

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

## References

1. Qin, Y.; Li, B. Effective infrared small target detection utilizing a novel local contrast method. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1890–1894. [CrossRef]
2. Zhuang, J.; Chen, W.; Guo, B.; Yan, Y. Infrared weak target detection in dual images and dual areas. *Remote Sens.* **2024**, *16*, 3608.
3. 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.
4. 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.* **2022**, *32*, 1745–1758. [CrossRef] [PubMed]
5. 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.* **2020**, *59*, 4481–4492.
6. 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.
7. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric contextual modulation for infrared small target detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 950–959.
8. 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. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5007613. [CrossRef]
9. 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]
10. 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.
11. Liu, Q.; Liu, R.; Zheng, B.; Wang, H.; Fu, Y. Infrared small target detection with scale and location sensitivity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16-22 June 2024; pp. 17490–17499.
12. Wang, K.; Du, S.; Liu, C.; Cao, Z. Interior attention-aware network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
13. Zhang, M.; Wang, Y.; Guo, J.; Li, Y.; Gao, X.; Zhang, J. IRSAM: Advancing segment anything model for infrared small target detection. In *Proceedings of the European Conference on Computer Vision*, Milan, Italy, September 29–October 4 2024; Springer: Cham, Switzerland, 2024; pp. 233–249.
14. Kou, R.; Wang, C.; Peng, Z.; Zhao, Z.; Chen, Y.; Han, J.; Huang, F.; Yu, Y.; Fu, Q. Infrared small target segmentation networks: A survey. *Pattern Recognit.* **2023**, *143*, 109788. [CrossRef]
15. Yang, B.; Zhang, X.; Zhang, J.; Luo, J.; Zhou, M.; Pi, Y. EFLNet: Enhancing feature learning network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5906511.
16. Pan, L.; Liu, T.; Cheng, J.; Cheng, B.; Cai, Y. AIMED-Net: An enhancing infrared small target detection net in UAVs with multi-layer feature enhancement for edge computing. *Remote Sens.* **2024**, *16*, 1776.
17. Peng, L.; Lu, Z.; Lei, T.; Jiang, P. Dual-Structure elements morphological filtering and local z-score normalization for infrared small target detection against heavy clouds. *Remote Sens.* **2024**, *16*, 2343. [CrossRef]
18. Hou, Q.; Wang, Z.; Tan, F.; Zhao, Y.; Zheng, H.; Zhang, W. RISTDnet: Robust infrared small target detection network. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 7000805.
19. Tong, X.; Su, S.; Wu, P.; Guo, R.; Wei, J.; Zuo, Z.; Sun, B. MSAFFNet: A multiscale label-supervised attention feature fusion network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5002616.
20. Dai, Y.; Li, X.; Zhou, F.; Qian, Y.; Chen, Y.; Yang, J. One-stage cascade refinement networks for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5000917.
21. Li, Z.; An, W.; Guo, G.; Wang, L.; Wang, Y.; Lin, Z. SpecDETR: A transformer-based hyperspectral point object detection network. *ISPRS J. Photogramm. Remote Sens.* **2025**, *226*, 221–246. [CrossRef]
22. Wei, H.; Yang, Y.; Sun, S.; Feng, M.; Song, X.; Lei, Q.; Hu, H.; Wang, R.; Song, H.; Akhtar, N.; et al. Mono3DVLT: Monocular-Video-Based 3D Visual Language Tracking. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 10-17 June 2025; pp. 13886–13896.
23. Jurosch, F.; Zeller, J.; Wagner, L.; Özsoy, E.; Jell, A.; Kolb, S.; Wilhelm, D. Video-based multi-target multi-camera tracking for postoperative phase recognition. *Int. J. Comput. Assist. Radiol. Surg.* **2025**, *20*, 1159–1166. [PubMed]

24. Habibi, M.; Delaram, Z.; Nourani, M.; Sullivan, D.H. Video-Based Human-Object Interaction Analysis for Patient Behavioral Monitoring. In Proceedings of the 2025 IEEE 13th International Conference on Healthcare Informatics (ICHI), Rende, Italy, 18–21 June 2025; pp. 414–422.
25. Wan, M.; Gu, G.; Qian, W.; Ren, K.; Maldague, X.; Chen, Q. Unmanned aerial vehicle video-based target tracking algorithm using sparse representation. *IEEE Internet Things J.* **2019**, *6*, 9689–9706. [CrossRef]
26. Howard, R.T.; Book, M.L.; Bryan, T.C. Video-based sensor for tracking three-dimensional targets. In *Proceedings of the Atmospheric Propagation, Adaptive Systems, and Laser Radar Technology for Remote Sensing, Barcelona, Spain, 31 January 2001*; SPIE: Bellingham, WA, USA, 2001; Volume 4167, pp. 242–251.
27. Zhang, Z.; Wang, C.; Song, J.; Xu, Y. Object tracking based on satellite videos: A literature review. *Remote Sens.* **2022**, *14*, 3674.
28. Wan, X.; Zhou, S.; Wang, J.; Meng, R. Multiple object tracking by trajectory map regression with temporal priors embedding. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 1377–1386.
29. Chen, S.; Wang, T.; Wang, H.; Wang, Y.; Hong, J.; Dong, T.; Li, Z. Vehicle tracking on satellite video based on historical model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 7784–7796. [CrossRef]
30. Othmani, M. A vehicle detection and tracking method for traffic video based on faster R-CNN. *Multimed. Tools Appl.* **2022**, *81*, 28347–28365. [CrossRef]
31. Ibrahim, N.; Darlis, A.R.; Herianto; Kusumoputro, B. Performance Analysis of YOLO-DeepSORT on Thermal Video-Based Online Multi-Object Tracking. In Proceedings of the 2023 3rd International Conference on Robotics, Automation and Artificial Intelligence (RAAI), Singapore, 14–16 December 2023; pp. 46–51.
32. Zhao, S.; Xu, T.; Li, H.; Wu, X.J.; Kittler, J. Visual complexity guided diffusion defender for video object tracking and recognition. *Pattern Recognit.* **2026**, *169*, 111867. [CrossRef]
33. Liu, Z.; Liang, S.; Guan, B.; Tan, D.; Shang, Y.; Yu, Q. Collimator-assisted high-precision calibration method for event cameras. *Opt. Lett.* **2025**, *50*, 4254–4257. [CrossRef] [PubMed]
34. Huang, H.; Chen, C.; Guan, B.; Tan, Z.; Shang, Y.; Li, Z.; Yu, Q. Ridge estimation-based vision and laser ranging fusion localization method for UAVs. *Appl. Opt.* **2025**, *64*, 1352–1361. [CrossRef]
35. Xu, C.; Qi, H.; Zheng, Y.; Peng, S. Real-time moving vehicle detection in satellite video based on historical differential information and grouping features. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5614716. [CrossRef]
36. Tang, D.; Tang, S.; Wang, Y.; Guan, S.; Jin, Y. A global object-oriented dynamic network for low-altitude remote sensing object detection. *Sci. Rep.* **2025**, *15*, 19071. [CrossRef]
37. Shi, Y.; Xu, G.; Liu, Y.; Chen, H.; Zhou, S.; Yang, J.; Dong, C.; Lin, Z.; Wu, J. Deep Learning-Based Real-Time Surf Detection Model During Typhoon Events. *Remote Sens.* **2025**, *17*, 1039. [CrossRef]
38. Li, Z.X.; Xu, Q.Y.; An, W.; He, X.; Guo, G.W.; Li, M.; Ling, Q.; Wang, L.G.; Xiao, C.; Lin, Z.P. A lightweight dark object detection network for infrared images. *J. Infrared Millim. Waves* **2025**, *44*, 285. [CrossRef]
39. 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.