remote sensing

MDPI

Article

FeedbackSTS-Det: Sparse-Frames-Based Spatio-Temporal Semantic Feedback Network for Moving Infrared Small Target Detection

Yian Huang [1], Qing Qin [2], Aji Mao [1], Xiangyu Qiu [1], Han Guo [3], Liang Xu [1], Xian Zhang [1] and Zhenming Peng [1,*]

[1] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; huangyian1@std.uestc.edu.cn (Y.H.); ajimao@std.uestc.edu.cn (A.M.); xiangyuqiu@std.uestc.edu.cn (X.Q.); liangx@std.uestc.edu.cn (L.X.); zhangxian@std.uestc.edu.cn (X.Z.)
[2] School of Television and Audiovisual Arts (Documentary School), Communication University of Zhejiang, Hangzhou 310018, China; 20190012@cuz.edu.cn
[3] School of Computer Science and Engineering (School of Cyber Security), University of Electronic Science and Technology of China, Chengdu 611731, China; 2024080908010@std.uestc.edu.cn
* Correspondence: zmpeng@uestc.edu.cn; Tel.: +86-28-83208946

Highlights

What are the main findings?

- A closed-loop semantic feedback strategy with forward and backward spatio-temporal refinement modules is proposed to enhance inter-frame information exchange, improving detection accuracy and reducing false alarms.
- An embedded sparse semantic module (SSM) is introduced to capture long-range temporal dependencies efficiently by grouping frames at intervals, achieving low computational overhead.

What are the implications of the main findings?

- The proposed FeedbackSTS-Det network demonstrates strong generalization ability and scene adaptability across multiple multi-frame infrared small target datasets.
- The sparse temporal modeling approach offers a lightweight solution for leveraging temporal cues in resource-constrained scenarios, facilitating practical deployment in real-time systems.

Abstract

Infrared small target detection (ISTD) has been a critical technology in various civilian and industrial applications over the past several decades, such as civilian patrol missions aboard UAVs or shipboard systems, and industrial inspection tasks like factory defect scanning. Nevertheless, moving infrared small target detection still faces considerable challenges: existing models suffer from insufficient spatio-temporal semantic correlation and are not lightweight-friendly, while algorithms that perform reliably across diverse scenarios are in great demand for real-world applications. To address these issues, we propose FeedbackSTS-Det, a sparse-frames-based spatio-temporal semantic feedback network. A closed-loop spatio-temporal semantic feedback strategy with paired forward and backward refinement modules that work cooperatively across the encoder and decoder is adopted to enhance information exchange between consecutive frames, effectively improving detection accuracy and reducing false alarms. Moreover, we introduce an embedded sparse semantic module (SSM), which operates by strategically grouping frames by interval, propagating semantics within each group, and reassembling the sequence to efficiently capture long-range temporal dependencies with low computational overhead.

Check for updates

Academic Editor: Tania Stathaki

Received: 21 April 2026
Revised: 5 June 2026
Accepted: 14 June 2026
Published: 18 June 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.

Extensive experiments on many widely adopted multi-frame infrared small target datasets demonstrate the consistent effectiveness of our proposed network across diverse scenes.

Keywords: infrared small target detection; sparse spatial-temporal semantics; feature alignment feedback; dynamic scenes

1. Introduction

Infrared small target detection (ISTD) has served as a fundamental and critical technology over the past several decades, playing an essential role in various applications such as civilian patrol missions aboard unmanned aerial vehicles (UAVs) [1,2] or shipboard systems [3], as well as industrial inspection tasks like factory defect scanning [4]. In many of these real-world scenarios, detection systems are deployed on resource-constrained platforms where size, weight, and power limitations impose strict constraints on computational complexity and memory footprint. These platforms require lightweight yet effective algorithms to enable real-time processing under limited onboard resources.

Although traditional single-frame infrared small target (SIRST) detection methods have achieved notable progress, they suffer from an inherent limitation: the inability to leverage temporal motion cues, resulting in insufficient robustness under low signal-to-noise ratio (SNR) scenarios. To address this bottleneck, multi-frame infrared small target detection (MIRSTD) has emerged, which significantly enhances detection performance by integrating temporal contextual information.

Existing MIRST detection approaches can be broadly categorized into two technical paradigms. The first is model-driven methods, which rely on mathematical priors for target–background separation. Early filtering-based approaches [5,6] and subsequent optimization-based methods leveraging low-rank and sparse decomposition [7–9] have demonstrated effectiveness in specific scenarios. More recent tensor-based extensions incorporate temporal information through spatio-temporal modeling [10–12]. Although theoretically principled, these methods often require meticulous manual tuning and struggle to adapt to complex, nonlinear clutter environments.

The second category comprises data-driven methods, which employ deep learning to automatically learn spatio-temporal features. Within single-frame detection, substantial efforts have focused on enhancing spatial feature discrimination through multi-scale fusion [1,13], dense connections [14,15], and feature grouping mechanisms [16]. Another important direction incorporates mathematical priors into network architectures to improve interpretability and robustness [17–20], while deep unfolding networks [21–23] further bridge model-driven and data-driven paradigms. For multi-frame detection, temporal information is leveraged through various strategies, including unsupervised sequence registration [24], motion modeling [25], deformable alignment [26], nonlinear spatio-temporal feature enhancement [27], and temporal anomaly reformulation [28]. Meanwhile, introducing powerful model frameworks from other domains, such as Mamba [29–31], has also become a common direction in the current field. Despite their advancements, data-driven methods face the following challenges:

Infrared small targets typically occupy only a few pixels, lack distinct shape or texture features, and are easily hidden in complex backgrounds. Without sufficient use of temporal information, effective target detection becomes very difficult. Nevertheless, excessively long sequences greatly increase computational cost and make lightweight design difficult to achieve. In addition, abundant false alarm sources and background clutter often make

algorithm performance unstable. Therefore, infrared small target detection algorithms require extremely high scene adaptability.

To tackle these challenges, this paper proposes a novel sparse-frames-based spatio-temporal feedback semantics network, termed FeedbackSTS-Det. The core of our approach is a spatio-temporal semantic feedback strategy that establishes a closed-loop semantic association mechanism. Specifically, it integrates a forward spatio-temporal semantic refinement module (FSTSRM) in the encoder and a backward spatio-temporal semantic refinement module (BSTSRM) in the decoder in a cooperative manner, both built upon a balanced 3D Res-UNet backbone. Additionally, we embed a sparse semantic module (SSM) within both refinement modules, which performs structured sparse temporal modeling to capture long-range dependencies with low computational cost. These designs facilitate robust implicit inter-frame registration and continuous spatio-temporal semantic refinement, ensuring effective false alarm suppression and consistent performance. Furthermore, extensive experiments on widely adopted multi-frame infrared small target datasets confirm the effectiveness of our proposed method under the evaluated scene configurations.

The main contributions of this work are summarized as follows:

- We propose a closed-loop semantic feedback strategy for infrared small target detection, comprising paired forward and backward refinement modules that operate across the encoder and decoder. This design enhances information exchange between consecutive frames, effectively improving detection accuracy and reducing false alarms.
- We introduce an embedded sparse semantic module (SSM), which operates by strategically grouping frames by interval, propagating semantics within each group, and reassembling the sequence to efficiently capture long-range temporal dependencies with low computational overhead.
- Our model exhibits strong detection capability and false alarm suppression across all experiments on many widely adopted multi-frame infrared small target datasets, demonstrating its robustness on the evaluated benchmarks.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review and critical analysis of related work. Section 3 describes the proposed network architecture. Section 4 presents experimental results and analysis. Finally, the discussions and conclusions are described in Sections 5 and 6.

2. Related Work

In this section, we provide a concise review of the main methods in the field of IRST, with a particular emphasis on research related to MIRST detection. Existing MIRST detection methods can be broadly categorized into model-driven methods and data-driven methods. The former primarily relies on filtering or handcrafted optimization models for feature extraction, while the latter utilizes deep learning methods to automatically learn feature representations from large-scale data.

2.1. Model-Driven Schemes for IRST

Model-driven infrared small target detection methods can be divided into two categories: filtering methods based on local priors and optimization methods based on non-local priors.

Filtering methods operate under the background consistency assumption, estimating the image background and enhancing targets via background subtraction. Classical methods like Top-Hat [5] and two-dimensional least mean square (TDLMS) [6] perform well in uniform backgrounds but struggle in complex cluttered environments. Subsequent

improvements include methods leveraging the human visual system (HVS) to exploit target saliency characteristics [32–34].

In contrast, optimization methods exploit the non-local correlation of backgrounds and target sparsity, achieving target–background separation through model formulation. Assuming low-rank background subspaces, Gao et al. [7] model ISTD as a low-rank and sparse decomposition problem. This inspired numerous subspace-based variants [35–37]. Further developments include designing the reweighted infrared patch tensor [8], formulating the non-convex rank approximation minimization [9], proposing the partial sum of tensor nuclear norm [38], and incorporating local contrast energy or hyper total variation to improve robustness [39,40].

To incorporate temporal information, sequence-based tensor methods have been developed, extending prior spatial models into the spatio-temporal domain. Representative approaches include extending multi-subspace learning to the tensor domain [10], constructing spatio-temporal tensors with overlapping patches [41], and extracting non-overlapping patches across frames via sliding windows [11]. Further improvements involve designing specialized regularization methods combined with non-convex tensor rank approximation [42], imposing constraints on Tucker decomposition factor matrices [12], and incorporating total variation on sparse representation matrices [43]. Advanced methods also leverage sparse regularization to enhance target–background distinction [44] or extend tensor decomposition to four-dimensional representations [45].

In summary, model-driven methods rely heavily on handcrafted priors that often fail to generalize across diverse scenes and require meticulous parameter tuning for each new scenario. This limitation has motivated the shift toward data-driven paradigms that learn spatio-temporal features directly from large-scale data.

2.2. Data-Driven Schemes for IRST

Model-driven methods often rely on manual parameter tuning and idealized mathematical models, whereas data-driven methods employ deep learning to automatically learn features from large-scale data, enabling end-to-end infrared small target detection.

Within this data-driven paradigm, single-frame detection focuses on extracting more discriminative spatial features from a single image. Specifically, Dai et al. [1] and Zhang et al. [13] enhance feature representation for small targets by designing multi-scale asymmetric fusion modules that synergize global context attention with local attention mechanisms targeting point-like objects. Similarly, Li et al. [14] and Wu et al. [15] strengthen internal connections among network modules and the flow of feature information through repeated stacking of model structures. Hou et al. [16] introduce a feature map grouping mechanism and assign higher weights to groups containing small targets to focus on critical information.

Beyond these purely data-driven designs, another important direction is to incorporate mathematical model priors into neural network architectures. For instance, Sun et al. [18] and Zhang et al. [46] embed Gaussian operator modules into networks based on the assumption that the grayscale distribution of infrared targets approximates a Gaussian distribution. Wu et al. [19] design learnable local saliency kernels to embed traditional filtering principles into networks, while Hu et al. [20] develop a specific convolution block that simulates central difference to extract image gradient features more robustly. In addition, many researchers have introduced deep unfolding networks (DUNs) based on robust principal component analysis [47] into ISTD [21,22,48], bridging model-driven and data-driven methodologies. These approaches achieve strong detection performance while maintaining a high degree of interpretability.

In contrast to single-frame approaches, MIRSTD methods place greater emphasis on mining and utilizing temporal information across consecutive frames. Hou et al. [24] propose a hierarchical temporal consistency loss function within an unsupervised end-to-end framework to resolve registration inconsistencies of infrared image sequences. Li et al. [25] extract spatial saliency features via a detection head, and then input the features of five consecutive frames into a self-designed direction-coded temporal U-shape module (DTUM) for motion modeling. The recurrent feature refinement framework (RFR) [26] employs deformable convolutions and a pyramid structure to enhance inter-frame registration and positional information extraction for long-sequence detection. Wu et al. [27] further employ nonlinear networks to enhance correlations between spatial-temporal features and performs detection in an unsupervised manner. Furthermore, Li et al. [28] reformulate infrared small target detection as a one-dimensional signal anomaly detection problem, where global temporal saliency and correlation characteristics distinguish targets from interference signals.

In addition to conventional deep learning paradigms, incorporating powerful frameworks from other domains has also attracted increasing attention. A state-space model with linear complexity, named Mamba [29], has emerged, capable of achieving performance comparable to or even better than the Transformer model [49] at a lower cost in long-sequence modeling tasks. It is thus natural to apply this high-efficiency capability to moving infrared small target detection, which is crucial for deployment. Some researchers have designed multi-stage Mamba architectures [30,31] to progressively construct complementary local and global features. Other researchers have integrated classic feature extraction methods commonly used in infrared target detection into the Mamba framework, for example, by developing a spatiotemporal attention mechanism via Fourier transform [50], or by capturing the intensity and directional information of pixel differences between a scanning position and its central neighborhood [51]. Further studies have, within the Mamba framework, consolidated dual-branch structures for global and local spatiotemporal feature extraction while reducing computational redundancy [52,53].

Overall, existing deep-learning-based methods for moving infrared small target detection all attempt to leverage learning from large-scale datasets to fully exploit the spatiotemporal feature information of infrared small targets. However, they generally rely too heavily on dense inter-frame feature extraction and predominantly focus on utilizing temporally preceding feature information to compensate for the detection of subsequent frames, while seldom using subsequent detection results to retroactively refine earlier ones. The method we propose in Section 3 will address the aforementioned issues.

3. Materials and Methods

3.1. Overall Framework

As illustrated in Figure 1, we present a sparse-frames-based spatio-temporal feedback semantic network (FeedbackSTS-Det), which addresses temporal modeling from a sparse sampling perspective. Built upon a balanced 3D Res-UNet backbone [54,55], our network processes sliding windows by partitioning the continuous temporal image data into shorter sequences (denoted as *Seq*) of length D, with each *Seq* sequentially extracted as input. This approach maintains consistency between the training and inference processes, ensuring the stability of model performance during actual deployment.

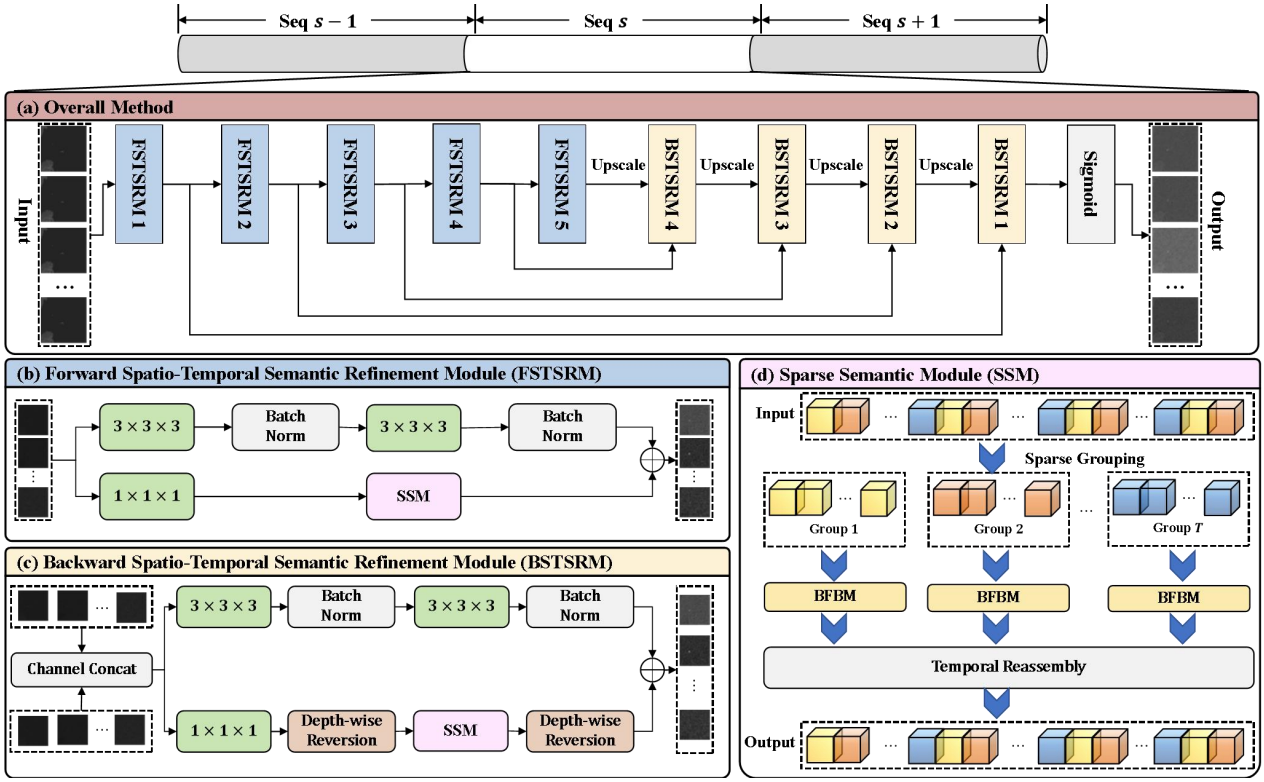


Figure 1. Overall procedure of FeedbackSTS-Det model for ISTD: (**a**) presents the overall framework; (**b**) denotes the forward spatio-temporal semantic refinement module (FSTSRM); (**c**) illustrates the backward spatio-temporal semantic refinement module (BSTSRM); and (**d**) describes the sparse semantic module (SSM).

The encoder phase incorporates five consecutive forward spatio-temporal semantic refinement modules (FSTSRM), while the decoder phase employs four successive backward spatio-temporal semantic refinement modules (BSTSRM). Both modules critically integrate our novel sparse semantic module (SSM), which operates by strategically grouping frames at intervals, propagating semantics within each group using the basic feedback module (BFM), and then reassembling the sequence. Collectively, FSTSRM and BSTSRM establish a spatio-temporal semantic feedback strategy.

The spatio-temporal semantic feedback strategy will be detailed in Section 3.2, whereas the SSM will be elaborated in Section 3.3. To alleviate the excessive growth of floating-point operations during iterative feature propagation, we deliberately set the base channel number of the overall framework to 8, ensuring that the deepest layer maintains a maximum channel count of 128. To achieve satisfactory segmentation performance, we adopt the Soft-IoU loss function [14,25] for training.

3.2. Spatio-Temporal Semantic Feedback Strategy

The spatio-temporal semantic feedback strategy establishes a closed-loop semantic association mechanism, which comprises the FSTSRM and BSTSRM described in detail below.

3.2.1. Forward Spatio-Temporal Semantic Refinement Module

The forward spatio-temporal semantic refinement module (FSTSRM) establishes the forward propagation path in our spatio-temporal feedback system, progressively building spatio-temporal semantic understanding across frames. Five consecutive FSTSRMs constitute the encoding layer of the overall framework.

Let $\mathbf{X}^{(l)} \in \mathbb{R}^{C\times D\times H\times W}$ denote the input feature map of the encoding layer at layer l. As illustrated in Figure 1b, forward refinement can be expressed as a bidirectional information integration process:

$$\mathbf{Y}^{(l)}_{\text{FSTSRM}} = \mathcal{B}_{\text{context}}(\mathbf{X}^{(l)}) \oplus \mathcal{P}_{\text{forward}}(\mathbf{X}^{(l)}) \tag{1}$$

where $\oplus$ denotes element-wise addition. Let $\text{Conv}_{3\times3\times3}(\cdot)$ and $\text{Conv}_{1\times1\times1}(\cdot)$ denote 3D convolutions with kernel sizes of 3 and 1, respectively. Specifically, $\mathcal{B}_{\text{context}}$ consists of two consecutive 3D convolutions with batch normalization to preserve local spatio-temporal structures, and its formulation is given in Equation (2). Meanwhile, the forward temporal propagation path $\mathcal{P}_{\text{forward}}$ employs a sparse semantic modeling strategy, which is defined in Equation (3).

$$\mathcal{B}_{\text{context}}(\mathbf{X}) = \text{BN}(\text{Conv}_{3\times3\times3}(\text{BN}(\text{Conv}_{3\times3\times3}(\mathbf{X})))) \tag{2}$$

$$\mathcal{P}_{\text{forward}}(\mathbf{X}) = \text{SSM}(\text{Conv}_{1\times1\times1}(\mathbf{X})) \tag{3}$$

where the operator $\text{SSM}(\cdot)$ represents the processing through the SSM, whose details are provided in Figure 1d and Section 3.3.

The design of FSTSRM enables complementary information integration. The context branch provides high-fidelity spatial cues, while the sparse propagation branch establishes long-range temporal associations.

3.2.2. Backward Spatio-Temporal Semantic Refinement Module

The backward spatio-temporal semantic refinement module (BSTSRM) implements retrospective refinement of FSTSRM by reversing temporal dependencies. Four consecutive BSTSRMs modules constitute the decoding layer of the overall framework.

Let $\mathbf{X}^{(l)}_{\text{dec}} \in \mathbb{R}^{C\times D\times H\times W}$ denote the input feature map of the l-th decoding layer, where $l \in [1, 5]$ and larger l indicates deeper layers. As illustrated in Figure 1a,c, the input of the BSTSRM at layer 5 is formed by concatenating the outputs of the two deepest FSTSRM along the channel dimension (Equation (4)). For $l \in [1, 4]$, the input concatenates the output of the BSTSRM from the deeper layer with the output of the FSTSRM at the same level along the channel dimension (Equation (5)).

$$\mathbf{X}^{(5)}_{\text{dec}} = \text{Concat}\left(\mathbf{Y}^{(4)}_{\text{FSTSRM}}, \text{Up}\left(\mathbf{Y}^{(5)}_{\text{FSTSRM}}\right)\right) \tag{4}$$

$$\mathbf{X}^{(l)}_{\text{dec}} = \text{Concat}\left(\mathbf{Y}^{(l)}_{\text{FSTSRM}}, \text{Up}\left(\mathbf{Y}^{(l+1)}_{\text{BSTSRM}}\right)\right) \tag{5}$$

where $\text{Up}(\cdot)$ denotes upsampling. Building upon this input, the backward refinement process is then expressed as:

$$\mathbf{Y}^{(l)}_{\text{BSTSRM}} = \mathcal{B}_{\text{context}}(\mathbf{X}^{(l)}_{\text{dec}}) \oplus \mathcal{P}_{\text{backward}}(\mathbf{X}_{\text{dec}}{}^{(l)}) \tag{6}$$

where the backward feature refinement path $\mathcal{P}_{\text{backward}}$ is further defined as:

$$\mathcal{P}_{\text{backwad}}(\mathbf{X}) = \mathcal{R}(\text{SSM}(\mathcal{R}(\text{Conv}_{1\times1\times1}(\mathbf{X})))) \tag{7}$$

where the depth-wise reversal operator $\mathcal{R}$ is the core operation of BSTSRM. Given an input tensor $\mathbf{X} \in \mathbb{R}^{C\times D\times H\times W}$, $\mathcal{R}$ is defined as:

$$\mathcal{R}(\mathbf{X}_{c,d,h,w}) = \mathbf{X}_{c,D-d+1,h,w}, \quad \forall d \in [1, D] \tag{8}$$

3.2.3. Feedback Strategy

As illustrated in Figure 1a, five consecutive FSTSRMs progressively extract spatio-temporal semantic features from low to high levels. By simultaneously preserving spatial context information and performing forward sparse propagation, they enable forward information flow, thereby constituting the feedforward pathway. Meanwhile, four consecutive BSTSRMs reverse the temporal dimension, allowing deep-layer spatio-temporal semantic information to propagate backward to shallower layers, thus forming the feedback pathway. Throughout the propagation process, continuous information exchange between the two pathways facilitates the overall spatio-temporal semantic feedback strategy. This feedback structure not only inherits the advantages of traditional networks by leveraging preceding frames to influence the detection results of subsequent frames, but also allows subsequent frames to correct errors in preceding ones, thereby enhancing system robustness, reducing false alarm rates, and improving detection accuracy.

Conventional sliding-window methods typically focus on using preceding frames to influence subsequent ones and rely on dense feature correction within each window, which lacks flexibility. In contrast, our spatio-temporal semantic feedback strategy promotes sufficient information exchange between temporally earlier and later features without incurring redundant computational growth.

3.3. Sparse Semantic Module

The sparse semantic module (SSM) is incorporated into the temporal feature extraction paths of both the FSTSRM and BSTSRM modules, with its specific structure illustrated in Figure 1d. It is designed to achieve efficient long-range temporal modeling by strategically processing a subset of frames, thereby reducing computational overhead while enhancing feature propagation across distant frames. The SSM operates in three consecutive stages: sparse grouping, intra-group spatio-temporal semantic propagation, and temporal reassembly.

3.3.1. Sparse Grouping

Given a sampling interval $T \in \mathbb{N}^+$ and an input feature sequence $I_1, I_2, \ldots, I_D$ of length D, the sparse grouping partitions the index set $\mathcal{D} = 1, 2, \ldots, D$ into T disjoint subsequences:

$$\mathcal{G}_k = \{I_i \mid i \in \mathcal{D}, i \equiv k \pmod{T}\}, \quad k = 1, 2, \ldots, T \tag{9}$$

Equivalently, the k-th group contains frames with indices $\{k, k+T, k+2T, \ldots, k+M_kT\}$, where $M_k = \lfloor (D-k)/T \rfloor$. This constructed index table is also of great help for reorganizing the grouped input feature groups.

In our model framework, there are five encoder layers and four decoder layers. During the feature propagation between frames at the same network layer, the scale of the feature maps remains unchanged, and the number of feature propagation operations is also $D - T$. Therefore, the computational cost is consistent across these operations. Let M_i denote the computational cost of feature propagation between frames at the i-th layer. Then, the overall computational cost of the network is given as follows:

$$\mathcal{T}(T) = \left(\sum_{i=1}^{4} 2M_i + M_5 \right) (D - T) \tag{10}$$

When $T = 1$, the model reduces to a sequential model without the sparse mechanism. The reduction in computational cost achieved by the sparse mechanism is given by Equation (11). It can be observed that as the sparse interval T increases, the degree of computational cost reduction becomes larger.

$$\mathcal{T}_{\text{red}}(T) = \mathcal{T}(1) - \mathcal{T}(T) = \left(\sum_{i=1}^{4} 2M_i + M_5\right)(T-1) \tag{11}$$

3.3.2. Intra-Group Spatio-Temporal Semantic Propagation

Although the sparse mechanism is an effective approach for reducing computational cost, simply decreasing the number of propagation steps inevitably undermines the effectiveness of feature propagation. Therefore, it is crucial to strengthen the propagation ability within each group. To address this, we propose the basic feedback module (BFBM), as illustrated in Figure 2a. For each group, all frames except the first one undergo two procedures, which are feature extraction (illustrated in Figure 2b) and feature alignment (illustrated in Figure 2c). Both procedures are adapted from the pyramid deformable alignment (PDA) module [26] and are deliberately decoupled in our design to reduce redundant computation.

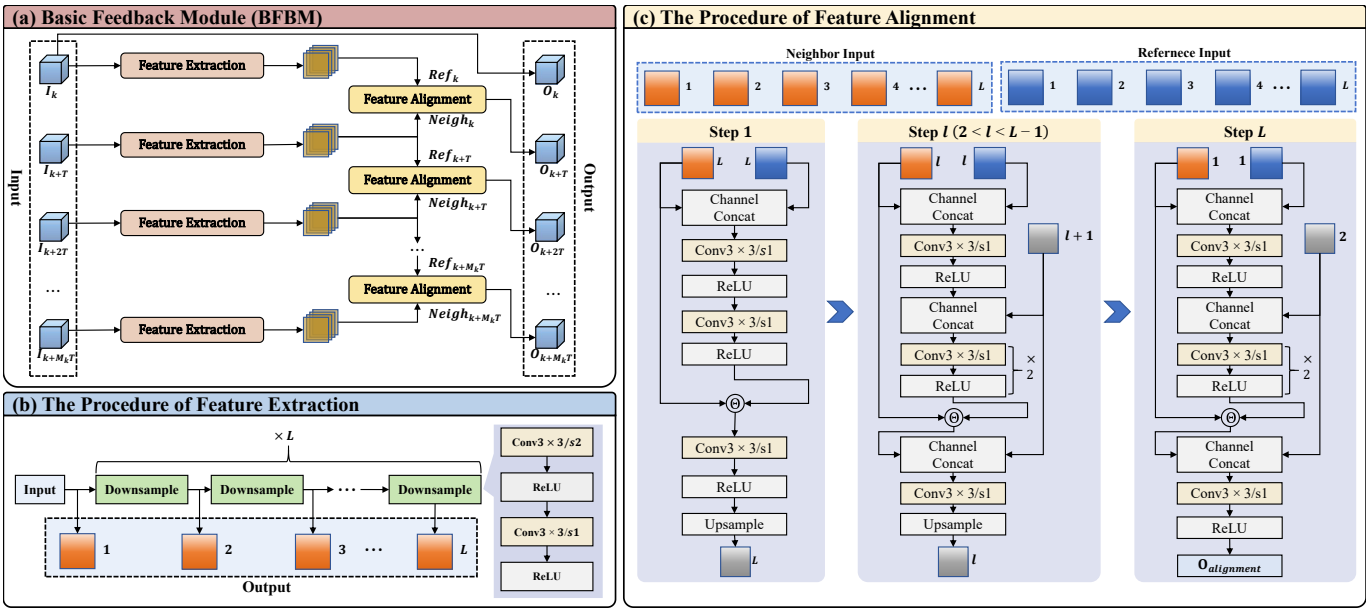


Figure 2. Basic feedback module (BFBM): (**a**) presents the framework of the BFBM module, which is embedded within the SSM; (**b**) details the procedure of feature extraction in BFBM; and (**c**) shows the procedure of the feature alignment module in BFBM.

Specifically, for a given feature group $\mathcal{G}_k$ obtained according to Equation (9), the initial frame I_k serves only as a reference. Consequently, the first output O_k within the group is obtained by directly passing I_k through without any special processing, as formulated in Equation (12). For the remaining outputs O_{k+mt}, where $m \in [1, M] \cap \mathbb{N}$, they are generated by first extracting features from the current input frame, which serves as the neighbor input, and its adjacent counterpart, which serves as the reference input. This is followed by the procedure of feature alignment. The corresponding computational process is given by Equation (13).

$$O_k = I_k \tag{12}$$

$$O_{k+mT} = \psi_{\text{FA}}^{(L)}\left(\psi_{\text{FE}}^{(L)}(I_{k+mT}), \psi_{\text{FE}}^{(L)}(I_{k+(m-1)T})\right) \tag{13}$$

where the operators $\psi_{\text{FE}}^{(l)}$ and $\psi_{\text{FA}}^{(l)}$ denote the procedures of feature extraction and feature alignment, respectively, both with the same number of layers L. As illustrated in Figure 2b, during feature extraction, the input is processed by a downsample module at each layer.

Each downsample module consists of two consecutive 2D convolutions with ReLU activation, where the strides are set to 2 and 1, respectively. Omitting the ReLU for brevity, the two convolutions with different strides are denoted as $\mathrm{Conv}_{3\times3,s2}$ and $\mathrm{Conv}_{3\times3,s1}$. The downsampling process is formally expressed in Equation (14).

$$\mathbf{F}_{\text{downsample}}(\mathbf{X}) = \mathrm{Conv}_{3\times3,s2}(\mathrm{Conv}_{3\times3,s1}(\mathbf{X})) \tag{14}$$

The feature maps of the neighbor at different scales $\{\mathbf{N}_1, \mathbf{N}_2, \ldots, \mathbf{N}_L\}$ and those of the reference $\{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_L\}$ are then jointly fed into the procedure of feature alignment, where the scale of feature maps decreases as the index increases.

As shown in Figure 2c, during the procedure of feature alignment, the feature maps from the reference input and the neighbor input are progressively aligned from the bottom-most feature map up to the top. The entire process consists of L steps. The core operation involved is deformable convolutions [56,57], which more effectively capture variations in pose, local deformation, and scale of objects. Omitting ReLU and other activations, let $\mathcal{C}^{(n)}$ denote n consecutive 3×3 convolutions with stride 1, and let $\theta(\cdot)$ denote the deformable convolution operation.

In the first step, the primary operation is auxiliary refinement of the bottom-level feature maps to facilitate feature alignment in the upper layers. The corresponding formulas are shown in Equations (15) and (16).

$$\mathbf{B}_L = \mathcal{C}^{(2)}(\mathrm{Concat}(\mathbf{N}_L, \mathbf{R}_L)) \tag{15}$$

$$\mathbf{S}_L = \mathrm{Up}\Big(\mathcal{C}^{(1)}(\theta(\mathbf{B}_L, \mathbf{R}_L))\Big) \tag{16}$$

In the following step l, where $l \in [2, L-1]$, an intermediate output $\mathbf{S}_l$ is generated after upsampling at the end of the step, and the model performs feature alignment by leveraging the upsampled output from the previous step. The corresponding formulas are given in Equations (17) and (18).

$$\mathbf{B}_l = \mathcal{C}^{(2)}\Big(\mathrm{Concat}\Big(\mathcal{C}^{(1)}(\mathrm{Concat}(\mathbf{N}_l, \mathbf{R}_l)), \mathbf{S}_{l+1}\Big)\Big) \tag{17}$$

$$\mathbf{S}_l = \mathrm{Up}\Big(\mathcal{C}^{(1)}(\mathrm{Concat}(\theta(\mathbf{B}_l, \mathbf{R}_l), \mathbf{S}_{l+1}))\Big) \tag{18}$$

Finally, the last step generates the aligned feature output $\mathbf{O}_{\text{alignment}}$, which is obtained as shown in Equations (19) and (20).

$$\mathbf{B}_1 = \mathcal{C}^{(2)}\Big(\mathrm{Concat}\Big(\mathcal{C}^{(1)}(\mathrm{Concat}(\mathbf{N}_1, \mathbf{R}_1)), \mathbf{S}_2\Big)\Big) \tag{19}$$

$$\mathbf{O}_{\text{alignment}} = \mathcal{C}^{(1)}(\mathrm{Concat}(\theta(\mathbf{B}_1, \mathbf{R}_1), \mathbf{S}_2)) \tag{20}$$

After processing all $M_k + 1$ frames in $\mathcal{G}_k$, the group yields a set of refined output features:

$$\mathcal{O}_k = \{O_k, O_{k+T}, \ldots, O_{k+M_k T}\} \tag{21}$$

3.3.3. Temporal Reassembly

In the final stage, the outputs from all T groups are interleaved and sorted according to their original temporal indices to reconstruct a complete, coherent feature sequence of length D:

$$\mathbf{Y}_{\text{SSM}} = \mathrm{Merge}(\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_T) \tag{22}$$

where $\mathrm{Merge}(\cdot)$ is a simple concatenation and sorting operation that reconstructs the full feature sequence $\{O_1, O_2, \ldots, O_D\}$, preserving the temporal order.

4. Results

In this section, we first describe the experimental setup, including evaluation metrics and implementation details. Then, we evaluate our method against other baseline and state-of-the-art approaches on both SIRST and MIRST tasks. Finally, we present ablation studies to validate the effectiveness of our design.

4.1. Datasets and Implementation

4.1.1. Datasets

To evaluate the performance of the proposed FeedbackSTS-Det, we employ two publicly available benchmark datasets. The NUDT-MIRSDT [25] dataset comprises 120 sequences with 12,000 images, covering diverse infrared scenes such as clouds, oceans, and land surfaces. Following the original setup, each sequence is generated by applying local perturbations and synthetic noise to real infrared imagery. The IRSatVideo-LEO [26] dataset contains 200 sequences with 91,022 images, constructed by combining real satellite imagery with simulated satellite motion trajectories. Both datasets incorporate synthetic augmentations, specifically perturbation injection in NUDT-MIRSDT and trajectory simulation in IRSatVideo-LEO, while providing high-quality pixel-level mask annotations. Due to the scarcity of publicly available real-world sequential data, the use of synthetic or semi-synthetic components is common in the multi-frame infrared small target detection community; nevertheless, these datasets are built upon real infrared imagery and complemented by dense pixel-level annotations of extremely high precision, making them representative of practical scenarios. Importantly, to ensure fair evaluation, we conduct all experiments strictly following the standard protocols without any dataset-specific customization. We further validate the robustness of our model under different sampling configurations and sequence lengths in the ablation studies (Section 4.3).

4.1.2. Metrics

In this paper, we adopt a mainstream evaluation framework [13,14] for infrared small target detection models, incorporating pixel-level metrics (intersection over union (IoU), F-measure, false alarm rate (Fa)) and object-level metrics (detection rate (Pd), along with the receiver operating characteristic (ROC)).

4.1.3. Implementation Details

For training, we use Soft-IoU loss [57] and the Adam optimizer with a MultiStepLR scheduler, starting at 1×10^{-5} and halving at epochs $[5, 10, 15, 20, 25, 30]$. Training lasts 30 epochs from scratch, implemented in PyTorch 2.4.1 on six NVIDIA RTX 4090 GPUs. Additionally, both datasets then undergo the same augmentations, including random cropping, horizontal flipping, vertical flipping, and rotation, before being resized to 256×256. The total batch size is set to 16.

In terms of data preprocessing, both datasets then undergo the same augmentations (horizontal/vertical/channel flips, transpositions) before resizing to 256×256. Moreover, to ensure a consistent evaluation protocol, we exclude the "Mix" folder from NUDT-MIRSDT, which contains simulated sequences, to focus evaluation on authentic infrared sequences from its "images" and "masks" folders.

Regarding the network architecture, we set the number of layers L in both the feature extraction and feature alignment procedures of the BFBM to 2, as mentioned in Section 3.3, which achieves a good balance between alignment accuracy and computational cost. Furthermore, the sampling step T of the SSM is a key hyperparameter. While larger values of T can further reduce computational cost, excessively reducing the number of propagation steps inevitably compromises the effectiveness of feature propagation, leading to degraded

detection performance. Therefore, T must balance computational efficiency with sufficient temporal feature propagation. Our subsequent experiments demonstrate that values of 2, 3, or 4 yield strong performance, effectively trading off between temporal coverage and processing efficiency.

4.2. SOTA Comparisons

We compare the proposed FeedbackSTS-Det against state-of-the-art methods in both single-frame and multi-frame infrared small target detection. For single-frame infrared small target detection, evaluated methods include three model-driven approaches: TopHat [5], NRAM [9], and PSTNN [38]; and seven data-driven methods: ACM [1], ALCNet [58], DNANet [14], RDIAN [18], ISTDU-Net [16], RPCANet [21], and L2SKNet_FPN [19]. For multi-frame infrared small target detection, comparisons include seven model-driven methods (MSLSTIPT [10], NFTDGSTV [12], RCTV [43], IMNN-LWEC [59], SRSTT [44], ASTTV-NTLA [42], 4DISDT [45]) and five data-driven methods (DNANet_DTUM [25], ALCNet_DTUM [25], ResUNet_DTUM [25], ACM_RFR [26], ResUNet_RFR [26]).

For data-driven MIRST methods, all metrics follow a uniform input sequence length of 10. Additionally, our networks with different sampling intervals t are denoted as FeedbackSTS-Det-Tt. When $t = 1$, the model bypasses interval sampling in SSM. All models are retrained from scratch on both the IRSatVideo-LEO [26] and NUDT-MIRSDT [25] datasets under identical settings for fair comparison.

4.2.1. Quantitative Results

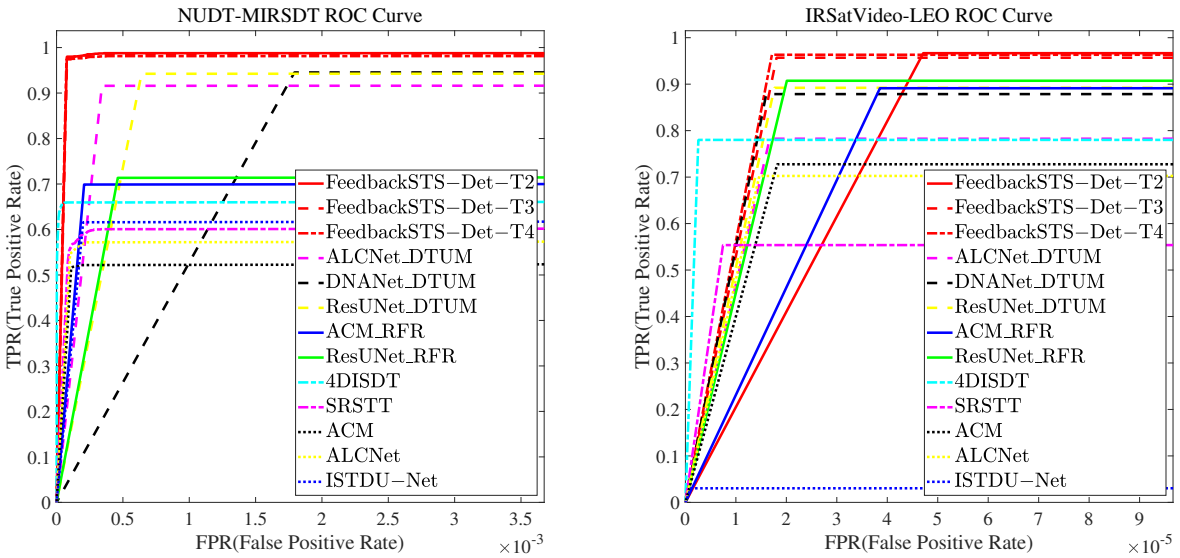
Table 1 presents a comprehensive comparison of state-of-the-art methods on IRSatVideo-LEO [26] and NUDT-MIRSDT [25] using multiple evaluation metrics. Our proposed FeedbackSTS-Det variants consistently demonstrate superior performance across both datasets, with each sampling interval configuration exhibiting distinct advantages. The FeedbackSTS-Det-T2 variant achieves the most balanced and outstanding performance overall. On the NUDT-MIRSDT [25] dataset, it obtains the best results in four key metrics, $mIoU$, F_1, P_d, and AUC, while maintaining the second-lowest false alarm rate. Similarly, on the IRSatVideo-LEO [26] dataset, it delivers competitive detection accuracy with an exceptionally low false alarm rate of 5.5×10^{-7}, indicating its strong robustness in practical scenarios. The FeedbackSTS-Det-T3 and FeedbackSTS-Det-T4 variants also show competitive performance, particularly in specific metrics. FeedbackSTS-Det-T3 achieves the second-best performance in P_d and AUC on NUDT-MIRSDT [25], while FeedbackSTS-Det-T4 excels on IRSatVideo-LEO [26] with the highest P_d of 0.9648 and AUC of 0.9816. This performance progression across sampling intervals reveals an important trade-off: larger intervals reduce computational complexity while maintaining competitive detection accuracy.

Notably, all three FeedbackSTS-Det variants significantly outperform both traditional model-driven methods and contemporary data-driven approaches. While methods like DNANet_DTUM [25] and ResUNet_RFR [26] show respectable results, they are consistently surpassed by our proposed framework across virtually all metrics. This performance advantage is achieved while maintaining reasonable computational requirements, with all FeedbackSTS-Det variants utilizing only 5.68 M parameters and FLOPs decreasing from 79.39 G to 68.50 G as the sampling interval increases. The consistent superiority of FeedbackSTS-Det across different sampling configurations validates the effectiveness of our architectural design and demonstrates its robustness for infrared small target detection in diverse operational scenarios.

Table 1. The results of $mIoU(\times 10^{-2})$, $F_1(\times 10^{-2})$, $P_d(\times 10^{-2})$, $F_a(\times 10^{-6})$, and $AUC(\times 10^{-2})$ are achieved by different methods on the datasets IRSatvideo-LEO [26] and NUDT-MIRSDT [25]. "Params." represents the number of parameters. FLOPs are calculated based on an input sequence scale of $11 \times 256 \times 256$. The best-performing results are indicated in red font, while the second-best results are shown in blue font. The up/down arrows indicate that higher/lower values are better.

Method	Params (M)	FLOPs (G)	NUDT-MIRSDT [25]					IRSatVideo-LEO [26]				
			$mIoU \uparrow$	$F_1 \uparrow$	$P_d \uparrow$	$F_a \downarrow$	$AUC \uparrow$	$mIoU \uparrow$	$F_1 \uparrow$	$P_d \uparrow$	$F_a \downarrow$	$AUC \uparrow$
TopHat [5]	-	-	4.32	8.29	15.42	70.20	68.66	4.18	8.03	27.46	3.26	63.73
NRAM [9]	-	-	2.98	5.78	8.68	26.80	59.89	1.51	2.97	9.45	0.20	54.68
PSTNN [38]	-	-	13.03	23.06	21.62	84.92	64.64	3.50	6.76	46.29	61.31	73.14
ACM [1]	0.398	4.426	36.94	53.95	51.00	39.91	76.08	30.53	46.78	72.87	7.33	86.38
ALCNet [58]	0.427	4.157	32.26	48.79	54.65	65.38	78.58	31.36	47.75	70.27	5.21	85.12
DNA-Net [14]	4.697	71.306	27.51	43.15	60.00	211.94	80.04	0.13	0.27	0.92	0.21	50.44
RDIAN [18]	0.217	40.902	19.33	32.40	61.06	509.90	80.74	0.10	0.19	6.73	2009.97	53.67
ISTDU-Net [16]	2.752	87.388	33.45	50.13	60.88	111.24	80.78	0.47	0.93	3.09	0.96	51.51
RPCANet [21]	0.680	490.255	0.02	0.05	10.47	495,831.30	73.74	0.00	0.00	0.65	499,730.39	63.20
L2SKNet_FPN [19]	1.071	66.085	21.12	34.88	43.06	92.03	71.85	1.09	2.15	47.67	362.75	73.23
MSLSTIPT [10]	-	-	13.76	24.19	21.66	125.26	76.25	4.68	8.94	36.71	12.41	68.35
NFTDGSTV [12]	-	-	2.41	4.71	5.50	307.22	64.46	0.99	1.97	50.16	191.84	75.06
RCTV [43]	-	-	0.34	0.68	0.18	1.40	52.55	8.74	16.07	39.31	0.93	69.65
IMNN-LWEC [59]	-	-	6.04	11.40	13.51	209.38	64.57	1.51	2.97	46.00	265.21	72.97
SRSTT [44]	-	-	12.18	21.72	21.16	444.36	80.01	12.68	22.51	55.40	6.63	77.68
ASTTV-NTLA [42]	-	-	0.78	1.55	0.36	52.54	53.84	0.80	1.59	50.88	206.00	75.42
4DISDT [45]	-	-	10.66	19.26	15.92	105.03	82.97	23.13	37.58	78.16	1.81	89.00
DNANet_DTUM [25]	1.205	97.757	10.44	18.91	94.29	1726.77	97.18	38.88	55.99	88.24	6.31	93.93
ALCNet_DTUM [25]	0.842	16.897	35.67	52.58	91.82	175.71	95.78	35.49	52.39	79.18	2.53	89.13
ResUNet_DTUM [25]	0.298	22.772	24.33	39.13	95.65	499.13	97.09	38.22	55.30	89.52	6.58	94.61
ACM_RFR [26]	0.504	40.842	29.98	46.13	69.24	135.79	84.92	24.68	39.59	89.18	16.88	94.56
ResUNet_RFR [26]	1.012	79.801	30.98	47.31	70.94	286.67	85.64	37.49	54.54	90.82	6.02	95.36
FeedbackSTS-Det-T2 (Ours)	5.678	79.386	52.24	68.63	97.41	14.42	99.38	43.50	60.63	95.28	0.55	97.46
FeedbackSTS-Det-T3 (Ours)	5.678	73.941	51.55	68.03	97.29	19.23	99.32	41.96	59.11	95.83	4.51	97.84
FeedbackSTS-Det-T4 (Ours)	5.678	68.496	51.95	68.38	96.65	18.45	99.05	43.42	60.54	96.48	3.07	98.16

The ROC results are presented in Figure 3. As shown in Figure 3a, all FeedbackSTS−Det variants achieve a faster increase in TPR, reaching 1 more quickly than other models. In the more challenging IRSatVideo-LEO dataset (Figure 3b), although FeedbackSTS-Det-T2 exhibits a slower initial TPR rise, it ultimately reaches the maximum value faster. Meanwhile, FeedbackSTS-Det-T3 and FeedbackSTS-Det-T4 maintain strong TPR growth throughout the entire phase, confirming their robust detection performance.



(**a**) (**b**)

Figure 3. ROC curves on two benchmark datasets: (**a**) ROC Curve on NUDT-MIRSDT [25]; (**b**) ROC Curve on IRSTVideo-LEO [26].

4.2.2. Qualitative Results

Qualitative results are presented in Figure 4. The selected examples are organized as follows: rows 1–2 illustrate easy scenes from IRSatVideo-LEO [26]; rows 3–4 depict medium scenes; rows 5–6 present challenging scenes from the same dataset; and rows 7–8 show sequence samples from NUDT-MIRSDT [25].

The scenes encompass satellite, aerial, and maritime imagery captured from a vertical perspective. It can be observed that traditional model-driven methods, including low-rank and sparse representation techniques such as PSTNN [38] and SRSTT [44], exhibit limited detection performance in these scenes. The typical single-frame low-rank and sparse infrared small target detection method, PSTNN [38], produces significant false alarms in most satellite nadir-view scenes, except for the sequence "EastNorthAsia25_43", and fails to detect targets correctly in "Sequence92" and "Sequence96". SRSTT [44], as a representative multi-frame low-rank and sparse detection method, generates substantial false alarms across all sequences and exhibits missed detections in several satellite view sequences. Data-driven multi-frame methods show noticeable improvement, yet still exhibit certain limitations. Except for "Sequence96", ALCNet_DTUM [25] produces false alarms in all other sequences, with particularly severe cases in "Sequence92". RFR-based [26] frameworks (e.g., ResUNet_RFR [26] and ACM_RFR [26]) yield significant false alarms in sequences such as "EastNorthAsia0_97", "EastAustralia3_94", "EastAustralia5_07", and "EastNorthAsia25_43". However, ACM_RFR [26] achieves relatively good detection performance on "Sequence96". In contrast, it can be observed that all variants of our FeedbackSTS-Det achieve consistently strong performance across all scenarios and difficulty levels, specifically in terms of accurate target localization, low false alarm rate, and morphological accuracy of detected targets.

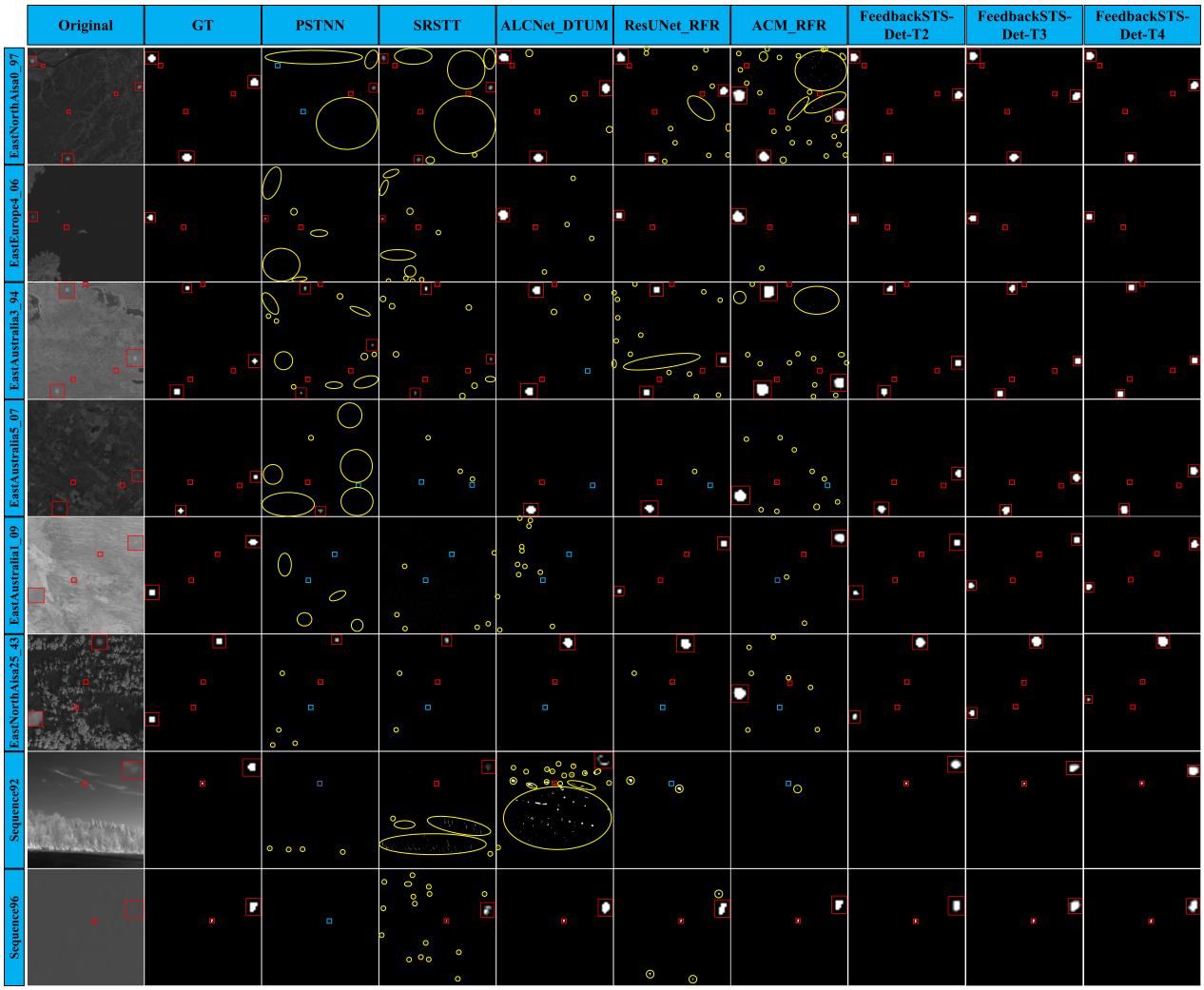


Figure 4. Qualitative results of different methods. For better visual presentation, the target regions are highlighted with red boxes and then displayed as zoomed-in views. Missed detections and false alarms are marked with blue and yellow circles, respectively.

4.3. Ablation Study

This subsection presents experiments on FeedbackSTS-Det and its variants, aiming to verify the potential advantages of our proposed modules and key design decisions. If not stated explicitly, we set the temporal sampling interval to 2.

4.3.1. Ablation on Spatio-Temporal Semantic Feedback Strategy

Introduced in Section 3.2, the spatio-temporal semantic feedback strategy is designed to strengthen inter-frame feature linkages for better infrared small target detection. Our approach is built upon a 3D U-Net backbone [55]. A systematic comparison of the ablated variants of this feedback design is provided in Table 2. This ablation study is designed to validate the effectiveness of our proposed framework by contrasting it against several alternative configurations, including unidirectional (All-Fwd, All-Bwd), partial semantic feedback (Part-FB1, Part-FB2), and feedback-disabled schemes (Dec-NoFB, Enc-NoFB).

As summarized in Table 2, while simpler configurations like Enc-NoFB achieve the lowest parameter count and FLOPs, our optimal method, Full-FB, strikes a critical balance. It maintains full feature interaction capabilities in both the encoder and decoder, incurring only a modest increase in computational cost compared to the most constrained variants. Subsequent experiments confirm that the proposed spatio-temporal semantic feedback

framework is a necessary and efficient component for accurate target localization, shape delineation, and false alarm suppression.

Table 2. Ablation study on feedback design variants: configuration and efficiency. "Params." represents the number of parameters. FLOPs are calculated based on an input sequence scale of $5 \times 256 \times 256$. The row corresponding to the proposed spatio-temporal semantic feedback framework is highlighted with a gray background. The down arrows indicate that lower values are better.

Function & Characteristic	Abbr.	Encoder	Decoder	Params ↓	FLOPs ↓
Decoder without semantic feedback	Dec-NoFB	$[+,+,+,+,+]$	$[\sim,\sim,\sim,\sim]$	5.046 M	26.979 G
Encoder without semantic feedback	Enc-NoFB	$[\sim,\sim,\sim,\sim,\sim]$	$[-,-,-,-]$	2.748 M	22.684 G
All forward semantic feedback	All-Fwd	$[+,+,+,+,+]$	$[+,+,+,+]$	5.678 M	30.144 G
All backward semantic feedback	All-Bwd	$[-,-,-,-,-]$	$[-,-,-,-]$	5.678 M	30.144 G
Partial semantic feedback (pattern 1)	Part-FB1	$[+,\sim,+,\sim,+]$	$[-,\sim,-,\sim]$	4.560 M	26.559 G
Partial semantic feedback (pattern 2)	Part-FB2	$[+,\sim,+,\sim,+]$	$[\sim,-,\sim,-]$	4.876 M	25.343 G
Full semantic feedback (Ours)	Full-FB	$[+,+,+,+,+]$	$[-,-,-,-]$	5.678 M	30.144 G

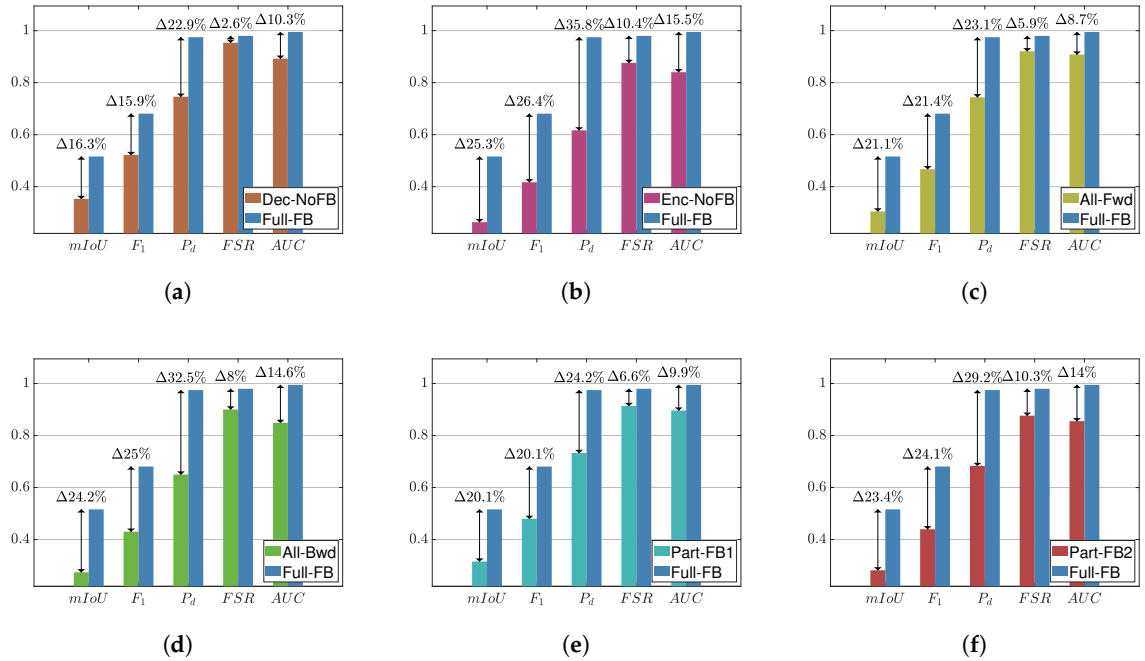
Symbols: $+$: FSTSRM, $\sim$: 3D Conv Block, $-$: BSTSRM.

For consistent performance evaluation where higher values indicate better results across all metrics, including $mIoU$, F_1, P_d, and AUC, we introduce the false alarm suppression rate (FSR), as formulated in Equation (19), where F_a denotes the false alarm rate. Consequently, a higher FSR value indicates better performance, consistent with the other metrics and sharing the same order of magnitude by using the scaling factor 10^3.

$$FSR = 1 - F_a \times 10^3 \tag{23}$$

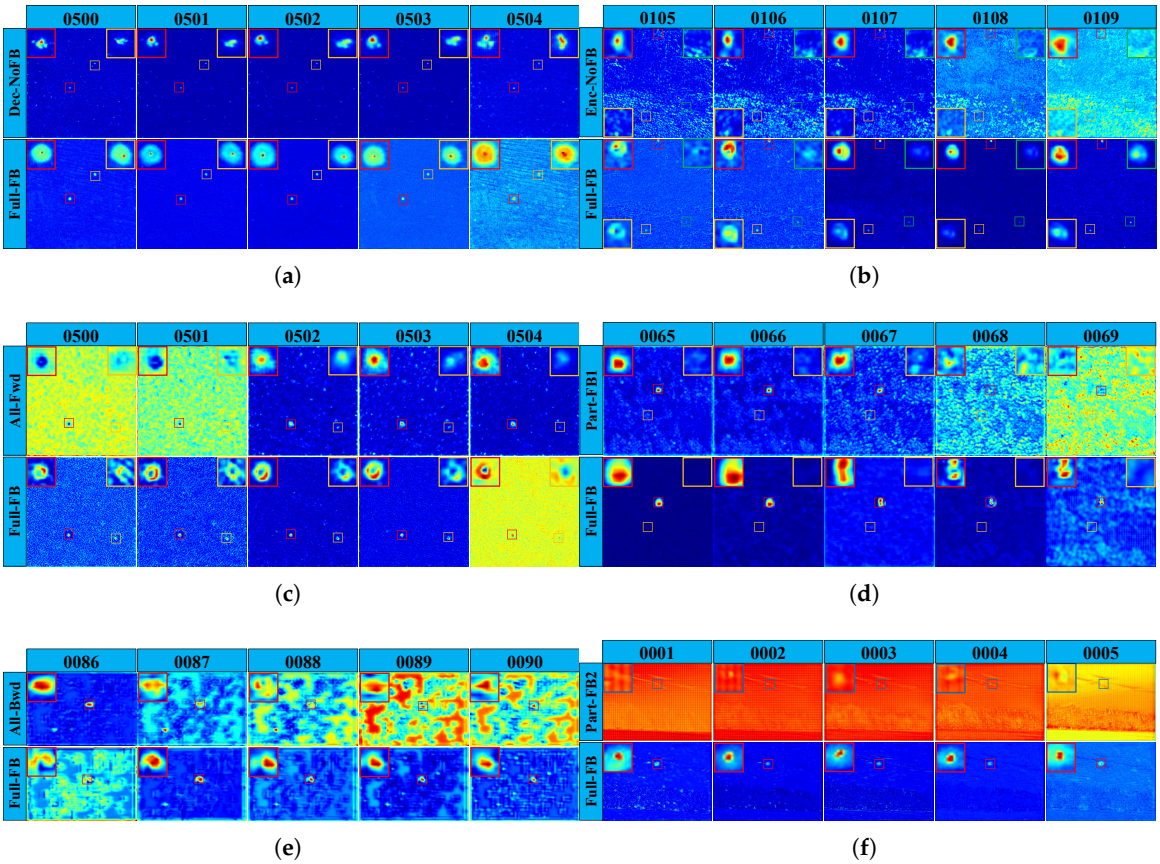
As illustrated in Figure 5, we benchmark all spatio-temporal semantic feedback variants from Table 2 against our optimal method, Full-FB, on the NUDT-MIRSDT [25] dataset. The evaluation uses five metrics ($mIoU$, F_1, P_d, FSR, AUC) under a fixed input size of $5 \times 256 \times 256$, and results show the clear superiority of our method across all metrics.

The visual comparisons corresponding to the methods in Table 2 are provided in Figure 6. For clarity, targets are annotated with bounding boxes positioned at image corners, with distinct colors assigned to different targets for easy cross-reference. The effectiveness of backward feedback in the decoder is demonstrated by comparing Full-FB with alternative designs. First, in Figure 6a, Full-FB shows progressively stronger suppression and clearer targets during backward feedback (frames #0504 to #0500), while Dec-NoFB performs poorly. Second, Figure 6b reveals that the All-Fwd method yields unclear targets in initial frames (#0500, #0501) during forward feedback, whereas Full-FB exhibits this limitation only in the initial frame (#0504) of its backward process. Moreover, in Figure 6c,e, targets in the second row (Full-FB) are consistently clearer than those in the first row (the compared methods), highlighting the benefit of forward feedback in the encoder. Finally, Figure 6d,f indicate that removing spatio-temporal semantic feedback modules in certain layers degrades target feature extraction, further underscoring the necessity of the complete feedback framework.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 5. Ablation study of different feedback design variants against Full-FB. Evaluation metrics include $mIoU$, F_1, P_d, FSR, and AUC, where Δ represents the performance difference relative to the Full-FB: (**a**) Dec-NoFB vs. Full-FB; (**b**) Enc-NoFB vs. Full-FB; (**c**) All-Fwd vs. Full-FB; (**d**) All-Bwd vs. Full-FB; (**e**) Part-FB1 vs. Full-FB; (**f**) Part-FB2 vs. Full-FB.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 6. Ablation experiments show the visual comparison of different spatio-temporal semantic feedback approaches. In this experiment, the better the background clutter is suppressed, the more

prominently the color patches appear in blue. The better the target regions are extracted, the more the color patches tend towards red. (**a**) Visual comparison of Dec-NoFB versus Full-FB, evaluated on the IRSatVideo-LEO [26]. (**b**) Visual comparison of Enc-NoFB versus Full-FB, evaluated on the IRSatVideo-LEO [26]. (**c**) Visual comparison of All-Fwd versus Full-FB, evaluated on the IRSatVideo-LEO [26]. (**d**) Visual comparison of Part-FB1 versus Full-FB, evaluated on the IRSatVideo-LEO [26]. (**e**) Visual comparison of All-Bwd versus Full-FB, evaluated on the NUDT-MIRSDT [25]. (**f**) Visual comparison of Part-FB2 versus Full-FB, evaluated on the NUDT-MIRSDT [25].

4.3.2. Ablation on Sparse Grouping Method

To evaluate the effectiveness of our sparse grouping strategy within the SSM, we redesign several alternative methods for the construction of sparse frames, which are summarized as follows. Among them, the last one is the method we ultimately adopt for grouping frames.

- Group-Rand-Seq: We randomly segment the input continuous frames into contiguous segments and perform alignment only between consecutive frames within each segment.
- Group-Rand-Step-Fixed: We randomly determine a global sampling interval and uniformly adopt this interval across the network to achieve skip-frame alignment.
- Group-Rand-Step-Rand: We assign a random sampling interval to each layer, enabling skip-frame alignment with layer-wise randomness.
- Group-Fixed-Step-T2: We manually preset the sampling interval to $T = 2$ as our final configuration to achieve skip-frame alignment.

As shown in Table 3, among all variants, Group-Fixed-Step-T2 achieves the best overall performance, attaining the highest $mIoU$ (48.97), F_1 (65.74), P_d (96.34), and AUC (98.09), while maintaining a competitive false alarm rate (1.19). Group-Rand-Step-Fixed delivers strong second-best results in P_d (92.35), F_a (1.18), and AUC (96.10), demonstrating the effectiveness of a globally consistent sampling interval. Notably, Group-Rand-Step-Rand achieves slightly higher $mIoU$ and F_1 than Group-Rand-Step-Fixed but lags in P_d and AUC, indicating that layer-wise random sampling may disrupt stable temporal propagation. In contrast, Group-Rand-Seq shows the weakest performance across all metrics, suggesting that purely contiguous segmentation without skip-frame propagation limits the model's ability to capture long-range dependencies. These results validate our final design choice of fixed-step sampling as the optimal configuration for balancing temporal coverage and propagation effectiveness.

Table 3. Ablation results of $mIoU(\times 10^{-2})$, $F_1(\times 10^{-2})$, $P_d(\times 10^{-2})$, $F_a(\times 10^{-6})$, and $AUC(\times 10^{-2})$ are achieved by different frame alignment grouping strategies on the dataset IRSatvideo-LEO [26]. The rows containing our final proposed framework are distinguished with gray background highlighting. The best-performing results are bolded, while the second-best results are underlined. The up/down arrows indicate that higher/lower values are better.

Framework	$mIoU \uparrow$	$F_1 \uparrow$	$P_d \uparrow$	$F_a \downarrow$	$AUC \uparrow$
Group-Rand-Seq	32.19	48.70	88.16	5.59	94.00
Group-Rand-Step-Fixed	43.49	60.62	<u>92.35</u>	**1.18**	<u>96.10</u>
Group-Rand-Step-Rand	<u>43.78</u>	<u>60.90</u>	88.63	1.23	94.26
Group-Fixed-Step-T2	**48.97**	**65.74**	**96.34**	<u>1.19</u>	**98.09**

4.3.3. Ablation on Sampling Interval in SSM

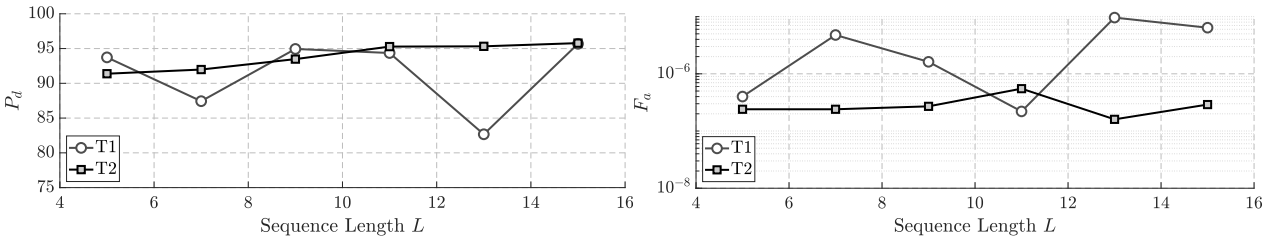
In this experiment, we compare the impact of different temporal sampling intervals of the SSM module on performance. For clarity, we use T1 to denote the baseline without temporal sampling, and T2, T3, and T4 as shorthands for FeedbackSTS-Det-T2, -T3,

and -T4, with sparse frame sampling intervals of 2, 3, and 4, respectively. The computational efficiency of the model under different settings is summarized in Table 4, which reports FLOPs and FPS for various sequence lengths L and temporal sampling intervals T based on an input image scale of 256×256. The results demonstrate that for a fixed T, longer sequences L lead to higher FLOPs and generally lower FPS. Conversely, increasing the temporal sampling interval T consistently reduces computational cost and improves inference speed across all sequence lengths. For instance, when the sequence length is set to $L = 9$, the computational cost measured in FLOPs decreases from 68.42 G for $T = 1$ to 52.08 G for $T = 4$, while the inference speed (FPS) increases from 7.5 to 13.7, representing a substantial improvement of 82.6% in processing efficiency. This analysis confirms that larger temporal sampling intervals enhance computational efficiency, especially for long sequences.

To evaluate the efficacy of the temporal sampling mechanism, we conduct a comparative analysis between the T2 condition (with a temporal sampling interval of length 2) and the T1 condition (without temporal sampling) on the IRSatVideo-LEO [26], using P_d and F_a as key metrics. As shown in Figure 7a, the P_d for T2 increases monotonically with the sequence length L, whereas the P_d for T1 exhibits significant and unstable fluctuations. It is observed from Figure 7b that the F_a for T2 remains consistently stable within a narrow range as L increases. In contrast, the F_a for T1 varies drastically and is markedly higher than that of T2. In summary, the sparse frame sampling module significantly enhances target detection capability while effectively suppressing false alarms over long sequences, demonstrating its robust ability to propagate inter-frame features compared to the baseline T1 condition.

Table 4. Ablation study of computational efficiency (FLOPs and FPS) for varying sequence lengths L and sparse frame sampling intervals T. Both FLOPs and FPS are calculated based on an input image scale of 256×256. The up/down arrows indicate that higher/lower values are better.

L	$T = 1$		$T = 2$		$T = 3$		$T = 4$	
	FLOPs ↓	**FPS ↑**	**FLOPs ↓**	**FPS ↑**	**FLOPs ↓**	**FPS ↑**	**FLOPs ↓**	**FPS ↑**
5	35.59 G	8.4	30.14 G	10.7	24.70 G	14.2	19.25 G	15.1
7	52.00 G	8.0	46.56 G	9.8	41.11 G	11.2	35.67 G	15.2
9	68.42 G	7.5	62.97 G	8.9	57.53 G	10.5	52.08 G	13.7
11	84.83 G	7.6	79.39 G	8.3	73.94 G	8.9	68.50 G	9.7
13	101.245 G	7.4	95.80 G	7.6	90.36 G	9.1	84.91 G	10.7
15	117.66 G	7.7	112.214 G	7.8	106.77 G	8.9	101.32 G	9.8



(**a**) (**b**)

Figure 7. Ablation study for the IRSatVideo-LEO [26] under no temporal sampling and T2 conditions: (**a**) P_d vs. L; (**b**) F_a vs. L.

To investigate the capability of different sparse frame sampling intervals in extracting targets under varying input window sizes, we evaluate the effect of the input sequence length L on the detection probability P_d across different sampling intervals using IRSatVideo-LEO [26], as shown in Figure 8. It could be observed that P_d increases with L

for all sampling intervals, validating that the sparse frame sampling module successfully enhances inter-frame dependencies in long sequences. Moreover, a longer sampling interval is generally associated with improved performance at an equivalent sequence length. An exception is noted at $L = 11$, where T4 underperforms relative to T3; however, T4 demonstrates superior performance at $L = 13$.

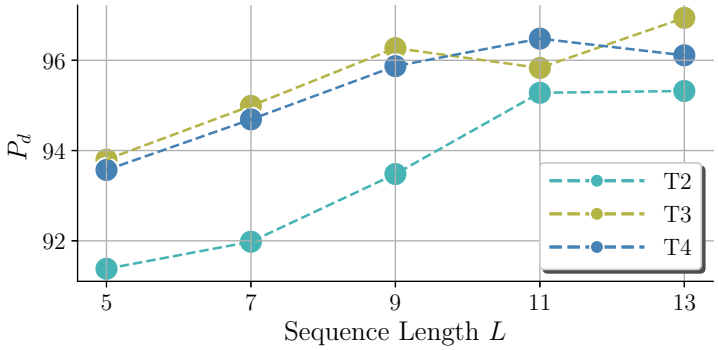


Figure 8. Ablation results of sequence length effect on P_d under different sampling intervals for the IRSatVideo-LEO [26].

In summary, the experimental results suggest that increasing the sampling interval moderately reduces the network's computational cost while improving its temporal propagation capability across long frame sequences. However, an overly large sampling interval T can compromise detection accuracy. Our experimental results confirm that choosing $T = 2$, 3, or 4 reduces computational cost to a certain extent while preserving or even improving the detection performance of the sequence. To be more specific, for scenarios prioritizing accuracy, $T = 2$ is recommended. For deployment on platforms with strict computational limits, such as UAVs or edge devices, a larger T, specifically 3 or 4, is preferable to balance efficiency and detection performance.

5. Discussion

Although the proposed FeedbackSTS-Det shows clear improvements in moving infrared small target detection, several aspects deserve further discussion. By incorporating a closed-loop spatio-temporal semantic feedback strategy with paired forward and backward refinement modules, our method enhances inter-frame information exchange and enables retrospective error correction. This bidirectional design outperforms both unidirectional and feedback-disabled variants as confirmed by our ablation study. Additionally, the embedded sparse semantic module (SSM) further reduces computational cost while improving long-range temporal dependency modeling, achieving a favorable trade-off between accuracy and efficiency. These strengths collectively contribute to robust performance across diverse scenes, including satellite nadir-view, aerial, and maritime imagery.

Despite the promising results, the proposed method has several limitations that may lead to unstable behavior under specific conditions.

First, the SSM relies on a fixed sampling interval T chosen from $\{2, 3, 4\}$, which is not adaptive to scene dynamics or target motion speed. Consequently, the model may struggle in scenarios where target motion is highly non-uniform, such as sudden acceleration or deceleration, as a fixed interval cannot optimally capture such varying motion patterns.

Second, the robustness of our model under extremely challenging conditions remains to be validated. Specifically, we identify the following potential failure scenarios: (a) low signal-to-noise ratio (SNR) conditions, where targets are nearly immersed in background noise and even temporal cues may become unreliable; (b) dense target swarms, where multiple small targets appear in close proximity, potentially causing confusion in the feature

alignment process; and (c) rapidly changing backgrounds, such as scenes with sudden illumination variations or fast-moving clutter, which may violate the temporal consistency assumptions underlying our feedback strategy.

Third, the current framework assumes fixed-length sliding windows. In scenarios where targets appear intermittently or with long gaps, the fixed window size may either miss relevant temporal context or introduce redundant computation.

To address these limitations, future work will focus on developing an adaptive sparse sampling mechanism that dynamically adjusts the sampling interval T based on the inferred motion type of infrared targets. Additionally, we plan to evaluate our model on more challenging benchmarks specifically designed for low-SNR and dense-target scenarios, as well as on real-world datasets, to further assess its practical applicability. Extending the framework to cross-domain adaptation with unlabeled target scenes and multi-source infrared data is also a promising direction. Furthermore, we intend to incorporate interpretable neural network theory to better understand when and why the model fails, thereby guiding further improvements. Building on this, we will further develop the model to fully leverage multi-frame radiometric characteristics, unlocking the potential of FeedbackSTS-Det for robust and scalable moving infrared small target detection in real-world applications.

6. Conclusions

This paper presented a novel sparse-frames-based spatio-temporal feedback network, termed FeedbackSTS-Det, for moving infrared small target detection. Its core is a closed-loop semantic feedback strategy with forward and backward refinement modules that work jointly, enabling information exchange between consecutive frames, thereby improving detection accuracy and reducing false alarms. An embedded sparse semantic module (SSM) performs structured sparse temporal modeling to capture long-range dependencies with low computational cost. Extensive experiments on widely adopted multi-frame infrared small target datasets demonstrate the effectiveness of our proposed method under the evaluated scene configurations. While the proposed method achieves strong performance, it has certain limitations, including the fixed sampling interval and the assumption of fixed-length sliding windows. Deployment on resource-constrained platforms also requires careful consideration of the trade-off between accuracy and efficiency. To address these issues, future work will focus on developing an adaptive sampling mechanism and integrating interpretable network designs.

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

Funding: This research was funded in part by the Natural Science Foundation of Sichuan Province of China under Grant 2025ZNSFSC0522, and in part by the National Natural Science Foundation of China under Grant 61571096.

Data Availability Statement: The NUDT-MIRSDT dataset used in this study is publicly available at https://github.com/TinaLRJ/Multi-frame-infrared-small-target-detection-DTUM, last accessed on 10 May 2025. The IRSatVideo-LEO dataset is publicly available at https://github.com/XinyiYing/RFR, last accessed on 15 May 2025. The source code and models for FeedbackSTS-Det are available at https://github.com/IDIP-Lab/FeedbackSTS-Det (accessed on 13 June 2026).

Acknowledgments: The authors would like to thank the University of Electronic Science and Technology of China for their support. The authors also thank Li et al. for providing the NUDT-MIRSDT dataset and acknowledge Ying et al. for providing the IRSatVideo-LEO dataset.

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

References

1. 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, Virtual, 5–9 January 2021; pp. 950–959.
2. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713. [CrossRef]
3. Han, Y.; Liao, J.; Lu, T.; Pu, T.; Peng, Z. KCPNet: Knowledge-driven context perception networks for ship detection in infrared imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5000219. [CrossRef]
4. Zhao, Y.; Pang, Y.; Zhang, L.; Liu, H.; Zuo, J.; Lu, H.; Zhao, X. Unimmad: Unified multi-modal and multi-class anomaly detection via moe-driven feature decompression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Denver, CO, USA, 3–7 June 2026; pp. 28502–28511.
5. Tom, V.T.; Peli, T.; Leung, M.; Bondaryk, J.E. Morphology-based algorithm for point target detection in infrared backgrounds. In *Proceedings of the Signal and Data Processing of Small Targets 1993, Orlando, FL, USA, 11–16 April 1993*; SPIE: Bellingham, WA, USA, 1993; Volume 1954, pp. 2–11.
6. Hadhoud, M.M.; Thomas, D.W. The two-dimensional adaptive LMS (TDLMS) algorithm. *IEEE Trans. Circuits Syst.* **1988**, *35*, 485–494. [CrossRef]
7. 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]
8. 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]
9. Zhang, L.; Peng, L.; Zhang, T.; Cao, S.; Peng, Z. Infrared small target detection via non-convex rank approximation minimization joint $l_{2,1}$ norm. *Remote Sens.* **2018**, *10*, 1821. [CrossRef]
10. 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.* **2020**, *59*, 3737–3752.
11. Wang, G.; Tao, B.; Kong, X.; Peng, Z. Infrared small target detection using nonoverlapping patch spatial–temporal tensor factorization with capped nuclear norm regularization. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5001417. [CrossRef]
12. Liu, T.; Yang, J.; Li, B.; Wang, Y.; An, W. Infrared small target detection via nonconvex tensor tucker decomposition with factor prior. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5617317. [CrossRef]
13. Zhang, T.; Li, L.; Cao, S.; Pu, T.; Peng, Z. Attention-guided pyramid context networks for detecting infrared small target under complex background. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 4250–4261. [CrossRef]
14. 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]
15. Wu, X.; Hong, D.; Chanussot, J. UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Trans. Image Process.* **2022**, *32*, 364–376. [PubMed]
16. 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]
17. 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.
18. 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]
19. Wu, F.; Liu, A.; Zhang, T.; Zhang, L.; Luo, J.; Peng, Z. Saliency at the helm: Steering infrared small target detection with learnable kernels. *IEEE Trans. Geosci. Remote Sens.* **2024**, *63*, 5000514. [CrossRef]
20. Hu, C.; Huang, Y.; Li, K.; Zhang, L.; Long, C.; Zhu, Y.; Pu, T.; Peng, Z. DATransNet: Dynamic attention transformer network for infrared small target detection. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 7001005. [CrossRef]
21. Wu, F.; Zhang, T.; Li, L.; Huang, Y.; Peng, Z. RPCANet: Deep unfolding RPCA based infrared small target detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 1–8 January 2024; pp. 4809–4818.
22. Liu, P.; Pang, L.; Peng, J.; Luo, Y.; Liu, J.; Cao, X. CTVNet: Gradient Prior-Guided Deep Unfolding Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5010514. [CrossRef]

23. Xiong, Z.; Zhou, F.; Wu, F.; Yuan, S.; Fu, M.; Peng, Z.; Yang, J.; Dai, Y. DRPCA-Net: Make robust PCA great again for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5005516. [CrossRef]
24. Hou, R.; Yan, P.; Duan, X.; Wang, X. Unsupervised image sequence registration and enhancement for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5620814. [CrossRef]
25. 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.* **2023**, *36*, 555–568.
26. 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]
27. Wu, F.; Liu, S.; Wang, H.; Tao, B.; Luo, J.; Peng, Z. Neural spatial–temporal tensor representation for infrared small target detection. *Pattern Recognit.* **2026**, *169*, 111929. [CrossRef]
28. Li, R.; An, W.; Wang, Y.; Ying, X.; Dai, Y.; Wang, L.; Li, M.; Guo, Y.; Liu, L. Probing Deep into Temporal Profile Makes the Infrared Small Target Detector Much Better. *arXiv* **2025**, arXiv:2506.12766.
29. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2023**, arXiv:2312.00752.
30. 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]
31. Liu, W.; Lu, X.; Zhang, J.; Li, D.; Zhang, X. MOU-Mamba: Multi-Order U-shape Mamba for infrared small target detection. *Opt. Laser Technol.* **2025**, *187*, 112851. [CrossRef]
32. Kim, S.; Lee, J. Scale invariant small target detection by optimizing signal-to-clutter ratio in heterogeneous background for infrared search and track. *Pattern Recognit.* **2012**, *45*, 393–406. [CrossRef]
33. Chen, C.P.; Li, H.; Wei, Y.; Xia, T.; Tang, Y.Y. A local contrast method for small infrared target detection. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 574–581. [CrossRef]
34. Wei, Y.; You, X.; Li, H. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognit.* **2016**, *58*, 216–226. [CrossRef]
35. He, Y.; Li, M.; Zhang, J.; An, Q. Small infrared target detection based on low-rank and sparse representation. *Infrared Phys. Technol.* **2015**, *68*, 98–109. [CrossRef]
36. Wang, X.; Peng, Z.; Kong, D.; He, Y. Infrared dim and small target detection based on stable multisubspace learning in heterogeneous scene. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5481–5493. [CrossRef]
37. Zhang, T.; Peng, Z.; Wu, H.; He, Y.; Li, C.; Yang, C. Infrared small target detection via self-regularized weighted sparse model. *Neurocomputing* **2021**, *420*, 124–148. [CrossRef]
38. Zhang, L.; Peng, Z. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sens.* **2019**, *11*, 382. [CrossRef]
39. Guan, X.; Zhang, L.; Huang, S.; Peng, Z. Infrared small target detection via non-convex tensor rank surrogate joint local contrast energy. *Remote Sens.* **2020**, *12*, 1520. [CrossRef]
40. Kong, X.; Yang, C.; Cao, S.; Li, C.; Peng, Z. Infrared small target detection via nonconvex tensor fibered rank approximation. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5000321. [CrossRef]
41. Liu, H.K.; Zhang, L.; Huang, H. Small target detection in infrared videos based on spatio-temporal tensor model. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8689–8700. [CrossRef]
42. Liu, T.; Yang, J.; Li, B.; Xiao, C.; Sun, Y.; Wang, Y.; An, W. Nonconvex tensor low-rank approximation for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5614718. [CrossRef]
43. Liu, T.; Yang, J.; Li, B.; Wang, Y.; An, W. Representative coefficient total variation for efficient infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5622918. [CrossRef]
44. Li, J.; Zhang, P.; Zhang, L.; Zhang, Z. Sparse regularization-based spatial–temporal twist tensor model for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5000417. [CrossRef]
45. Wu, F.; Yu, H.; Liu, A.; Luo, J.; Peng, Z. Infrared small target detection using spatiotemporal 4-D tensor train and ring unfolding. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5002922. [CrossRef]
46. Zhang, L.; Luo, J.; Huang, Y.; Wu, F.; Cui, X.; Peng, Z. Mdigcnet: Multi-directional information-guided contextual network for infrared small target detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *18*, 2063–2076.
47. Goldfarb, D.; Qin, Z. Robust low-rank tensor recovery: Models and algorithms. *SIAM J. Matrix Anal. Appl.* **2014**, *35*, 225–253. [CrossRef]
48. Liu, P.; Peng, J.; Luo, Y.; Fu, J.; Li, J.; Cao, X. DDFet: Infrared Small Target Detection via a Dual-Domain Fused Deep Unfolding Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 23646–23657. [CrossRef]
49. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.

50. Zhang, M.; Ouyang, Y.; Gao, F.; Guo, J.; Zhang, Q.; Zhang, J. MOCID: Motion context and displacement information learning for moving infrared small target detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 10022–10030.
51. Zhang, M.; Li, X.; Gao, F.; Guo, J. Irmamba: Pixel difference mamba with layer restoration for infrared small target detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 10003–10011.
52. Zhang, S.; Zhang, H.; Cui, Y.; Xiong, R.; He, C.; Feng, D. SMMT: Siamese Motion Mamba with Self-attention for Thermal Infrared Target Tracking. In *Proceedings of the International Conference on Intelligent Computing*; Springer: Singapore, 2025; pp. 242–253.
53. Qiu, X.; Luo, J.; Tian, W.; He, F.; Huang, Y.; Hu, C.; Peng, Z. Local-Motion Mamba with Frequency-Guided for Moving Infrared Small Target Detection. *IEEE Geosci. Remote Sens. Lett.* **2026**, *23*, 6003905. [CrossRef]
54. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
55. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2016; pp. 424–432.
56. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
57. Huang, Y.; Tang, Z.; Chen, D.; Su, K.; Chen, C. Batching soft IoU for training semantic segmentation networks. *IEEE Signal Process. Lett.* **2019**, *27*, 66–70.
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. Luo, Y.; Li, X.; Chen, S.; Xia, C.; Zhao, L. IMNN-LWEC: A novel infrared small target detection based on spatial–temporal tensor model. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5004022. [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.