Spatio-Temporal Context Learning with Temporal Difference Convolution for Moving Infrared Small Target Detection

Houzhang Fang[1*], **Shukai Guo**[1], **Qiuhuan Chen**[1], **Yi Chang**[2], **Luxin Yan**[2]

[1]School of Computer Science and Technology, Xidian University, China
[2]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China
houzhangfang@outlook.com, {shukaiguo, cqh}@stu.xidian.edu.cn, {yichang, yanluxin}@hust.edu.cn

Abstract

Moving infrared small target detection (IRSTD) plays a critical role in practical applications, such as surveillance of unmanned aerial vehicles (UAVs) and UAV-based search system. Moving IRSTD still remains highly challenging due to weak target features and complex background interference. Accurate spatio-temporal feature modeling is crucial for moving target detection, typically achieved through either temporal differences or spatio-temporal (3D) convolutions. Temporal difference can explicitly leverage motion cues but exhibits limited capability in extracting spatial features, whereas 3D convolution effectively represents spatio-temporal features yet lacks explicit awareness of motion dynamics along the temporal dimension. In this paper, we propose a novel moving IRSTD network (TDCNet), which effectively extracts and enhances spatio-temporal features for accurate target detection. Specifically, we introduce a novel temporal difference convolution (TDC) re-parameterization module that comprises three parallel TDC blocks designed to capture contextual dependencies across different temporal ranges. Each TDC block fuses temporal difference and 3D convolution into a unified spatio-temporal convolution representation. This re-parameterized module can effectively capture multi-scale motion contextual features while suppressing pseudo-motion clutter in complex backgrounds, significantly improving detection performance. Moreover, we propose a TDC-guided spatio-temporal attention mechanism that performs cross-attention between the spatio-temporal features extracted from the TDC-based backbone and a parallel 3D backbone. This mechanism models their global semantic dependencies to refine the current frame's features, thereby guiding the model to focus more accurately on critical target regions. To facilitate comprehensive evaluation, we construct a new challenging benchmark, IRSTD-UAV, consisting of 15,106 real infrared images with diverse low signal-to-clutter ratio scenarios and complex backgrounds. Extensive experiments on IRSTD-UAV and public infrared datasets demonstrate that our TDCNet achieves state-of-the-art detection performance in moving target detection.

Dataset and code — https://github.com/IVPLabs/TDCNet

1 Introduction

Moving infrared small target detection (IRSTD) aims to locate small and dim targets in infrared images, often under complex backgrounds and low signal-to-clutter ratio (SCR) conditions. It plays a crucial role in a wide range of applications, including surveillance of unmanned aerial vehicles (UAVs) (Fang et al. 2022, 2023a,b, 2025b; Zhang et al. 2025) and space-based monitoring (Du et al. 2022). In such scenarios, targets are typically tiny and low-contrast in the spatial domain, making them easily overwhelmed by complex dynamic background clutter. These challenges often result in missed detections and false alarms.

To address the above challenges, numerous infrared small target detection methods have been proposed and can

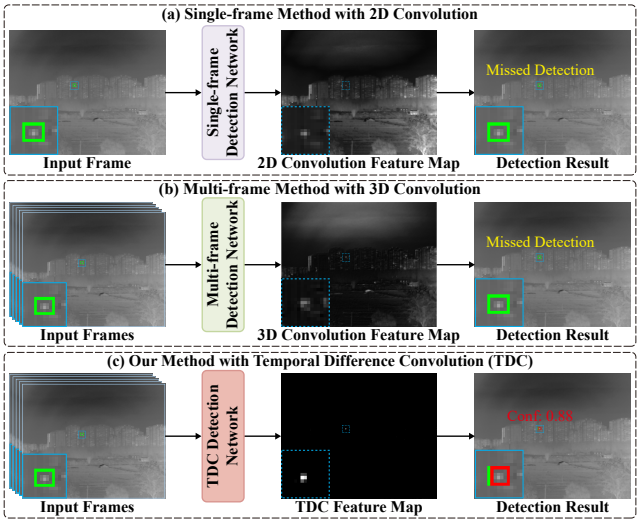

Figure 1: Three representative categories of methods for moving infrared small target detection. (a) Single-frame methods (Liu et al. 2024) employ 2D convolution, which lacks temporal context and often fails to distinguish targets from background clutter. (b) Multi-frame methods (Peng et al. 2025) typically utilize 3D convolution to extract spatio-temporal features, but they often overlook explicitly leveraging motion cues, resulting in limited detection performance. (c) Our method introduces temporal difference convolution (TDC) to explicitly capture motion-contextual information while representing spatio-temporal features, thereby effectively suppressing complex backgrounds and enhancing the detection performance of moving infrared small targets.

*Corresponding author.

be broadly divided into two categories: single-frame approaches (Fang et al. 2023a,b, 2025b; Liu et al. 2024) and multi-frame approaches (Chen et al. 2024; Tong et al. 2024; Zhang et al. 2025; Peng et al. 2025). The former focuses on constructing complex network architectures to extract spatial features but lacks the capability to model temporal motion patterns in complex backgrounds, often leading to missed detections or false alarms (Fang et al. 2022; Yang et al. 2025; Zhang et al. 2024; Liu et al. 2024). Accurate spatio-temporal feature modeling is crucial for moving IRSTD. Accordingly, the latter incorporates multi-frame inputs and leverages temporal difference modeling (Wang et al. 2021; Yan et al. 2023; Xiao et al. 2023) or spatio-temporal (3D) convolutions (Peng et al. 2025; Li et al. 2025b) to extract spatio-temporal target features. The temporal difference operation can explicitly capture temporal contextual information but has a limited capability to extract spatial features. In contrast, 3D convolution can effectively represent features in three dimensions, yet it lacks explicit awareness of motion dynamics along the temporal dimension. This limitation hampers its capability to capture subtle frame-to-frame variations at the pixel level, which is particularly critical in detecting weak small targets under low-SCR scenarios (Huang et al. 2024; Zhang et al. 2025; Peng et al. 2025; Li et al. 2025b).

To overcome the above limitations, we propose a novel moving IRSTD network (TDCNet), which effectively extracts and enhances spatio-temporal features for accurate target detection. In general, infrared small targets have weak features in the spatial domain and are susceptible to complex background interference. However, moving infrared small targets typically exhibit strong motion-contextual dependencies along the temporal dimension. This observation motivates us to exploit such contextual dependencies to suppress complex background interference and more effectively model the spatio-temporal features. In this work, we introduce the temporal difference convolution reparameterization (TDCR) module that integrates three parallel temporal difference convolution (TDC) blocks to model short-, mid-, and long-term motion-contextual dependencies, respectively. Each TDC block fuses temporal difference and 3D convolution into a unified spatio-temporal convolution representation, which is designed to effectively capture motion-contextual dependencies within a specified temporal range. This allows the TDC block to suppress background clutter as it simultaneously enhances the discrimination of spatio-temporal features for infrared small targets. The multi-branch architecture of the TDCR module during training is equivalently converted into a single-branch structure for efficient inference. This design enables the TDCR module to effectively capture multi-scale motion-contextual dependencies while suppressing pseudo-motion clutter in complex backgrounds without incurring additional computational cost during inference, significantly improving detection performance.

Moreover, we propose a TDC-guided spatio-temporal attention mechanism that performs cross-attention between the spatio-temporal features extracted from the TDC-based backbone and a parallel 3D convolution backbone. The TDC-based backbone emphasizes the salient target regions while suppressing interference from complex backgrounds. By leveraging this property, the proposed mechanism effectively captures global semantic dependencies between the two feature streams and refines the spatio-temporal feature representation of the current frame, guiding the model to focus more accurately on critical target regions and thereby enhancing detection performance. Furthermore, we construct a new IRSTD benchmark, termed IRSTD-UAV, which contains 15,106 frames captured across diverse UAV types and complex backgrounds. Extensive experiments on both IRSTD-UAV and public benchmark IRDST (Sun et al. 2023) demonstrate that TDCNet achieves state-of-the-art (SOTA) detection performance, significantly outperforming existing single-frame and multi-frame methods under low SCR and complex backgrounds.

The contributions of this work can be summarized as:

- We introduce a novel moving IRSTD network (TDCNet) that can effectively capture spatio-temporal features while suppressing complex backgrounds for accurate detection.
- We are the first to propose TDC that fuses temporal difference and spatio-temporal convolution into a unified 3D convolution representation, enabling effectively capture motion-contextual dependencies within a specified temporal range.
- We introduce a novel TDC-guided spatio-temporal attention (TDCSTA) mechanism that models semantic relationships between TDC-enhanced features and parallel 3D convolutional features. This mechanism is leveraged to refine the representations of critical target regions in the current frame, thereby enhancing the detection performance in complex backgrounds.

2 Related Work

2.1 Moving Infrared Small Target Detection

Existing moving IRSTD methods mainly differing in how they handle spatial and temporal information. A widely adopted strategy is to apply 2D convolutional networks independently on each frame within a temporal sequence (Yan et al. 2023; Chen et al. 2024). However, the lack of inter-frame interaction constrains their capability to model spatio-temporal continuity. Conversely, temporal difference methods focus exclusively on frame-wise intensity variations to capture motion cues (Du et al. 2022; Yan et al. 2023), but struggle to extract the spatial semantic representations essential for robust detection. To effectively leverage both spatial and temporal information, recent approaches either adopt staged pipelines that first extract spatial features via 2D convolutions and then apply temporal modeling (Zhang et al. 2025; Zhu et al. 2025) or employ 3D convolutions to jointly capture spatio-temporal features (Li et al. 2025a,b). However, these methods often suffer from limited motion-awareness or insufficient spatio-temporal context modeling in complex backgrounds. In contrast, we propose TDC block that fuses temporal difference and spatio-temporal convolution into a unified 3D convolution representation, effectively

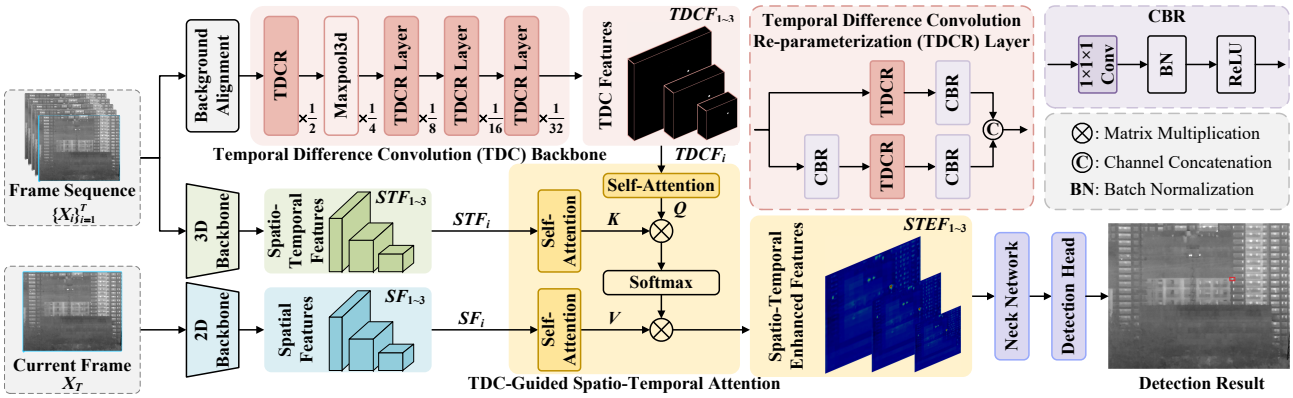

Figure 2: Overview of the proposed TDCNet. The input consists of a frame sequence $\{X_i\}_{i=1}^{T}$ and the current frame X_T. The temporal difference convolution (TDC) backbone utilizes the temporal difference convolution re-parameterization layer to extract TDC features from $\{X_i\}_{i=1}^{T}$. The 2D backbone processes X_T to extract spatial features, while the 3D backbone handles $\{X_i\}_{i=1}^{T}$ to extract spatio-temporal features. The TDC-guided spatio-temporal attention module refines these feature streams to generate spatio-temporal enhanced features, which are aggregated by the neck and detection head to produce the final result.

capturing motion-contextual dependencies for robust moving IRSTD under complex backgrounds.

2.2 Spatio-Temporal Context Modeling

Temporal difference, 3D convolution, and transformer-based models are fundamental techniques for spatio-temporal modeling in video analysis, widely applied to tasks such as action recognition and video understanding (Zhao, Xiong, and Lin 2018; Wang et al. 2021; Bertasius, Wang, and Torresani 2021). Temporal difference captures inter-frame variations to highlight motion cues (Ng and Davis 2018; Xie et al. 2023), while 3D convolution jointly learns spatial and temporal features (Zhou et al. 2018; Li et al. 2020). Transformer-based models further introduce temporal self-attention to enable long-range dependency modeling (Arnab et al. 2021; Selva et al. 2023). However, each method focuses on a limited aspect: temporal difference lacks semantic context, and both 3D convolution and transformer-based models often overlook explicit motion cues. In this work, we propose a unified spatio-temporal network that combines multi-scale motion-contextual modeling via TDCR and spatio-temporal feature enhancement via TDCSTA for robust moving IRSTD.

3 The Proposed Method

3.1 Overall Architecture

In this study, we propose a novel moving IRSTD network, TDCNet, as illustrated in Figure 2. It first introduces a temporal difference convolution (TDC) backbone. Then, a TDC-guided spatio-temporal attention module refines feature representations by applying self-attention to three distinct feature streams and performing cross-attention with TDC features as the query to selectively enhance spatio-temporal features. Finally, we construct a challenging benchmark dataset, IRSTD-UAV, to validate the effectiveness of our method.

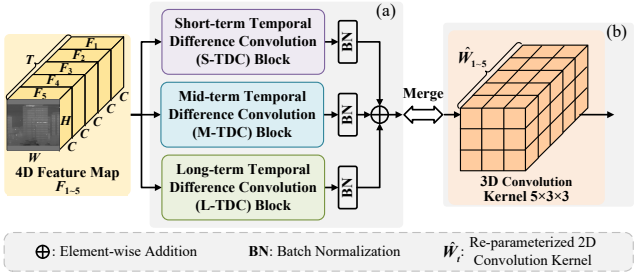

Figure 3: Overview of the proposed temporal difference convolution re-parameterization (TDCR) module, which equivalently transforms three parallel TDC blocks (a) into a single 3D convolution representation (b).

3.2 Temporal Difference Convolution Backbone

Inspired by the 3D backbone design of STMENet (Peng et al. 2025), the TDC backbone is introduced to extract spatio-temporal contextual features. Before the frame sequence is fed into the TDC backbone, a background alignment process is applied to suppress camera motion (Shen et al. 2024). By progressively stacking TDCR layers, the spatio-temporal contextual features from earlier stages are further refined with multi-scale temporal ranges, enabling the model to learn more discriminative representations of small moving targets embedded in complex infrared scenes.

3.3 Temporal Difference Convolution Re-parameterization Module

As illustrated in Figure 3, we propose a novel TDCR module to enhance spatio-temporal contextual feature modeling capability over multiple temporal scales. During training, the TDCR consists of three parallel branches: short-term TDC (S-TDC) block, mid-term TDC (M-TDC) block, and long-term TDC (L-TDC) block (Figure 4). Each branch is specifically designed to capture temporal dependencies at different

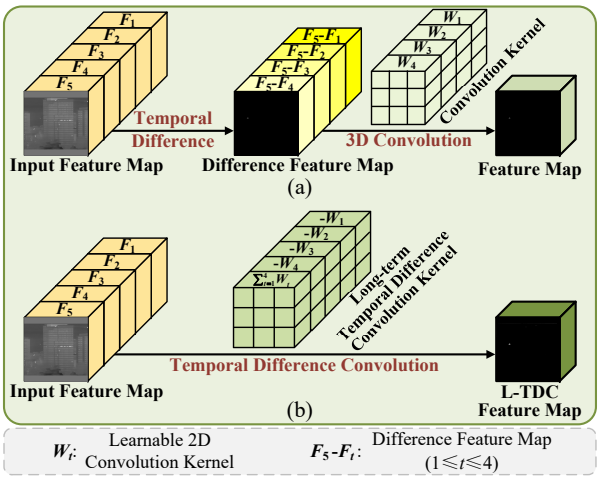

Figure 4: (a) Sequential combination of existing temporal difference and 3D convolution, and (b) Our proposed long-term temporal difference convolution (L-TDC) block, which fuses temporal difference modeling and 3D convolution into a unified spatio-temporal convolution representation.

temporal scales. The outputs of these blocks are independently normalized by batch normalization layers and then aggregated through summation. During inference, we re-parameterize the three branches into a unified single 3D convolution to simplify the inference pipeline while preserving the multi-scale temporal modeling capability.

Temporal Difference Convolution. Accurate spatio-temporal feature modeling is essential for robust moving IRSTD in infrared sequences. Traditional approaches typically rely on either temporal difference operations or 3D convolutions. The temporal difference directly models the motion information by computing frame-wise differences, providing strong awareness of motion dynamics but suffering from weak spatial feature representation (Du et al. 2022). In contrast, 3D convolution effectively extracts spatio-temporal features but lacks explicit motion awareness in cluttered backgrounds (Peng et al. 2025). To leverage the strengths of both methods, we propose the TDC block, which fuses temporal difference and 3D convolution into a unified spatio-temporal convolution representation. Specifically, to explicitly capture the motion-contextual dependencies between frames, we reformulate the traditional 3D convolutional weights $W \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times T \times H \times W}$, where C_{in} and C_{out} denote the number of input and output channels, respectively, and T, H, and W represent the kernel size along the temporal, height, and width dimensions. As presented in Figure 4, we take the L-TDC block as an example. The input feature map $F \in \mathbb{R}^{T \times C \times H \times W}$ to the L-TDC block consists of a sequence of frames $\{F_t\}_{t=1}^{5}$, where each $F_t \in \mathbb{R}^{C \times H \times W}$. Here, F_5 denotes the current frame, while F_1 to F_4 are preceding frames. The L-TDC block aims to capture long-term motion-contextual dependencies by computing differences between the current frame and all previous frames. To achieve this, W is decomposed along the temporal dimension into a set of 2D convolution kernels $\{W_t\}_{t=1}^{4}$, where each $W_t \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times H \times W}$ models the frame-wise difference at time step t. Its output is defined as:

$$\begin{aligned} F_l &= (-W_1) * F_1 + (-W_2) * F_2 + (-W_3) * F_3 \\ &\quad + (-W_4) * F_4 + \left(\textstyle\sum_{t=1}^{4} W_t\right) * F_5 \\ &= \textstyle\sum_{t=1}^{4} \left(W_t * (F_5 - F_t)\right). \end{aligned} \tag{1}$$

Here, $*$ denotes the convolution operation. F_l is mathematically equivalent to the summation of convolutions between W_t and the temporal difference feature map $F_5 - F_t$. However, as shown in Figure 4, it is important to emphasize that our TDC does not explicitly perform a difference operation. Instead, it implicitly fuses temporal difference and 3D convolution into a unified spatio-temporal convolution representation. This formulation explicitly encodes the long-term temporal difference and rich spatio-temporal features between the current frame and all previous frames, capturing long-term motion-contextual dependencies.

S-TDC and M-TDC are derived similarly to L-TDC, each targeting motion modeling at different temporal scales. The S-TDC block focuses on short-term motion by computing differences between consecutive frames:

$$\begin{aligned} F_s &= (-W_2) * F_1 + (W_2 - W_3) * F_2 + (W_3 - W_4) * F_3 \\ &\quad + (W_4 - W_5) * F_4 + W_5 * F_5 \\ &= \textstyle\sum_{t=2}^{5} \left(W_t * (F_t - F_{t-1})\right). \end{aligned} \tag{2}$$

This short-term motion modeling design enhances the network's sensitivity to fine-grained and fast-changing motion patterns, enabling effective capture of subtle variations between consecutive frames. Meanwhile, the M-TDC block captures intermediate motion context by computing differences between frames with two-frame intervals, complementing short- and long-term modeling with its distinct temporal scope:

$$\begin{aligned} F_m &= (-W_3) * F_1 + (-W_4) * F_2 + (W_3 - W_5) * F_3 \\ &\quad + W_4 * F_4 + W_5 * F_5 \\ &= \textstyle\sum_{t=3}^{5} \left(W_t * (F_t - F_{t-2})\right). \end{aligned} \tag{3}$$

This design enables the network to capture mid-term motion contextual dependencies while mitigating redundant motion or noise. Together, the three TDC blocks capture complementary spatio-temporal features at different temporal scales, thereby strengthening the overall motion context modeling capability.

As a result, the TDCR module captures multi-scale motion-contextual dependencies through three parallel TDC branches: $\tilde{F}_s = \text{BN}_s(F_s)$, $\tilde{F}_m = \text{BN}_m(F_m)$, $\tilde{F}_l = \text{BN}_l(F_l)$, where $\text{BN}_{\{s,m,l\}}$ are their corresponding batch normalization layers. The three outputs are then aggregated as the final output of the TDCR module: $F_{TDCR} = \tilde{F}_s + \tilde{F}_m + \tilde{F}_l$.

Re-parameterization of the Multi-scale TDC Branches. We first fuse convolution and BN operations within each TDC branch via parameter transformation (Kobayashi and

Ye 2024): $\hat{W}_i = \gamma_i \cdot W_i/\sigma_i$, $\hat{b}_i = \gamma_i \cdot (b_i - \mu_i)/\sigma_i + \beta_i$, where W_i and b_i denote the convolutional kernel weights and biases, and $\gamma_i, \beta_i, \mu_i, \sigma_i$ are the parameters of batch normalization. Leveraging the homogeneity and additivity of convolution, we then merge the three TDC branches into a single 3D convolution: $\hat{W}_{\text{TDCR}} = \sum_{i\in\{s,m,l\}} \hat{W}_i$, $\hat{b}_{\text{TDCR}} = \sum_{i\in\{s,m,l\}} \hat{b}_i$. The resulting re-parameterized TDCR module can be expressed as: $\text{TDCR}(F) = \hat{W}_{\text{TDCR}} * F + \hat{b}_{\text{TDCR}}$. This re-parameterization improves intra-model inference efficiency while preserving the benefits of multi-scale motion context modeling.

3.4 TDC-Guided Spatio-Temporal Attention Module

As illustrated in the bottom center of Figure 2, we propose a TDC-guided spatio-temporal attention (TDCSTA) module to refine the feature representation of small moving targets in cluttered infrared scenes. Unlike conventional methods that directly fuse multi-frame features, TDCSTA introduces a tri-branch architecture to decouple and specialize in distinct spatio-temporal cues, thereby enabling more structured and effective feature interaction. Specifically, TDCSTA operates on three feature streams extracted from the final three stages of their respective backbones: the temporal difference convolution features ($TDCF_{1\sim3}$) from the TDC backbone, the spatio-temporal features ($STF_{1\sim3}$) from the 3D backbone and the spatial features ($SF_{1\sim3}$) from the 2D backbone. By capturing global semantic dependencies and enabling selective feature interaction, TDCSTA generates spatio-temporal enhanced features, helping the model focus more accurately on small targets and improving detection performance. The enhanced features are then passed to the neck and detection head to produce the final detection result.

Self-Attention for Semantic Expressiveness Enhancement. To enhance the semantic representation capability of each feature stream and effectively suppress irrelevant background clutter, we apply a self-attention mechanism independently to the three feature streams: $TDCF_i$, STF_i, and SF_i, at each stage i. We partition each feature stream $FS \in \mathbb{R}^{T\times C\times H\times W}$ into non-overlapping 3D local windows of size $P \times M \times M$, and compute self-attention with both regular and shifted window partitioning applied (Liu et al. 2022). Formally, the self-attention is defined as:

$$\text{SA}(FS) = \text{Softmax}(QK^\top/\sqrt{d} + B)V, \tag{4}$$

where $Q, K, V \in \mathbb{R}^{PM^2\times d}$ are linear projections of the input tokens within each window, d is the embedding dimension and B is the relative positional bias. We apply this mechanism to each feature stream as follows:$\hat{F}_{TDCF,i} = \text{SA}(TDCF_i)$, $\hat{F}_{STF,i} = \text{SA}(STF_i)$, $\hat{F}_{SF,i} = \text{SA}(SF_i)$.

Cross-Attention for TDC-Guided Semantic Dependency Modeling. To explicitly model semantic dependencies guided by motion-aware features, we employ a cross-attention mechanism where $\hat{F}_{TDCF,i}$ serves as the query, while $\hat{F}_{STF,i}$ and $\hat{F}_{SF,i}$ serve as key and value, respectively. The output of the cross-attention mechanism is the spatio-temporal enhanced features (STEF), defined as:

$$STEF_i = \text{Softmax}(Q_i K_i^\top/\sqrt{d} + B_i)V_i, \tag{5}$$

where Q_i is derived from $\hat{F}_{TDCF,i}$, K_i from $\hat{F}_{STF,i}$, and V_i from $\hat{F}_{SF,i}$. By leveraging the discriminative motion-contextual cues encoded in $\hat{F}_{TDCF,i}$, which highlight salient target regions while suppressing complex background interference, this mechanism enables the model to focus on semantically relevant regions across temporal and spatial dimensions, thereby enhancing semantic dependency modeling and refining the spatio-temporal representation of the current frame.

4 Experimental Results and Analysis

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate our method on two real infrared benchmarks: a self-constructed IRSTD-UAV dataset and a public IRDST dataset (Sun et al. 2023). The IRSTD-UAV dataset contains 17 real infrared video sequences with 15,106 frames, featuring small targets and complex backgrounds such as buildings, trees, and clouds. More details of our dataset are provided in the supplementary material (Fang et al. 2025a).

Evaluation Metrics. For evaluation, we adopt standard metrics, including precision (P), recall (R), F_1-score (F_1), and average precision (AP_{50}), all computed at an intersection-over-union (IoU) threshold of 0.5. Real-time performance is measured in frames per second (FPS), while computational complexity is assessed using the number of parameters (Params) and floating point of operations (FLOPs).

4.2 Implementation Details

All experiments are conducted on a single NVIDIA RTX 3090 GPU with CUDA 12.4 and PyTorch 2.7. The training is performed using the Adam optimizer with a learning rate of 0.001 and a weight decay of 1×10^{-4}. We first pre-train the 2D backbone on still images and the 3D backbone on multi-frame inputs. Both are then frozen, and the TDC backbone along with the TDCSTA module is trained on video sequences. The input frames are resized to 640×640, and five consecutive frames were used as input during training and inference. We adopt the IoU loss for regression and binary cross-entropy loss for objectness and classification: $\mathcal{L} = \mathcal{L}_{reg} + \mathcal{L}_{obj} + \mathcal{L}_{cls}$. For single-frame methods, we include YOLO11-L (Jocher, Qiu, and Chaurasia 2024), YOLOv12-L (Tian, Ye, and Doermann 2025), and Hyper-YOLO-M (Feng et al. 2025) as general-purpose convolutional neural network (CNN)-based detectors, along with MSHNet (Liu et al. 2024) and PConv (YOLOv8) (Yang et al. 2025), which are specifically designed for IRSTD. Additionally, we include SCTransNet (Yuan et al. 2024) as a infrared-specific transformer-based baseline. For multi-frame methods, we select infrared-specific CNN-based methods including TMP (Zhu et al. 2024), SSTNet (Chen et al. 2024), MOCID (Zhang et al. 2025), STMENet (Peng et al. 2025), RFR (Ying et al. 2025), and DTUM (Li et al. 2025b).

Type	Method	Pub'Year	IRSTD-UAV				IRDST				Params (M)	FLOPs (G)	FPS
			P	R	F_1	AP_{50}	P	R	F_1	AP_{50}			
Single-frame	YOLO11-L	2024	96.26	95.73	95.99	91.20	96.70	96.00	96.35	92.10	26.0	111.8	61.7
	MSHNet	CVPR'2024	86.92	84.94	85.92	87.62	82.31	77.64	79.91	63.21	4.1	76.3	63.3
	SCTransNet	TGRS'2024	96.71	89.19	92.80	85.36	96.80	90.18	93.37	92.35	11.2	80.9	17.8
	YOLOv12-L	2025	96.50	94.99	95.74	90.54	97.29	95.63	96.45	92.09	27.1	104.8	30.5
	PConv (YOLOv8)	AAAI'2025	94.51	94.00	94.25	88.47	96.21	95.91	96.06	91.88	23.8	88.6	58.8
	Hyper-YOLO-M	TPAMI'2025	96.21	95.08	95.64	91.04	96.68	96.90	96.79	92.53	32.4	119.0	**87.3**
Multi-frame	TMP	ESA'2024	37.90	61.70	46.96	23.00	86.65	81.36	83.92	70.01	16.4	92.9	25.0
	SSTNet	TGRS'2024	88.98	84.19	86.52	74.60	88.56	81.92	85.11	71.55	11.9	123.6	22.6
	MOCID	AAAI'2025	97.28	94.85	96.05	91.32	**98.92**	96.86	97.88	94.74	13.1	98.7	11.5
	STMENet	EAAI'2025	86.70	88.04	87.36	75.97	87.78	84.22	85.96	73.40	10.4	45.8	48.5
	ResUNet+RFR	TGRS'2025	56.24	82.47	66.87	46.20	83.92	76.10	79.82	61.45	0.9	73.8	72.6
	ResUNet+DTUM	TNNLS'2025	93.83	83.17	88.18	81.37	82.87	87.79	85.26	71.48	**0.3**	**41.4**	13.9
	Ours	-	**97.99**	**96.27**	**97.12**	**93.83**	98.12	**97.71**	**97.91**	**94.79**	24.8	95.7	18.5

Table 1: Quantitative comparison of the proposed method with SOTA methods on the IRSTD-UAV and IRDST datasets. **Bold** and underline indicate the best and the second best results, respectively.

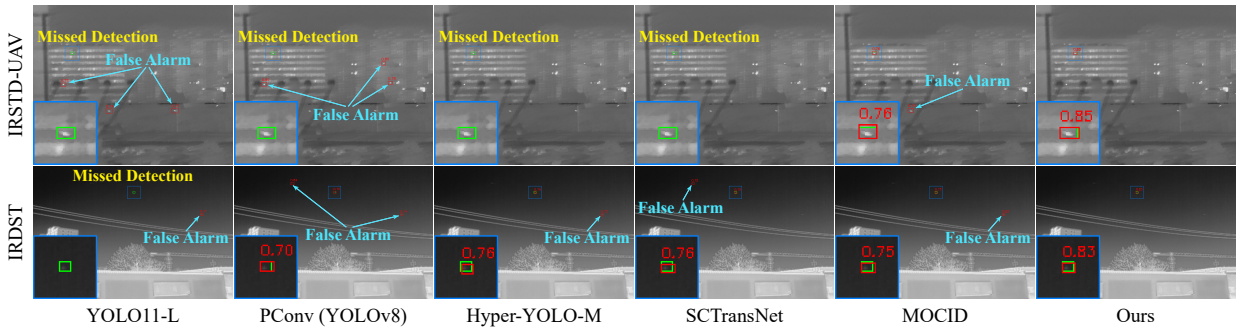

Figure 5: Visual comparison of results from SOTA methods and TDCNet on the IRSTD-UAV and IRDST dataset. Boxes in green and red represent ground-truth and detected targets, respectively.

4.3 Quantitative Results

As shown in Table 1, our proposed TDCNet achieves SOTA performance on P, R, F_1, and AP_{50} on the IRSTD-UAV dataset, and on R, F_1, and AP_{50} on IRDST. TDCNet outperforms all single-frame methods such as MSHNet and Hyper-YOLO-M, which suffer from limited robustness in cluttered infrared scenes due to the absence of temporal modeling. Among multi-frame methods, TDCNet achieves the highest R, F_1, and AP_{50}. Other methods like MOCID and SCTransNet are less effective in complex scenes due to insufficient motion modeling and suboptimal spatio-temporal representation. TDCNet achieves a lower computational cost of 95.7G FLOPs and a reasonable inference speed of 18.5 FPS.

4.4 Qualitative Results

As demonstrated in Figure 5, our TDCNet exhibits superior detection performance across two challenging infrared scenarios from the IRSTD-UAV and IRDST datasets. Even in the presence of strong background clutter, such as urban structures and light-like distractors, TDCNet effectively highlights true UAV targets while suppressing false alarms. This is because the TDCR module can effectively capture multi-scale motion-contextual dependencies, while the TDCSTA selectively enhances target-relevant features while suppressing irrelevant background clutter. YOLO11-L, Hyper-YOLO-M, and PConv (YOLOv8) all struggle under complex infrared scenes, frequently missing true targets and generating false alarms due to their lack of temporal modeling and motion-aware feature representation. SCTransNet suffers from false alarms due to the absence of explicit motion context guidance. Although MOCID incorporates motion context, it fails to capture multi-scale temporal dependencies, which limits its capability to suppress clutter in complex backgrounds. More visual results are provided in the supplementary material (Fang et al. 2025a).

4.5 Ablation Study

In this section, we report ablation studies. More experiments are provided in the supplementary material (Fang et al. 2025a).

Impact of the proposed TDCR and TDCSTA. As shown in Table 2, both TDCR and TDCSTA independently contribute to performance improvements over the baseline. Specifically, TDCR improves P to 97.61 and AP_{50} to 92.50, while TDCSTA improves R to 95.96 and F_1 to 96.74. In combination, TDCR and TDCSTA achieve superior results, highlighting their complementary benefits. To better understand their effects, we visualize the heatmaps in Figure 6. Compared to the base model, TDCR yields more focused and distinguishable activations on targets. After applying

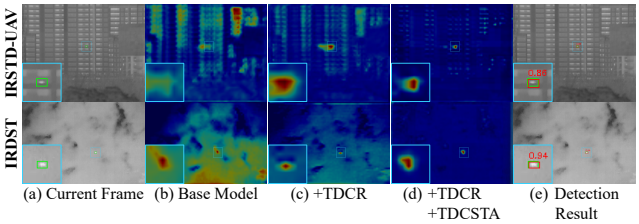

Figure 6: Heatmap visualization illustrating the progressive enhancement in background suppression achieved by TDCR and TDCSTA.

TDCR	TDCSTA	P	R	F_1	AP_{50}
×	×	86.70	88.04	87.36	75.97
√	×	97.61	95.93	96.76	92.50
×	√	97.53	95.96	96.74	92.35
√	√	**97.99**	**96.27**	**97.12**	**93.83**

Table 2: Ablation study of the TDCR and TDCSTA modules.

Method	P	R	F_1	AP_{50}	FLOPs (G)
TD	94.36	90.24	92.25	89.73	**41.351**
3D Conv	86.70	88.04	87.36	75.97	45.842
TD + 3D Conv	95.15	92.62	93.87	89.81	45.854
TDC	**97.61**	**95.93**	**96.76**	**92.50**	45.749

Table 3: Ablation study of TD, 3D Conv, TD+3D Conv, and TDC.

TDCSTA, irrelevant background activations are noticeably suppressed, further enhancing target saliency in cluttered infrared scenes.

Impact of TDC. As presented in Table 3, temporal difference (TD) alone suffers from limited recall (R) and AP_{50}, as it leverages only frame-wise intensity variations while discarding most spatial contextual information. In contrast, 3D convolution produces lower F_1 and AP_{50} due to its limited capability to model explicit temporal dependencies. Simply combining TD and 3D convolution yields some performance gains due to its reliance on single-scale spatio-temporal context. In contrast, our method achieves greater performance improvements, increasing AP_{50} from 89.81 to 92.50, without introducing additional computational cost. This is because our proposed TDC fuses temporal difference and 3D convolution into a unified and learnable representation that captures multi-scale spatio-temporal context dependencies across different temporal ranges.

Impact of Different Spatio-Temporal Contextual Features. Table 4 shows that incorporating spatio-temporal contextual features at different temporal scales leads to notable performance gains. The S-TDC block improves P to 96.19 and F_1 to 94.91 by capturing fine-grained short-term spatio-temporal contextual features. M-TDC improves R to 95.79 and F_1 to 95.65. L-TDC captures long-range dependencies, achieving a P of 97.49 and an AP_{50} of 92.35. When all branches are combined, the model reaches the best results across all metrics, confirming that multi-scale temporal modeling provides complementary motion cues essential for robust small target detection.

Method	P	R	F_1	AP_{50}
w/o TDC	86.70	88.04	87.36	75.97
w/ S-TDC	96.19	93.67	94.91	90.31
w/ M-TDC	95.51	95.79	95.65	90.46
w/ L-TDC	97.49	95.11	96.29	92.35
w/ All	**97.61**	**95.93**	**96.76**	**92.50**

Table 4: Ablation study of different TDC blocks.

Re-param	F_1	AP_{50}	Params (M)	FLOPs (G)
w/o	96.76	92.50	24.85	102.96
w/	96.76	92.50	**24.76**	**95.67**

Table 5: Ablation study of re-parameterization in TDCR.

Query	Key	Value	P	R	F_1	AP_{50}
TDCF	STF	SF	**97.53**	**95.96**	**96.74**	**92.35**
STF	TDCF	SF	93.11	90.15	91.61	87.22
SF	STF	TDCF	92.84	89.73	91.26	86.39

Table 6: Ablation study of the TDCSTA module with different combinations of query, key, and value in the cross-attention mechanism.

Impact of Re-parameterization in TDCR. According to Table 5, re-parameterization reduces Params from 24.85M to 24.76M and FLOPs from 102.96G to 95.67G while maintaining consistent detection performance, demonstrating improved efficiency without sacrificing accuracy.

Impact of TDCSTA. From Table 6, we observe that setting temporal difference convolution features (TDCF) as query, with spatio-temporal features (STF) and spatial features (SF) as key and value, yields the best performance with F_1 of 96.74 and AP_{50} of 92.35. Replacing the query with STF or SF leads to noticeable drops in all metrics, confirming that TDCF provides more discriminative guidance in TDCSTA, which is essential for accurate target localization in cluttered scenes.

5 Conclusion

This paper introduces a novel model TDCNet for moving IRSTD. TDCNet incorporates two key designs: the TDCR module and the TDCSTA mechanism. TDCR module captures multi-scale temporal contextual features while suppressing complex backgrounds without incurring additional computational cost during inference. TDCSTA mechanism models semantic relationships between two 3D feature streams to refine the representation of critical target regions in the current frame. These components effectively enhance spatio-temporal feature representation, enabling TDCNet to outperform existing methods on the IRSTD-UAV and public IRDST datasets.

Acknowledgments

This work was supported by the Open Research Fund of the National Key Laboratory of Multispectral Information Intelligent Processing Technology under Grant 61421132301 and was supported in part by the projects of the National Natural Science Foundation of China under Grants No. 62371203 and 62301228.

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. ViViT: A Video Vision Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6816–6826.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, 813–824.

Chen, S.; Ji, L.; Zhu, J.; Ye, M.; and Yao, X. 2024. SSTNet: Sliced Spatio-Temporal Network With Cross-Slice ConvLSTM for Moving Infrared Dim-Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–12.

Du, J.; Lu, H.; Zhang, L.; Hu, M.; Chen, S.; Deng, Y.; Shen, X.; and Zhang, Y. 2022. A Spatial-Temporal Feature-Based Detection Framework for Infrared Dim Small Target. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–12.

Fang, H.; Ding, L.; Wang, L.; Chang, Y.; Yan, L.; and Han, J. 2022. Infrared Small UAV Target Detection Based on Depthwise Separable Residual Dense Network and Multiscale Feature Fusion. *IEEE Transactions on Instrumentation and Measurement*, 71: 1–20.

Fang, H.; Guo, S.; Chen, Q.; Chang, Y.; and Yan, L. 2025a. Spatio-Temporal Context Learning with Temporal Difference Convolution for Moving Infrared Small Target Detection. https://arxiv.org/abs/2511.09352.

Fang, H.; Liao, Z.; Wang, L.; Li, Q.; Chang, Y.; Yan, L.; and Wang, X. 2023a. DANet: Multi-scale UAV Target Detection with Dynamic Feature Perception and Scale-aware Knowledge Distillation. In *Proceedings of the 31st ACM International Conference on Multimedia (ACMMM)*, 2121–2130.

Fang, H.; Liao, Z.; Wang, X.; Chang, Y.; and Yan, L. 2023b. Differentiated attention guided network over hierarchical and aggregated features for intelligent UAV surveillance. *IEEE Transactions on Industrial Informatics*, 19(9): 9909–9920.

Fang, H.; Wang, X.; Li, Z.; Wang, L.; Li, Q.; Chang, Y.; and Yan, L. 2025b. Detection-Friendly Nonuniformity Correction: A Union Framework for Infrared UAV Target Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11898–11907.

Feng, Y.; Huang, J.; Du, S.; Ying, S.; Yong, J.-H.; Li, Y.; Ding, G.; Ji, R.; and Gao, Y. 2025. Hyper-YOLO: When Visual Object Detection Meets Hypergraph Computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 2388–2401.

Huang, Y.; Zhi, X.; Hu, J.; Yu, L.; Han, Q.; Chen, W.; and Zhang, W. 2024. LMAFormer: Local Motion Aware Transformer for Small Moving Infrared Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–17.

Jocher, G.; Qiu, J.; and Chaurasia, A. 2024. Ultralytics YOLO. https://github.com/ultralytics/ultralytics.

Kobayashi, T.; and Ye, J. 2024. Spatio-temporal Filter Analysis Improves 3D-CNN For Action Classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6972–6981.

Li, F.; Rao, P.; Sun, W.; Su, Y.; and Chen, X. 2025a. A Low Signal-to-Noise Ratio Infrared Small-Target Detection Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18: 8643–8658.

Li, R.; An, W.; Xiao, C.; Li, B.; Wang, Y.; Li, M.; and Guo, Y. 2025b. Direction-Coded Temporal U-Shape Module for Multiframe Infrared Small Target Detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1): 555–568.

Li, X.; Lin, T.; Liu, X.; Zuo, W.; Li, C.; Long, X.; He, D.; Li, F.; Wen, S.; and Gan, C. 2020. Deep Concept-wise Temporal Convolutional Networks for Action Localization. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, 4004–4012.

Liu, Q.; Liu, R.; Zheng, B.; Wang, H.; and FU, Y. 2024. Infrared Small Target Detection with Scale and Location Sensitivity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17490–17499.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video Swin Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3192–3201.

Ng, J. Y.-H.; and Davis, L. S. 2018. Temporal Difference Networks for Video Action Recognition. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1587–1596.

Peng, S.; Ji, L.; Chen, S.; Duan, W.; and Zhu, S. 2025. Moving infrared dim and small target detection by mixed spatio-temporal encoding. *Engineering Applications of Artificial Intelligence*, 144: 110100.

Selva, J.; Johansen, A. S.; Escalera, S.; Nasrollahi, K.; Moeslund, T. B.; and Clapés, A. 2023. Video Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 12922–12943.

Shen, X.; Cai, Z.; Yin, W.; Müller, M.; Li, Z.; Wang, K.; Chen, X.; and Wang, C. 2024. GIM: Learning Generalizable Image Matcher From Internet Videos. In *Proceedings of the International Conference on Learning Representation (ICLR)*.

Sun, H.; Bai, J.; Yang, F.; and Bai, X. 2023. Receptive-Field and Direction Induced Attention Network for Infrared Dim Small Target Detection With a Large-Scale Dataset IRDST. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Tian, Y.; Ye, Q.; and Doermann, D. 2025. YOLOv12: Attention-Centric Real-Time Object Detectors. arXiv:2502.12524.

Tong, X.; Zuo, Z.; Su, S.; Wei, J.; Sun, X.; Wu, P.; and Zhao, Z. 2024. ST-Trans: Spatial-Temporal Transformer for Infrared Small Target Detection in Sequential Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–19.

Wang, L.; Tong, Z.; Ji, B.; and Wu, G. 2021. TDN: Temporal Difference Networks for Efficient Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1895–1904.

Xiao, J.; Wu, Y.; Chen, Y.; Wang, S.; Wang, Z.; and Ma, J. 2023. LSTFE-Net: Long Short-Term Feature Enhancement Network for Video Small Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14613–14622.

Xie, Z.; Chen, J.; Wu, K.; Guo, D.; and Hong, R. 2023. Global Temporal Difference Network for Action Recognition. *IEEE Transactions on Multimedia*, 25: 7594–7606.

Yan, P.; Hou, R.; Duan, X.; Yue, C.; Wang, X.; and Cao, X. 2023. STDMANet: Spatio-Temporal Differential Multiscale Attention Network for Small Moving Infrared Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

Yang, J.; Liu, S.; Wu, J.; Su, X.; Hai, N.; and Huang, X. 2025. Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9202–9210.

Ying, X.; Liu, L.; Lin, Z.; Shi, Y.; Wang, Y.; Li, R.; Cao, X.; Li, B.; Zhou, S.; and An, W. 2025. Infrared Small Target Detection in Satellite Videos: A New Dataset and a Novel Recurrent Feature Refinement Framework. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–18.

Yuan, S.; Qin, H.; Yan, X.; Akhtar, N.; and Mian, A. 2024. SCTransNet: Spatial-Channel Cross Transformer Network for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–15.

Zhang, M.; Ouyang, Y.; Gao, F.; Guo, J.; Zhang, Q.; and Zhang, J. 2025. MOCID: Motion Context and Displacement Information Learning for Moving Infrared Small Target Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(10): 10022–10030.

Zhang, M.; Yang, H.; Guo, J.; Li, Y.; Gao, X.; and Zhang, J. 2024. IRPruneDet: efficient infrared small target detection via wavelet structure-regularized soft channel pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7224–7232.

Zhao, Y.; Xiong, Y.; and Lin, D. 2018. Trajectory Convolution for Action Recognition. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2208–2219.

Zhou, Y.; Sun, X.; Zha, Z.-J.; and Zeng, W. 2018. MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 449–458.

Zhu, S.; Ji, L.; Chen, S.; and Duan, W. 2025. Spatial–temporal-channel collaborative feature learning with transformers for infrared small target detection. *Image and Vision Computing*, 154: 105435.

Zhu, S.; Ji, L.; Zhu, J.; Chen, S.; and Duan, W. 2024. TMP: Temporal Motion Perception with spatial auxiliary enhancement for moving Infrared dim-small target detection. *Expert Systems with Applications*, 255: 124731.