Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Infrared Small Target Detection via Candidate-Gated Background Modeling and Fixed-Pixel Spatiotemporal Saliency

**Xin Li**

kdlixin@163.com

Nanjing University of Science and Technology ZiJin College

**Research Article**

**Keywords:** infrared image sequences, dim small-target detection, spatiotemporal saliency, adaptive background modeling, multi-frame detection, image processing

**Posted Date:** November 18th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11018983/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Infrared Small Target Detection via Candidate-Gated Background Modeling and Fixed-Pixel Spatiotemporal Saliency

Xin Li[1]

[1] School of Electronic Engineering and Optoelectronic Technology, Zijin College, Nanjing University of Science and Technology, Nanjing 210023, China

*Corresponding author(s). E-mail(s): kdlixin@163.com;

**Abstract.** Detecting dim small targets in infrared image sequences remains challenging because point-like targets are easily confused with structured clutter, while platform motion degrades inter-frame correspondence. This paper presents a fixed-pixel spatiotemporal saliency framework with candidate-gated adaptive background modeling. Robust quantile normalization and quality-controlled subpixel registration are first employed to reduce radiometric variation and reject unreliable frame alignment. A multi-scale spatial point response and a locally normalized positive residual are combined to generate a soft target likelihood, which adaptively controls the background learning rate and reduces the risk of weak-target absorption. At temporally aligned pixel locations, temporal residual energy, stability, and run-length continuity are accumulated over accepted frames. Spatial and temporal evidence is then integrated through weighted geometric fusion, followed by robust local constant false-alarm rate detection, multi-frame confirmation, and non-maximum suppression. Experiments on six infrared

sequences, including three self-collected UAV sequences and three public ATR sequences, show that the proposed method achieves the highest signal-to-clutter ratio gain (SCRG) on four sequences and ranks among the top three methods on all six. Its mean SCRG is 7.4603, which is 38.34% higher than that of the strongest baseline. ROC comparisons further demonstrate favorable performance in the low-false-alarm region. These results indicate that jointly exploiting spatial target compactness and temporally consistent residual responses improves dim small-target detection in challenging infrared image sequences.

**Keywords:** infrared image sequences; dim small-target detection; spatiotemporal saliency; adaptive background modeling; multi-frame detection; image processing

# 1 Introduction

Infrared small target detection is a fundamental technique in infrared search and tracking systems and has been widely applied in early-warning surveillance, maritime monitoring, remote sensing, and precision guidance. Unlike ordinary object-detection tasks, infrared small targets generally occupy only a few pixels and exhibit weak in-tensity, limited shape information, and little discernible texture. Their visual characteristics can also change considerably with imaging distance, atmospheric transmission, sensor noise, and background conditions. Meanwhile, clouds, sea waves, buildings, vegetation, and thermal noise may produce target-like structures with comparable brightness or local contrast. Consequently, accurately distinguishing genuine tar-gets from complex background interference remains a challenging problem [1].

Existing infrared small target detection methods can generally be divided into single-frame and multi-frame approaches. Single-frame methods detect targets from spatial information within an individual infrared image. Early approaches mainly re-lied on morphological filtering, local contrast enhancement, and handcrafted back-ground-suppression models. These methods are usually computationally efficient and interpretable, but their effectiveness is closely related to the validity of their assumptions about target size and background smoothness. Low-rank and sparse decomposition methods subsequently formulated the background and targets as low-rank and sparse components, respectively. Although these methods improve robustness under certain structured backgrounds, they may suffer from high computational complexity and degraded performance when strong edges or repetitive structures violate the low-rank assumption.

With the development of deep learning, convolutional neural networks and encoder–decoder architectures have become important solutions for infrared small target detection. By learning multilevel representations directly from training data, deep models can capture more discriminative target features than handcrafted operators. Attention mechanisms, feature-pyramid structures, dense skip connections, and boundary-aware supervision have further improved the preservation of small targets in deep networks. Nevertheless, repeated downsampling may still remove weak target responses, while limited training data and differences between imaging conditions can reduce the generalization capability of learned models.

Compared with single-frame approaches, multi-frame methods exploit temporal continuity and motion consistency in infrared sequences. A genuine target generally maintains a relatively coherent trajectory across adjacent frames, whereas noise and many

background disturbances exhibit weaker temporal consistency. Therefore, temporal information can provide valuable evidence for suppressing spatially target-like false alarms. However, this advantage is weakened when the imaging platform or background moves. In such cases, global background displacement, local deformation, parallax, and occlusion may disrupt pixel-level correspondence between frames. Effective multi-frame detection consequently requires both reliable spatial representation and adaptive temporal association, while avoiding the accumulation of registration errors and background interference.

The main contributions of this study are summarized as follows.

1. Candidate-gated adaptive background modeling. A soft target likelihood is jointly constructed from the multi-scale spatial response and the locally normalized background residual. The likelihood adaptively modulates the pixel-wise learning rate, enabling the background model to follow gradual scene variations while preventing persistent weak targets from being progressively absorbed.

2. Fixed-pixel temporal-profile consistency modeling. After quality-controlled subpixel registration, residual energy, temporal stability, and run-length continuity are jointly modeled within a short accepted-frame window. This representation distinguishes persistent target responses from impulsive and flickering noise without re-quiring inter-pixel trajectory search.

3. Joint spatiotemporal saliency generation and robust target confirmation. The point-like spatial response and fixed-pixel temporal profile are integrated through a weighted geometric fusion, with run-length continuity providing additional modulation. The resulting saliency map is further processed by robust local CFAR, confirmation, and non-

maximum suppression, forming a complete detection frame-work from target enhancement to coordinate output.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the proposed method. Section 4 presents the experimental settings and results. Section 5 discusses the main findings and limitations. Section 6 concludes the paper.

# 2 Related Work

## 2.1 Traditional Single-Frame Infrared Small Target Detection

Traditional single-frame methods mainly distinguish targets from backgrounds according to local intensity differences, structural priors, or statistical characteristics. Morphological filtering is one of the earliest strategies for suppressing slowly varying backgrounds. Bai and Zhou improved the conventional top-hat transform by rede-signing morphological structuring elements, thereby enhancing small targets while reducing background residuals [2]. However, morphological methods are sensitive to the predefined shape and scale of the structuring elements, making it difficult to maintain stable performance when target sizes vary.

Local contrast methods estimate target saliency by comparing a central region with its surrounding neighborhood. Chen et al. proposed the local contrast measure, which enhances isolated bright targets through local intensity differences [3]. Wei et al. subsequently introduced a multiscale patch-based contrast measure to improve the detection of targets with different spatial sizes [4]. Deng et al. developed a weighted local difference measure that assigns different contributions to neighboring regions

according to their spatial relationships [5]. Han et al. further employed a tri-layer window to jointly characterize the target region, local background, and surrounding contextual area [6]. These methods are computationally convenient and do not require training samples. Nevertheless, strong edges, cloud boundaries, sea clutter, and other high-contrast structures can generate responses similar to those of genuine targets, resulting in false alarms.

Low-rank and sparse decomposition methods model an infrared image or patch tensor as the combination of a low-rank background component and a sparse target component. Gao et al. proposed the infrared patch-image model, which transforms nonlocal image patches into a data matrix and separates the target from the low-rank background [7]. Dai and Wu extended this idea to a reweighted patch-tensor model incorporating nonlocal and local priors [8]. Zhang and Peng introduced a partial-sum tensor nuclear norm to better approximate tensor rank and preserve small-target components [9]. Zhu et al. combined target-aware nonlocal low-rank modeling with saliency-filter regularization to improve target-background separation [10]. More recently, deep denoising priors have also been integrated with low-rank modeling to enhance representation capability [11]. Although these methods provide an interpretable optimization framework, their performance depends on whether the background satisfies the assumed low-rank property. Their iterative optimization procedures may al-so impose considerable computational costs.

## 2.2 Deep-Learning-Based Single-Frame Detection

Deep-learning-based methods learn infrared target representations directly from annotated images. Wang et al. formulated infrared small target detection as a seg-mentation problem and introduced adversarial learning to balance missed detections and false alarms [12]. Dai et al. proposed asymmetric contextual modulation to com-bine high-level semantic information with low-level spatial details [13]. They subsequently developed an attentional local contrast network that embeds traditional local contrast concepts into a trainable neural architecture [14]. These studies demonstrated that incorporating task-specific priors into deep networks can be beneficial for weak-target representation.

To preserve small-target features, Li et al. proposed a densely nested attention network with repeated feature aggregation and attention-guided fusion [15]. Wu et al. introduced UIU-Net, which embeds a small U-Net within a larger U-shaped architecture to enhance multilevel feature interaction [16]. Zhang et al. proposed ISNet and emphasized the importance of target shape and boundary information through a Tay-lor finite-difference-inspired edge block [17]. Zhang et al. developed an attention-guided pyramid context network to aggregate contextual information at multiple scales under complex backgrounds [18]. Liu et al. further investigated the scale and location sensitivity of infrared targets and constructed a framework that integrates shape, scale, and positional information [19].

Although these networks have achieved strong performance on public datasets, most of them process each frame independently. Consequently, their decisions are mainly based on spatial appearance and may become unstable when a target is extremely weak

or when a background structure resembles the target. Moreover, infrared small targets contain very few pixels, and their features can be attenuated or lost during repeated downsampling. The limited amount and diversity of annotated infrared data also make these models vulnerable to domain shifts between different sensors, scenes, and imaging conditions.

### 2.3 Multi-Frame and Spatiotemporal Detection

Multi-frame approaches exploit motion continuity and temporal correlation to distinguish moving targets from spatial interference. Li et al. proposed a spatiotemporal saliency method for airborne infrared sequences, combining spatial saliency with temporal motion information to improve robustness against complex back-grounds [20]. Deng et al. constructed a spatial–temporal local contrast filter to enhance moving point targets by jointly measuring spatial and temporal differences [21]. Du and Hamdulla subsequently proposed a spatial–temporal local difference measure that extracts moving targets through neighborhood differences across consecutive frames [22]. These methods demonstrate that temporal information can effectively eliminate transient noise and static target-like structures. However, fixed temporal neighborhoods may become unreliable when target velocity changes or when the background undergoes noticeable motion.

Tensor-based methods organize consecutive infrared frames into multidimensional data and describe temporal redundancy through low-rank constraints. Sun et al. proposed a spatial–temporal infrared patch-tensor model with weighted Schatten p-norm minimization [23]. Liu et al. represented infrared videos using a spatiotemporal tensor model to separate sparse moving targets from correlated back-grounds [24]. Pang et al.

further introduced spatial–temporal tensor modeling with saliency filter regularization, incorporating target saliency into the decomposition process [25]. Such methods provide a structured way to exploit temporal redundancy, but iterative tensor optimization often incurs substantial computational overhead. Their performance may also deteriorate when camera motion, parallax, or dynamic clutter destroys the assumed low-rank background structure.

Recent deep networks have attempted to learn spatiotemporal representations in an end-to-end manner. Yan et al. proposed a spatiotemporal differential multiscale attention network that combines temporal differencing with multiscale feature enhancement [26]. Chen et al. developed SSTNet, which slices a video sequence along multiple temporal directions and employs cross-slice ConvLSTM modules to extract motion-aware features [27]. Lu et al. introduced a long-term spatial–temporal tensor model for infrared small target detection under dynamic backgrounds and established a motion-aware patch-tensor representation to accommodate background displacement [28]. Compared with handcrafted temporal filters, these approaches can model more complicated motion patterns. Nevertheless, reliable temporal association re-mains difficult under low signal-to-clutter ratios, irregular target motion, platform vi-bration, and large background displacement. Therefore, the effective integration of spatial target cues and motion-adaptive temporal information remains an important research problem.

## 3 Methods

Let $I_{\mathrm{t}}(p)$ denote the $\mathrm{t}-th$ infrared frame at pixel $p$. After global subpixel registration, the residual displacement of the target within a short analysis window is

assumed not to exceed approximately 0.5 pixel. Temporal evidence can therefore be accumulated at a fixed pixel rather than along a searched trajectory.

The proposed method consists of three main stages. First, radiometric normalization, subpixel registration, and registration-quality control establish a common radiometric and geometric coordinate system. Second, a multi-scale spatial response and a polarized background residual drive the candidate-gated background model and fixed-pixel temporal-profile model. Finally, the spatial and temporal evidence is fused into a saliency map, followed by robust local CFAR, multi-frame confirmation, and non-maximum suppression. The overall framework of the proposed infrared small target detection method is illustrated in **Fig. 1**.

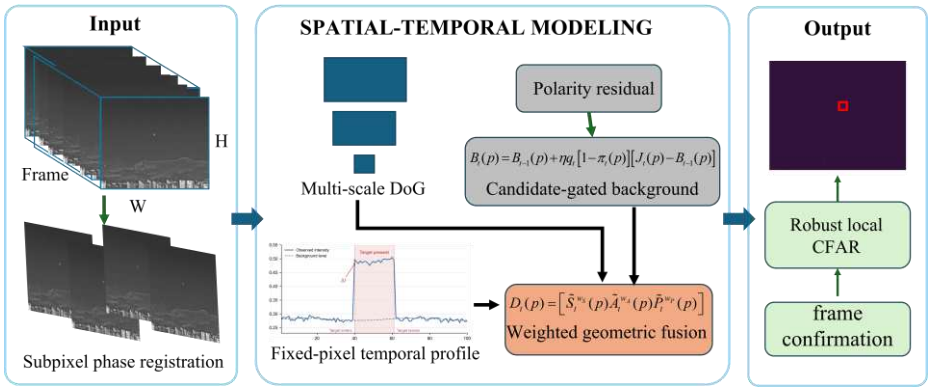



**Fig. 1.** Overall framework of the proposed fixed-pixel infrared small target detection method. After radiometric normalization and quality-controlled subpixel registration, the spatial point response and fixed-pixel temporal profile are modeled through candidate-gated background updating and temporal consistency analysis. Their complementary evidence is fused into a spatiotemporal saliency map, which is subsequently converted into target coordinates through robust local CFAR, $M$ / $N$ confirmation, and non-maximum suppression.

### 3.1 Radiometric Normalization and Registration

Frame-wise gain and offset variations may generate widespread background residuals. To suppress these variations, the current frame is matched to the reference frame using robust quantiles:

$$J_t^{\mathrm{rad}}(p) = \left[ I_t(p) - Q_{0.50}(I_t) \right] \frac{Q_{0.98}(I_{\mathrm{ref}}) - Q_{0.02}(I_{\mathrm{ref}})}{Q_{0.98}(I_t) - Q_{0.02}(I_t) + \varepsilon} + Q_{0.50}(I_{\mathrm{ref}}) \quad (1)$$

where $Q_{\mathrm{a}}(\square)$ denotes the a-quantile. Compared with min–max normalization, quantile matching is less sensitive to hot pixels and strong outliers.

The normalized frame is registered to the reference frame using phase correlation. The normalized cross-power spectrum and correlation surface are

$$C_t(u,v) = \frac{F_t(u,v) F_{\mathrm{ref}}^*(u,v)}{\left| F_t(u,v) F_{\mathrm{ref}}^*(u,v) \right| + \varepsilon}, c_t(x,y) = \mathrm{F}^{-1}\left\{ C_t(u,v) \right\} \quad (2)$$

where $F_t$ and $F_{\mathrm{ref}}^*$ are the Fourier transforms of the current and reference frames. The correlation-peak location provides the global translation, and a parabolic fit around the peak gives the subpixel correction.

Registration reliability is measured by the peak-to-sidelobe ratio:

$$\mathrm{PSR}_t = \frac{c_{\max} - \mu_{\mathrm{side}}}{\sigma_{\mathrm{side}} + \varepsilon} \quad (3)$$

where $c_{\max}$ is the correlation peak, while $\mu_{\mathrm{side}}$ and $\sigma_{\mathrm{side}}$ are the mean and standard deviation of the sidelobe region. The PSR is mapped to a quality factor:

$$q_t = \mathrm{clip}\left( \frac{\mathrm{PSR}_t - \tau_{\mathrm{reject}}}{\tau_{\mathrm{good}} - \tau_{\mathrm{reject}}}, 0, 1 \right) \tag{4}$$

When $q_t$, the frame is excluded from background updating, temporal accumulation, and multi-frame confirmation. A valid-region mask additionally removes interpolation borders and incomplete local neighborhoods.

### 3.2 Multi-Scale Spatial Point Response

A weak infrared target generally exhibits a compact point-like structure. Its spatial response is extracted by multi-scale difference-of-Gaussians filtering and normalized by the local background fluctuation:

$$S_t(p) = \max_{\sigma \in \Sigma} \frac{\left[ \left( G_\sigma - G_{\alpha\sigma} \right) * \left( \rho J_t \right) \right]^+ (p)}{\sigma_{\mathrm{loc},t}(p) + \varepsilon} \tag{5}$$

where $\left[ \mathrm{x} \right]^+ = \max(\mathrm{x}, 0)$, $\sum$ is the scale set, and $\sigma_{\mathrm{loc},t}$ is the local standard deviation. The scale set was $\{0.8, 1.2, 1.6\}$ pixels, with the larger Gaussian scale set to $1.8\sigma$. The local standard deviation was calculated within an $11 \times 11$ neighborhood.

The use of multiple scales improves the response to targets occupying different numbers of pixels, while local normalization reduces the influence of spatially nonuniform background clutter.

The initial background is estimated from quality-accepted initialization frames:

$$B_0(p) = median_{t \in \mathrm{I}_0, q_t > 0} J_t(p) \tag{6}$$

The initial background was estimated from 30 quality-accepted frames. Stable point anomalies already present in the initial background are marked as static clutter:

$$M_{\text{static}}(p) = \mathbf{1}\left[ S_{B_0}(p) > \tau_{\text{hot}} \right] \tag{7}$$

This mask mainly suppresses fixed bright points and hot pixels.

A conventional exponential background model applies the same learning rate to all pixels. Consequently, a weak target that remains at one registered position may be gradually incorporated into the background.

The polarized positive residual and its locally standardized strength are defined as

$$R_t(p) = \left[ \rho\left( J_t(p) - B_{t-1}(p) \right) \right]^{+} \tag{8}$$

$$Z_t(p) = \min\left[ \frac{R_t(p)}{\sigma_{\text{loc},t}(p) + \varepsilon}, Z_{\max} \right] \tag{9}$$

The spatial response and residual strength are converted into soft target likelihoods:

$$\pi_t^S(p) = \text{clip}\left( \frac{S_t(p) - \tau_S^{\text{low}}}{\tau_S^{\text{high}} - \tau_S^{\text{low}}}, 0, 1 \right) \tag{10}$$

$$\pi_t^R(p) = \text{clip}\left( \frac{Z_t(p) - \tau_R^{\text{low}}}{\tau_R^{\text{high}} - \tau_R^{\text{low}}}, 0, 1 \right) \tag{11}$$

$$\pi_t(p) = \max\left[ \pi_t^S(p), \pi_t^R(p) \right] \tag{12}$$

The background is then updated by

$$B_t(p) = B_{t-1}(p) + \eta q_t \left[ 1 - \pi_t(p) \right] \left[ J_t(p) - B_{t-1}(p) \right] \tag{13}$$

where $\eta$ is the nominal learning rate, which was set to 0.015 in all experiments. The proposed candidate-gated adaptive background update mechanism is illustrated in **Fig. 2**.

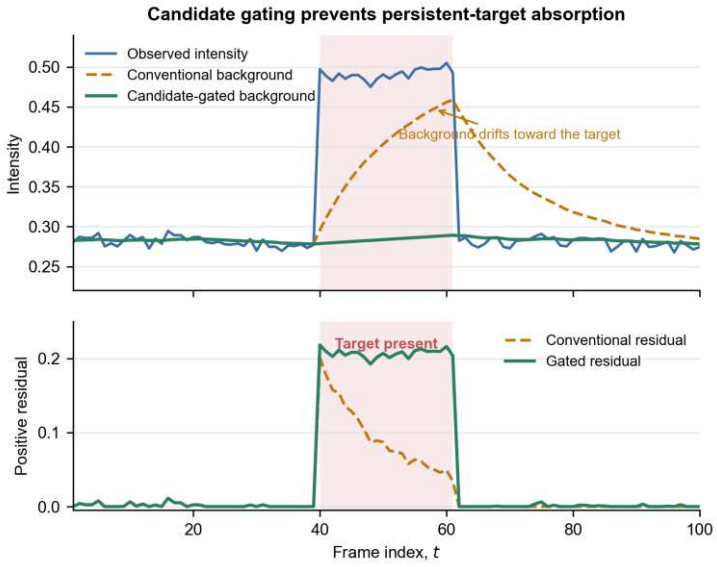



**Fig. 2.** Candidate-gated adaptive background update mechanism. A soft target likelihood is jointly constructed from the multi-scale spatial response and the locally normalized background residual, and is then used to modulate the pixel-wise background learning rate. Background pixels retain their adaptation capability, whereas target-like pixels are updated slowly, thereby mitigating the absorption of persistent weak targets.

When a pixel is unlikely to contain a target, $\pi_t(p)$ is small and the background maintains its adaptation capability. When the pixel exhibits a strong target-like response, $\pi_t(p)$ increases and the background learning rate is reduced. This mechanism suppresses target absorption without completely freezing the background model.

### 3.3 Fixed-Pixel Temporal-Profile Consistency

Let $\mathrm{W}_t$ contain the most recent $n_t \le L$ quality-accepted residual observations. The temporal window length was set to $L = 6$. The mean residual energy is

$$A_t(p) = \frac{1}{n_t} \sum_{i \in \mathrm{W}_t} Z_i(p) \tag{14}$$

and the temporal dispersion is

$$\sigma_{Z,t}(p) = \sqrt{\frac{1}{n_t} \sum_{i \in \mathrm{W}_t} Z_i^2(p) - A_t^2(p)} \tag{15}$$

The temporal stability is defined as

$$P_t(p) = \min\left[ \frac{A_t(p)}{\sigma_{Z,t}(p) + \varepsilon_s}, P_{\max} \right] \tag{16}$$

A high value of $A_t$ indicates persistent residual energy, whereas a high value of $P_t$ indicates that this energy remains stable rather than fluctuating randomly.

The run-length continuity is calculated as

$$\ell_t(p) = \begin{cases} \ell_{t-1}(p) + 1, & S_t(p) \geq \tau_S \text{ and } Z_t(p) \geq \tau_Z, \\ [0, & \text{otherwise.} \end{cases} \tag{17}$$

These three quantities describe different temporal properties. Residual energy measures persistence, stability suppresses flickering responses, and run length imposes local continuity. Their joint use distinguishes persistent targets from isolated or intermittent noise.

### 3.4 Joint Spatiotemporal Saliency Generation

The spatial response, temporal energy, and temporal stability are independently mapped to $[0,1]$, producing $\tilde{S}_t^{w_S}(p)$, $\tilde{A}_t^{w_A}(p)$, $\tilde{P}_t^{w_P}(p)$.They are fused by a weighted geometric mean:

$$D_t(p)=\left[\tilde{S}_t^{w_S}(p)\tilde{A}_t^{w_A}(p)\tilde{P}_t^{w_P}(p)\right]^{\frac{1}{w_S+w_A+w_P}}\psi\left(\ell_t(p)\right) \tag{18}$$

where the continuity modulation term is

$$\psi\left(\ell_t(p)\right)=0.5+0.5\,\mathrm{clip}\left(\frac{\ell_t(p)}{\ell_{\min}},0,1\right) \tag{19}$$

This joint saliency mechanism is deliberately multiplicative rather than additive. It acts as a soft logical AND among the spatial and temporal evidence. A strong value in only one feature cannot completely compensate for weak support from the remaining features.

For example, a hot pixel or structural edge may produce a strong spatial response but may be rejected by the temporal profile or static-clutter mask. Conversely, flickering noise may produce a temporary residual response but lacks stable point-like morphology and temporal continuity. A genuine target is more likely to obtain simultaneous support from spatial compactness, residual energy, temporal stability, and continuity.

Therefore, the fused saliency map emphasizes weak targets with consistent spatio-temporal evidence while suppressing single-feature false alarms.

### 3.5 Robust Decision and Multi-Frame Confirmation

A spatially uniform threshold is unsuitable for infrared scenes with nonuniform clutter. The local median and median absolute deviation are therefore calculated as

$$m_t(p)=median_{q\in\Omega_p}D_t(q) \tag{20}$$

$$\mathrm{MAD}_t(p)=median_{q\in\Omega_p}\left|D_t(q)-m_t(p)\right| \tag{21}$$

where $\Omega_p$ denotes the local neighborhood centered at $p$. The local robust CFAR threshold is defined as:

$$T_t(p) = \max\left[m_t(p) + 1.4826k\,\mathrm{MAD}_t(p), T_{\min}\right] \tag{22}$$

A $31\times31$ local window and a CFAR coefficient of $\mathrm{k} = 4.0$ were used, with the minimum saliency threshold set to 0.25. The single-frame hit map is

$$H_t(p) = \mathbf{1}\left[D_t(p) \geq T_t(p)\right] M_{\text{valid}}(p)\left[1 - M_{\text{static}}(p)\right] \tag{23}$$

The confirmation parameters were set to $M = 4$ and $N = 8$.A candidate is confirmed when it is detected in the current frame and has at least $M$ hits within the most recent $N$ accepted frames:

$$C_t(p) = H_t(p)\mathbf{1}\left[\sum_{i=0}^{N-1} H_{t-i}(p) \geq M\right] \tag{24}$$

Finally, non-maximum suppression retains the strongest response within each small neighborhood and converts the confirmed response into a target center.

For a retained target position, the confidence score is calculated as

$$\mathrm{Conf}_t(p) = D_t(p) q_t \min\left(\frac{h_t(p)}{M}, 1\right) \tag{25}$$

where $h_t(p)$ is the number of hits in the confirmation window. This quantity is an engineering reliability score rather than a calibrated posterior probability of the target category.

# 4 Experimental Results and Analysis

All experiments were implemented in MATLAB R2024b and run on a desktop computer equipped with an Intel Core i5-14600KF CPU, 32 GB of RAM. Unless otherwise stated, the same parameter settings were used for all six sequences.

For a fair comparison, all methods were evaluated using the same image sequences and ground-truth annotations. The parameters of the comparison methods were set according to their original publications or publicly available implementations. SCRG values were calculated using identical target and local-background regions, and ROC curves were generated by varying the detection threshold over the same evaluation frames.

## 4.1 Dataset Description

A total of six infrared image sequences were used to evaluate the proposed method. Sequences 1–3 were independently collected by our team using a long-wave infrared camera with a pixel pitch of 12 μm mounted on a UAV. During data acquisition, the UAV remained in continuous flight, resulting in variations in flight altitude, viewing angle, and background content. The targets appearing in these sequences were real targets captured under actual imaging conditions, and no synthetic target insertion was performed. Sequences 4–6 were selected from the publicly available ATR dataset [29], which was collected by the ATR Key Laboratory of the National University of Defense Technology between 2017 and 2019. These sequences were acquired using an infrared sensor mounted on a high-tower dual-axis gimbal system and contain diverse backgrounds and observation perspectives. The combination of UAV-based aerial sequences and publicly available ground-based sequences provides a varied test set

involving different imaging platforms, viewing conditions, and background characteristics. Representative frames and enlarged target regions from the six sequences are presented in **Fig. 3**, and their detailed information is summarized in **Table 1**.

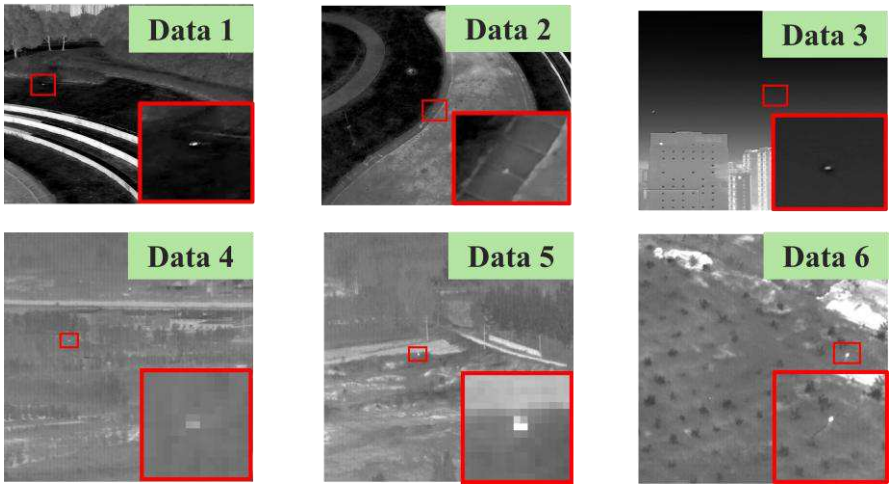



**Fig. 3.** Representative frames from the six infrared test sequences, including three self-collected UAV sequences and three sequences selected from the ATR dataset [29]. Enlarged views of the corresponding small targets are shown below each frame.

**Table 1.** Summary of the Infrared Sequences Used for Evaluation.

| Sequence | Image size | Number of frames | Average SCR | Background type |
|---|---|---|---|---|
| Seq.1 | 1024×1280 | 500 | 5.8565 | Trees, Ground |
| Seq.2 | 1024×1280 | 500 | 3.6930 | Road, Trees |
| Seq.3 | 1024×1280 | 500 | 4.3727 | Buildings, Sky |
| Seq.4 | 256×256 | 400 | 5.0441 | Trees, Buildings |
| Seq.5 | 256×256 | 400 | 2.2899 | Trees, Buildings |

| | | | | |
|---|---|---|---|---|
| Seq.6 | 256×256 | 400 | 5.2720 | Grassland, Buildings |

### 4.2 Evaluation Metrics

To quantitatively assess the detection performance, the signal-to-clutter ratio gain (SCRG) is employed to measure the improvement in target saliency after processing. The signal-to-clutter ratio (SCR) and SCRG are defined as follows:

$$SCR = \left|\frac{f_T - \mu_B}{\sigma_B}\right|, SCRG = \frac{SCR_{out}}{SCR_{in}} \tag{26}$$

where $f_T$ denotes the intensity of the target, while $\mu_B$ and $\sigma_B$ represent the mean intensity and standard deviation of the local background, respectively. The size of the background neighborhood is set to approximately three times that of the target region. $SCR_{in}$ and $SCR_{out}$ denote the SCR values calculated from the original and processed images, respectively. A larger SCRG indicates that the method provides stronger target enhancement and background clutter suppression.

The receiver operating characteristic (ROC) curve is also used to evaluate the overall detection performance. It describes the relationship between the target detection rate (TDR) and the false alarm rate (FAR) under different detection thresholds. The two metrics are calculated as follows:

$$TDR = \frac{N_{TD}}{N_{GT}}, \mathrm{FA}R = \frac{N_{FA}}{N_P} \tag{27}$$

where $N_{TD}$ is the number of correctly detected targets, $N_{GT}$ is the total number of ground-truth targets, $N_{FA}$ represents the number of false-alarm pixels, and $N_P$ denotes the total number of pixels in the evaluated images. A higher TDR and a lower FAR indicate better detection performance.

### 4.3 Baseline Methods

Six representative single-frame methods are selected for comparison, including absolute average difference weighted by cumulative directional derivatives (AADCDD) [30], local contrast measure (LCM) [3], local intensity and gradient (LIG) [31], SOG [34], tri-layer local contrast measure (TLLCM) [6], and weighted strengthened local contrast measure (WSLCM) [32]. These methods cover derivative-based characterization, intensity–gradient modeling, and local-contrast enhancement, thereby providing diverse benchmarks for evaluating target enhancement and background-clutter suppression.

Detection performance is evaluated in terms of the probability of detection and false-alarm rate, and receiver operating characteristic (ROC) curves are obtained by varying the detection threshold. Since the proposed method jointly exploits spatial and temporal information from infrared image sequences, four representative multi-frame methods are selected for comparison, including the non-convex tensor low-rank approximation method with asymmetric spatial–temporal total variation regularization (ASTTV-NTLA) [33], spatial–temporal local contrast filter (STLCF) [21], spatial–temporal local difference measure (STLDM) [22], and spatial–temporal infrared patch-tensor model with weighted Schatten p-norm minimization (WSNM-STIPT) [23]. Deep-learning-based methods were not included because they mainly address single-frame segmentation, require task-specific training data, and are not directly compatible with the multi-frame coordinate-output setting considered here.

### 4.4 Comparative Results and Analysis

To comprehensively evaluate the target-enhancement and clutter-suppression capabilities of the proposed method, quantitative and qualitative comparisons were conducted on six infrared sequences. The SCRG results are reported in Table 2. The normalized response heatmaps for Sequence 1, the three-dimensional response maps for Sequence 4, and the ROC curves for all six sequences are presented in **Fig. 4**–**6**, respectively.

As shown in Table 1, the average SCR values of the six sequences range from 2.2899 to 5.8565, covering different levels of target visibility. The proposed method achieves the highest SCRG on Sequences 1, 3, 4, and 5, while ranking second on Sequence 2 and third on Sequence 6. The improvement can be interpreted in relation to the processing stages described in Section 3. Robust quantile-based radiometric normalization first reduces frame-wise gain and offset variations. Quality-controlled subpixel registration then establishes the pixel correspondence required by the fixed-pixel temporal model, while frames with unreliable registration quality are excluded from temporal accumulation and background updating. This processing reduces the possibility that global platform motion or inaccurate registration will be incorrectly interpreted as target motion.

**Table 2.** Quantitative comparison of the signal-to-clutter ratio gain (SCRG) achieved by different methods on six infrared sequences.

| Method | Seq.1 | Seq.2 | Seq.3 | Seq.4 | Seq.5 | Seq.6 |
|---|---|---|---|---|---|---|
| AADCDD | 1.4722 | 1.3561 | 4.7010 | 0.9686 | 2.4586 | 5.1589 |
| LCM | 0.000 | 0.0000 | 0.3480 | 4.6953 | 0.8128 | 0.4700 |

| | | | | | | |
|---|---|---|---|---|---|---|
| LIG | 1.4793 | 0.9967 | 3.2631 | 4.0883 | 8.5174 | 6.6622 |
| SOG | 1.0121 | 0.8861 | 5.0500 | 2.9109 | 1.2326 | 2.2226 |
| TLLCM | 0.9129 | 0.5305 | 3.3606 | 7.6562 | 4.2893 | 11.6848 |
| WSLCM | 0.0000 | 0.2568 | 3.3260 | 1.0222 | 9.2577 | 11.7370 |
| ASTTV-NTLA | 1.1213 | 1.0380 | 4.1065 | 4.1341 | 4.5501 | 3.7432 |
| STLCF | 0.9504 | 0.4698 | 1.2641 | 1.4538 | 0.4128 | 0.6733 |
| STLDM | 1.2249 | 0.2105 | 0.8455 | 8.9893 | 0.5867 | 3.0958 |
| WSNM-STIPT | 0.3234 | 0.7549 | 8.9284 | 10.8767 | 7.6308 | 3.8422 |
| Ours | 1.7363 | 1.2640 | 9.4706 | 11.7899 | 10.0665 | 10.4346 |

As shown in **Fig. 4**, several comparison methods retain curved structures, diffuse responses, or scattered salient regions in the response heatmaps of Sequence 1. In contrast, the proposed method concentrates the response within the target box while maintaining a low response over most of the background. This result indicates that the joint spatial–temporal saliency mechanism can preserve the compact target response and suppress target-like background structures. The visual result is consistent with the highest SCRG of 1.7363 obtained by the proposed method on Sequence 1.

The three-dimensional response maps for Sequence 4 are presented in **Fig. 5**. Several comparison methods produce multiple background peaks, continuous response ridges, or an elevated response pedestal. Although the target may be enhanced by these methods, the presence of competing peaks decreases the separability between the true target and background interference.

The proposed method produces a sharp and spatially isolated response peak within the target region, while the surrounding response surface remains relatively low and flat. This response distribution is consistent with the multiplicative spatiotemporal fusion mechanism adopted in the proposed method. The weighted geometric fusion acts as a soft logical AND: a response must receive simultaneous support from spatial compactness, residual energy, temporal stability, and run-length continuity to remain salient. Therefore, a strong response produced by only one feature cannot completely compensate for weak evidence in the remaining features.

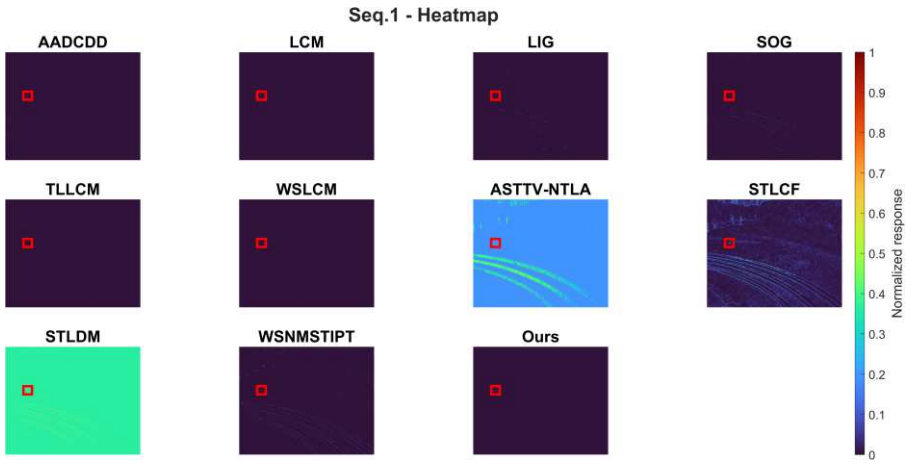



**Fig. 4.** Normalized response heatmaps obtained by different methods on Sequence 1.

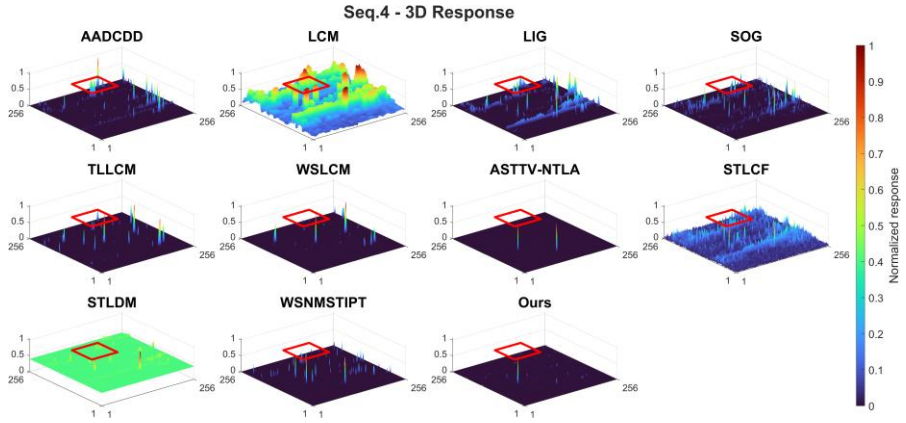



**Fig. 5.** Three-dimensional normalized response maps obtained by different methods on Sequence 4.

As shown in **Fig. 6**, the ROC curves of the proposed method remain close to the upper-left corner across the six sequences. This indicates that the method can achieve a high target detection rate at relatively low FAR values. The advantage is especially evident in the low-false-alarm region, where the proposed method rapidly reaches or maintains a high target detection rate.

In addition to the spatiotemporal saliency generation, the robust decision strategy contributes to the observed detection behavior. The local median and median absolute deviation are used to construct a spatially adaptive CFAR threshold, making the decision process less sensitive to nonuniform background statistics and strong outliers. The subsequent multi-frame confirmation suppresses candidates that do not persist across accepted frames, while non-maximum suppression removes redundant neighboring responses.

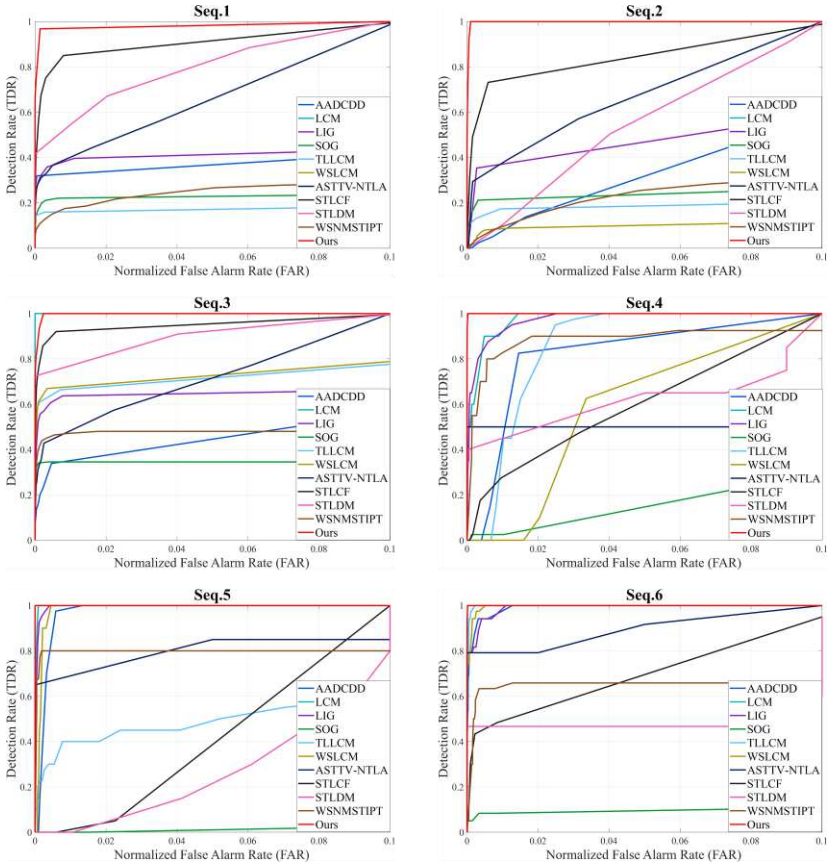



**Fig. 6.** ROC curves of different methods on the six infrared sequences. The six subplots correspond to Sequences 1–6, respectively. A curve closer to the upper-left corner indicates a higher target detection rate at a lower normalized false-alarm rate.

### 4.5 Ablation Study

To investigate the contribution of each major component, an incremental ablation study was conducted on Sequence 1. During the experiment, radiometric normalization, registration-quality control, CFAR detection, multi-frame confirmation, non-maximum suppression, and all parameter settings were kept unchanged. Variant A retains only the

multi-scale spatial response and uses conventional background updating. Variant B introduces the candidate-gated adaptive background update into Variant A. Variant C further incorporates the fixed-pixel temporal profile, including temporal residual energy, stability, and run-length continuity, while the spatial and temporal features are combined using a weighted arithmetic mean. Variant D replaces the arithmetic fusion with the proposed weighted geometric fusion and represents the complete model. In Table 3, "✓" and "×" indicate that the corresponding component is enabled and disabled, respectively, whereas "—" indicates that spatiotemporal fusion is not applicable because the temporal branch is not used.

**Table 3.** Ablation results of the main components on Sequence 1.

| Variant | Multi-scale spatial response | Candidate-gated background update | Fixed-pixel temporal profile | Weighted geometric fusion | SCRG |
|---|---|---|---|---|---|
| A | ✓ | × | × | — | 0.956 |
| B | ✓ | ✓ | × | — | 1.348 |
| C | ✓ | ✓ | ✓ | × | 1.543 |
| D | ✓ | ✓ | ✓ | ✓ | 1.736 |

As shown in Table 3, Variant A achieves an SCRG of 0.956 when only the multi-scale spatial response is employed. After introducing the candidate-gated background update, the SCRG increases to 1.348, corresponding to an absolute improvement of 0.392 and a relative improvement of 41.00%. This result indicates that adaptively reducing the

background learning rate at target-like pixels can alleviate the absorption of persistent weak targets and preserve their residual contrast. Within the adopted incremental ablation configuration, candidate-gated background updating provides the largest individual improvement.

When the fixed-pixel temporal profile is further incorporated, Variant C achieves an SCRG of 1.543, which is 14.47% higher than that of Variant B. This improvement demonstrates that temporal residual energy, stability, and run-length continuity provide complementary information for distinguishing persistent targets from spatially target-like or temporally unstable clutter. However, because Variant C employs arithmetic fusion, a strong response from one feature may partially compensate for weak evidence from another feature.

The complete model achieves the highest SCRG of 1.736. Compared with Variant C, the weighted geometric fusion produces a further improvement of 12.51%. This result supports the use of multiplicative fusion as a soft logical AND, which emphasizes pixels receiving consistent support from both spatial and temporal features while suppressing single-feature false responses. Overall, the complete model improves the SCRG by 81.59% relative to the spatial-only Variant A, demonstrating the complementary effects of candidate-gated background modeling, fixed-pixel temporal-profile consistency, and joint spatiotemporal saliency generation on Sequence 1.

# 5 Discussion

The experimental results demonstrate that the proposed method provides consistent target enhancement across different imaging platforms and background conditions. It

achieves the highest SCRG on four of the six sequences and remains among the top three methods on every sequence, with a mean SCRG improvement of 38.34% over the strongest baseline. The heatmap and three-dimensional response results further show that the method concentrates the target response while suppressing structured background residuals and competing peaks. Its ROC curves remain close to the upper-left corner, indicating a favorable balance between target detection and false-alarm suppression. Although phase correlation compensates mainly for translation, rotation, scale variation, and parallax remain limitations and may degrade performance under complex UAV motion.

The performance improvement can be attributed to the complementary roles of the proposed components. Candidate-gated background updating preserves persistent weak-target residuals that may otherwise be absorbed by conventional background adaptation. Fixed-pixel temporal profiling then uses residual energy, stability, and run-length continuity to reject spatially salient but temporally inconsistent interference. Finally, weighted geometric fusion requires simultaneous spatial and temporal support, limiting the influence of a strong but unreliable individual cue. This interpretation is supported by the ablation results, in which the SCRG on Sequence 1 increases progressively from 0.956 for the spatial-only configuration to 1.348 after candidate gating, 1.543 after temporal-profile modeling, and 1.7363 for the complete model.

The proposed method nevertheless depends on sufficiently accurate registration, because strong parallax, local deformation, or abrupt target motion may violate the fixed-pixel assumption. In addition, the present evaluation contains six sequences, and the ablation study is limited to Sequence 1. Future work will therefore investigate local motion compensation and adaptive temporal association, extend the evaluation to

additional sensors and datasets, and provide a systematic analysis of computational efficiency and parameter sensitivity.

# 6 Conclusions

This paper presented an infrared small target detection method based on candidate-gated adaptive background modeling and fixed-pixel spatiotemporal saliency. Robust radiometric normalization and quality-controlled subpixel registration first establish reliable correspondence between consecutive frames. A multi-scale spatial point response and locally normalized positive residual are then used to construct a soft target likelihood that adaptively controls the background learning rate. Temporal residual energy, stability, and run-length continuity are accumulated at fixed pixels and integrated with the spatial response through weighted geometric fusion. Robust local CFAR, multi-frame confirmation, and non-maximum suppression subsequently convert the fused saliency map into target coordinates.

Experiments on six infrared sequences show that the proposed method achieves the highest SCRG on four sequences and ranks among the top three methods on all six. Its average SCRG of 7.4603 is 38.34% higher than the best baseline average. The ROC results demonstrate favorable detection performance under low false-alarm conditions. Furthermore, the ablation experiment on Sequence 1 shows that the complete model improves SCRG from 0.956 to 1.7363, corresponding to an improvement of 81.59% over the spatial-only configuration. These results verify the complementary roles of candidate-gated background updating, fixed-pixel temporal-profile modeling, and multiplicative spatiotemporal fusion. Future work will focus on improving robustness to

local motion and parallax, reducing parameter dependence, and evaluating computational efficiency on broader infrared datasets.

## 7 Declarations

**Availability of data and materials**: The public ATR dataset used in this study is described in Reference [29]. The self-collected infrared sequences and associated annotations used in this study are available from the corresponding author upon reasonable request.

**Competing interests**: The author declares that he has no competing interests.

Funding: This research received no external funding.

**Authors' contributions**: XL developed the methodology, implemented the algorithms, conducted the experiments, analyzed the results, and drafted and revised the manuscript. The author read and approved the final manuscript.

**Acknowledgements**: Not applicable.

## References

1. M Zhao, W Li, L Li, J Hu, P Ma, R Tao, Single-frame infrared small-target detection: A survey. IEEE Geosci. Remote Sens. Mag. 10, 87–119 (2022). doi:10.1109/MGRS.2022.3145502.
2. X Bai, F Zhou, Analysis of new top-hat transformation and the application for infrared dim small target detection. Pattern Recognit. 43, 2145–2156 (2010). doi:10.1016/j.patcog.2009.12.023.

3. CLP Chen, H Li, Y Wei, T Xia, YY Tang, A local contrast method for small infrared target detection. IEEE Trans. Geosci. Remote Sens. 52, 574–581 (2014). doi:10.1109/TGRS.2013.2242477.
4. Y Wei, X You, H Li, Multiscale patch-based contrast measure for small infrared target detection. Pattern Recognit. 58, 216–226 (2016). doi:10.1016/j.patcog.2016.04.002.
5. H Deng, X Sun, M Liu, C Ye, X Zhou, Small infrared target detection based on weighted local difference measure. IEEE Trans. Geosci. Remote Sens. 54, 4204–4214 (2016). doi:10.1109/TGRS.2016.2538295.
6. J Han, S Moradi, I Faramarzi, C Liu, H Zhang, Q Zhao, A local contrast method for infrared small-target detection utilizing a tri-layer window. IEEE Geosci. Remote Sens. Lett. 17, 1822–1826 (2020). doi:10.1109/LGRS.2019.2954578.
7. C Gao, D Meng, Y Yang, Y Wang, X Zhou, AG Hauptmann, Infrared patch-image model for small target detection in a single image. IEEE Trans. Image Process. 22, 4996–5009 (2013). doi:10.1109/TIP.2013.2281420.
8. Y Dai, Y Wu, 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. 10, 3752–3767 (2017). doi:10.1109/JSTARS.2017.2700023.
9. L Zhang, Z Peng, Infrared small target detection based on partial sum of the tensor nuclear norm. Remote Sens. 11, 382 (2019). doi:10.3390/rs11040382.
10. H Zhu, H Ni, S Liu, G Xu, L Deng, TNLRS: Target-aware non-local low-rank modeling with saliency filtering regularization for infrared small target detection. IEEE Trans. Image Process. 29, 9546–9558 (2020). doi:10.1109/TIP.2020.3028457.
11. T Liu, Q Yin, J Yang, Y Wang, W An, Combining deep denoiser and low-rank priors for infrared small target detection. Pattern Recognit. 135, 109184 (2023). doi:10.1016/j.patcog.2022.109184.

12. H Wang, L Zhou, L Wang, Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images, in Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019, pp. 8508–8517. doi:10.1109/ICCV.2019.00860.
13. Y Dai, Y Wu, F Zhou, K Barnard, Asymmetric contextual modulation for infrared small target detection, in Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2021, pp. 949–958. doi:10.1109/WACV48630.2021.00099.
14. Y Dai, Y Wu, F Zhou, K Barnard, Attentional local contrast networks for infrared small target detection. IEEE Trans. Geosci. Remote Sens. 59, 9813–9824 (2021). doi:10.1109/TGRS.2020.3044958.
15. B Li, C Xiao, L Wang, Y Wang, Z Lin, M Li, W An, Y Guo, Dense nested attention network for infrared small target detection. IEEE Trans. Image Process. 32, 1745–1758 (2023). doi:10.1109/TIP.2022.3199107.
16. X Wu, D Hong, J Chanussot, UIU-Net: U-Net in U-Net for infrared small object detection. IEEE Trans. Image Process. 32, 364–376 (2023). doi:10.1109/TIP.2022.3228497.
17. M Zhang, R Zhang, Y Yang, H Bai, J Zhang, J Guo, 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, 19–24 June 2022, pp. 867–876. doi:10.1109/CVPR52688.2022.00095.
18. T Zhang, L Li, S Cao, T Pu, Z Peng, Attention-guided pyramid context networks for detecting infrared small target under complex background. IEEE Trans. Aerosp. Electron. Syst. 59, 4250–4261 (2023). doi:10.1109/TAES.2023.3238703.
19. Q Liu, R Liu, B Zheng, H Wang, Y Fu, Infrared small target detection with scale and location sensitivity, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition, Seattle, WA, USA, 16–22 June 2024, pp. 17490–17499. doi:10.1109/CVPR52733.2024.01656.

20. Y Li, Y Zhang, J-G Yu, Y Tan, J Tian, J Ma, A novel spatio-temporal saliency approach for robust dim moving target detection from airborne infrared image sequences. Inf. Sci. 369, 548–563 (2016). doi:10.1016/j.ins.2016.07.042.
21. L Deng, H Zhu, C Tao, Y Wei, Infrared moving point target detection based on spatial–temporal local contrast filter. Infrared Phys. Technol. 76, 168–173 (2016). doi:10.1016/j.infrared.2016.02.010.
22. P Du, A Hamdulla, Infrared moving small-target detection using spatial–temporal local difference measure. IEEE Geosci. Remote Sens. Lett. 17, 1817–1821 (2020). doi:10.1109/LGRS.2019.2954715.
23. Y Sun, J Yang, M Li, W An, Infrared small target detection via spatial–temporal infrared patch-tensor model and weighted Schatten p-norm minimization. Infrared Phys. Technol. 102, 103050 (2019). doi:10.1016/j.infrared.2019.103050.
24. H-K Liu, L Zhang, H Huang, Small target detection in infrared videos based on spatio-temporal tensor model. IEEE Trans. Geosci. Remote Sens. 58, 8689–8700 (2020). doi:10.1109/TGRS.2020.2989825.
25. D Pang, P Ma, T Shan, W Li, R Tao, Y Ma, T Wang, STTM-SFR: Spatial–temporal tensor modeling with saliency filter regularization for infrared small target detection. IEEE Trans. Geosci. Remote Sens. 60, 5623418 (2022). doi:10.1109/TGRS.2022.3172745.
26. P Yan, R Hou, X Duan, C Yue, X Wang, X Cao, STDMANet: Spatio-temporal differential multiscale attention network for small moving infrared target detection. IEEE Trans. Geosci. Remote Sens. 61, 5602516 (2023). doi:10.1109/TGRS.2023.3241311.
27. S Chen, L Ji, J Zhu, M Ye, X Yao, SSTNet: Sliced spatio-temporal network with cross-slice ConvLSTM for moving infrared dim-small target detection. IEEE Trans. Geosci. Remote Sens. 62, 5000912 (2024). doi:10.1109/TGRS.2024.3350024.

28. D Lu, W An, Q Ling, D Cao, H Wang, M Li, Z Lin, LSTT: Long-term spatial–temporal tensor model for infrared small target detection under dynamic background. Remote Sens. 16, 2746 (2024). doi:10.3390/rs16152746.

29. B Hui, Z Song, H Fan, P Zhong, W Hu, X Zhang, J Ling, H Su, W Jin, Y Zhang, Y Bai, A dataset for infrared detection and tracking of dim-small aircraft targets under ground/air background. China Sci. Data 5, 286–297 (2020). doi:10.11922/csdata.2019.0074.zh.

30. S Aghaziyarati, S Moradi, H Talebi, Small infrared target detection using absolute average difference weighted by cumulative directional derivatives. Infrared Phys. Technol. 101, 78–87 (2019). doi:10.1016/j.infrared.2019.06.003.

31. H Zhang, L Zhang, D Yuan, H Chen, Infrared small target detection based on local intensity and gradient properties. Infrared Phys. Technol. 89, 88–96 (2018). doi:10.1016/j.infrared.2017.12.018.

32. J Han, S Moradi, I Faramarzi, H Zhang, Q Zhao, X Zhang, N Li, Infrared small target detection based on the weighted strengthened local contrast measure. IEEE Geosci. Remote Sens. Lett. 18, 1670–1674 (2021). doi:10.1109/LGRS.2020.3004978.

33. T Liu, J Yang, B Li, C Xiao, Y Sun, Y Wang, W An, Non-convex tensor low-rank approximation for infrared small target detection. IEEE Trans. Geosci. Remote Sens. 60, 5614718 (2022). doi:10.1109/TGRS.2021.3130310.

34. Y Xu, P Liu, W Qian, J Zhang, X Kong, M Wan, Small and dim target detection under strong clutter based on similarly of Gaussian and motion outlier significance using moving infrared camera. IEEE Sens. J. 25, 13357–13373 (2025). doi:10.1109/JSEN.2025.3543839.