Non-Convex Tensor Low-Rank Approximation for Infrared Small Target Detection

Ting Liu, Jungang Yang, Boyang Li, Chao Xiao, Yang Sun, Yingqian Wang, Wei An

Abstract—Infrared small target detection is an important fundamental task in the infrared system. Therefore, many infrared small target detection methods have been proposed, in which the low-rank model has been used as a powerful tool. However, most low-rank-based methods assign the same weights for different singular values, which will lead to inaccurate background estimation. Considering that different singular values have different importance and should be treated discriminatively, in this paper, we propose a non-convex tensor low-rank approximation (NTLA) method for infrared small target detection. In our method, NTLA regularization adaptively assigns different weights to different singular values for accurate background estimation. Based on the proposed NTLA, we propose asymmetric spatial-temporal total variation (ASTTV) regularization to achieve more accurate background estimation in complex scenes. Compared with the traditional total variation approach, ASTTV exploits different smoothness intensities for spatial and temporal regularization. We design an efficient algorithm to find the optimal solution of our method. Compared with some state-of-the-art methods, the proposed method achieves an improvement in terms of various evaluation metrics. Extensive experimental results in various complex scenes demonstrate that our method has strong robustness and low false-alarm rate. Code is available at https://github.com/LiuTing20a/ASTTV-NTLA.

Index Terms—Non-convex tensor low-rank, asymmetric spatial-temporal total variation, infrared small target detection.

I. INTRODUCTION

Infrared small target detection is an important technique in many military and civilian fields, such as buried landmine detection, night navigation, precision guided weapons and missiles [1]–[3]. However, due to the long imaging distance of infrared detection systems, the targets usually lack texture features or fixed shape. In addition, the target has low signal-to-clutter ratio (SCR) because it is always immersed in complex noises and strong clutters scenes. In summary, infrared small target detection is an important and challenging problem.

In the past decades, many scholars have devoted themselves to the research of infrared small target detection and proposed different methods for this task. These methods can be classified into single-frame and sequential infrared small target detection. Single-frame infrared small target detection can be divided into three categories according to different assumptions. The first category supposes that the background changes slowly in the infrared image and there is a high correlation between adjacent pixels. Based on this assumption, many background suppression (BS)-based methods [4]–[6] were proposed. These methods use filter to suppress the background noise and clutter, and then use the intensity threshold to extract small targets. The BS-based methods obtain satisfactory computation efficiency. However, they achieve relatively poor detection performance with high false alarm rates under discontinuous backgrounds, clutter and noise. Considering that the small target is more visually salient than its surrounding background, human visual system (HVS)-based [7]–[9] methods have been proposed. For example, Chen et al. [8] introduced a novel local contrast method (LCM). To improve the detection performance, a series of improved LCM (ILCM) [9] methods have been proposed, such as multiscale relative LCM [10], weighted strengthened local contrast measure (WSLCM) [11] and multiscale tri-layer local contrast measure (TLLCM) [12]. However, in some cases of highlighted background or ground background, clutter and target may be similar on the saliency map, which will degrade the performance of target detection.

The second category uses the nonlocal self-correlation between background patches in infrared images to construct low-rank sparse decomposition (LRSD) model, which is a branch of the popular low-rank representation (LRR) in recent years. By constructing local patches, Gao et al. [13] first designed a novel infrared patch-image model (IPI). Then the existing small target detection is transformed into a robust principal component analysis (RPCA) problem [14]. Subsequently, many LRSD methods have been proposed. However, due to the limitation of nuclear norm minimization (NNM), it will lead to over-shrinkage problem. To handle the above problems, Guo et al. [15] designed a novel reweighted IPI (ReWIPI) method, in which weighted nuclear norm minimization (WNNM) was introduced to suppress sparse non-target pixels. However, it can only alleviate the over-shrinkage problem. In [16], a non-convex rank approximation model (NRAM) was proposed, in which \( l_2,1 \) norm was used to constrain clutter. In recent years, considering the importance of model robustness, a series of multi-subspace structure methods have been proposed, such as low-rank and sparse representation (LRSR) [17], stable multi-subspace learning methods (SMSL) [18] and self-regularized weighted sparse (SRWS) [19]. Encouraged by the powerfulness of TV regularization, Wang et al. [20] introduced the TV regularization into the existing model (TV-PCP).

Compared with matrix domain, multi-directional tensor domain can exploit the inner relationship of the data from more views. Therefore, Dai et al. [21] developed a new reweighted infrared patch-tensor (RIPT) method. The RIPT

This work was supported in part by the National Natural Science Foundation of China under Grant 61972435, Grant 61401474, and Grant 61921001. (Corresponding author: Jungang Yang)

T. Liu, J. Yang, B. Li, C. Xiao, Y. Sun, Y. Wang, and W. An are with the College of Electronic Science, National University of Defense Technology, Changsha, China. (e-mail: liuting@nudt.edu.cn, yangjungang@nudt.edu.cn).
model achieves infrared target detection from the perspective of the tensor domain, and can achieve relatively good performance. However, due to the limitation of sum of nuclear norm (SNN), RIPT method achieves less competitive performance in complex scenes. Considering the above problem, Sun et al. [22] and Zhang et al. [23] exploited different tensor nuclear norms to improve the detection results of the RIPT model. Guan et al. [24] combined local contrast with nonconvex tensor rank surrogate (NTRS) to infrared small target detection. In [25], non-convex triple tensor factorisation was used for target detection. Further, Kong et al. [26] proposed a new nonconvex tensor fibered rank approximation model. The third category is deep learning based method. Recently, deep learning based methods have attracted extensive attention due to its powerful feature learning ability. It is widely used in infrared small target detection [27]–[30]. Although they achieved improved performance, the main challenge of deep learning is that infrared small target lacks shape features and remarkable texture, which makes feature learning difficult.

Although the above single-frame methods have achieved good results, it is very necessary to consider the temporal information because single-frame image lacks sufficient information to detect the small target. Therefore, many sequential small target detection methods based on spatial-temporal information have been proposed. For example, Sun et al. [31] introduced spatial-temporal TV regularization and weighted tensor nuclear norm into the existing IPT model (STTVWNIPT). Liu et al. [32] achieved infrared video detection via a novel spatial-temporal tensor model (IVSTTM). Further, inspired by [18], Sun et al. [33] introduced the multiple subspace learning strategy into the existing IPT model (MSLSTIPT) to improve its robustness in complex scenes. Considering the importance of spatial-temporal information, this paper mainly focuses on infrared sequential small target detection. Although the existing methods have achieved relatively good results, there are still several problems to be solved.

First, due to the limitation of nuclear norm minimization (NNM), IVSTTM method will lead to an over-shrinkage problem. To solve this problem, the weighted nuclear norm minimization (WNNM) is introduced [31]. In fact, the introduction of the WNNM method can only slightly alleviate the problem of over-shrinkage. Further, Sun et al. [33], [34] introduced weighted Schatten p-norm minimization (WSNM) to obtain more accurate background estimation. In summary, the first problem is how to obtain accurate background estimation in complex background. Second, noise is important interference in real scene, which may lead to false alarm in target detection. In recent years, TV regularization is widely exploited in infrared small target detection. However, the classic TV only considers the spatial information, and its computational complexity is high. Although STTV exploits spatial-temporal information, but it treats spatial and temporal information equally, which will degrade the detection performance.

To solve the above problems, we propose a non-convex tensor low-rank approximation method, which combines NTLA and ASTTV regularization. Firstly, considering the importance of accurate background estimation, the NTLA regularization is introduced. Compared with the $l_1$ norm, NTLA regularization can more closely approximate to the $l_0$ norm through Laplace function (see Fig. 1). Meanwhile, it can adaptively assign weights to each singular value. Therefore, it is helpful to obtain more accurate background estimation. Additionally, considering that target is temporally consistant among successive frames and spatially smooth in local area. We exploit the ASTTV to thoroughly describe background feature, which helps obtain more accurate background estimation in complex scenes. Compared with STTV regularization, ASTTV regularization assigns different smoothness strength to adjacent frames. Therefore, the ASTTV regularization constraint on the background helps to better utilize the spatial-temporal information and detect the target more flexibly. In addition, considering the heavy noise in real scenes, we introduce Frobenius norm into the model. Finally, asymmetric spatial-temporal total variation regularized non-convex tensor low-rank approximation is proposed, named the ASTTV-NTLA model. Fig. 2 shows the framework of the ASTTV-NTLA method. The main contributions of our method is summarized as follows.

1) We propose a non-convex tensor low-rank approximation method for sequential infrared small target detection. Different from existing low-rank methods, NTLA regularization adaptively assigns different weights to all singular values through Laplace function, which helps to obtain an accurate background estimation.

2) To capture both spatial and temporal information, we introduce ASSTV regularization into the LRSD model. Furthermore, the ASTTV constraint on the background helps to preserve the details of the image and remove noise. Therefore, it can achieve better performance in complex background scenes.

3) We integrate NTLA regularization, ASTTV regularization and Frobenius norm for infrared small target detection, and develop an algorithm to solve the ASTTV-NTLA model. The experimental results show that the proposed method can achieve promising detection performance in various scenes.

The rest of sections this paper is organized as follows. Section II describes research work in related fields. We summarize some notations and preliminaries in Section III. In Section IV, the ASTTV-NTLA is proposed, and its optimization procedure is designed. Section V shows the experimental results and performance evaluation along with discussion and analyses. Finally, conclusion is given in Section VI.

II. NOTATIONS AND PRELIMINARIES

In this paper, we use lowercase letters (e.g., $x$), boldface lowercase letters (e.g., $x$) and boldface capital letters (e.g., $X$) represent scalars, vectors and matrices, respectively. Considering that tensors are multi-index arrays, we use Euler script (e.g., $A^\epsilon$) to represent them. Readers can refer to [35]–[39] for more details about TNN and t-SVD.

A. Adaptive Thresholding Using Laplace Function

The existing methods generally use TNN to capture the low-rank features of background. Moreover, the operation step of t-SVD is to calculate matrix SVDs of the frontal slices in...
the Fourier domain. The t-SVD [39] operation on the tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is expressed as follows:

$$\mathcal{X} = \mathcal{U} \ast \Sigma \ast \mathcal{V}^T,$$

(1)

where $\mathcal{U}$ and $\mathcal{V}$ represent orthogonal tensors, and $\Sigma$ represents the f-diagonal tensor. The size of $\mathcal{U}$, $\mathcal{V}$ and $\Sigma$ are $n_1 \times n_1 \times n_3$, $n_2 \times n_2 \times n_3$ and $n_1 \times n_2 \times n_3$, respectively. $\ast$ represents the t-product. Small singular values correspond to other sparse disturbances or noise, which can be removed by setting appropriate thresholds. Then, we can use the remaining larger singular values to reconstruct a low-rank tensor. First, we need to solve the Fourier transform for the third dimension of $\mathcal{X}$. Then, we assume that $\mathcal{X}$ is the result of the Fourier transform along the third dimension. Now, the multirank is a vector whose $k$th component gives the rank of the $k$th frontal slice, such as $\text{rank}(\mathcal{X}) = (\text{rank}(\bar{X}_1), \text{rank}(\bar{X}_2), \ldots, \text{rank}(\bar{X}_n))$, where the $k$th frontal slice is denoted as $X^k$ [40], [41]. The sum of the singular values of all the frontal slices (i.e., TNN) is expressed as

$$\|\mathcal{X}\|_s = \frac{1}{n_3} \sum_{k=1}^{n_3} \|\bar{X}_k\|_s.$$  

(2)

It is well known that TNN is a surrogate for tensor multirank [42]. The main shortcoming of TNN is that different singular values have the same importance. However, singular values of natural images have clear physical meaning and should be treated differently. To handle the above problems, Xu et al. [43] introduced the Laplacian function in TNN, which can automatically assign weights according to the importance of singular values. It is defined as follows:

$$\|\mathcal{X}\|_{\gamma,s} = \frac{1}{n_3} \sum_{k=1}^{n_3} \sum_{i,j,k} \phi(\sigma_i(\bar{X}_k)),$$

(3)

where $\epsilon$ is a positive constant and $\sigma_i(\cdot)$ represents the $i$th singular value. Laplacian function is represented by $\phi(x) = 1 - e^{-x/\epsilon}$. Compared with the $l_1$ norm, the Laplace function can better approximate the $l_0$ norm (see Fig.1). For the following optimization problems:

$$\arg \min_Z \|Z\|_{\gamma,s} + \frac{\eta}{2} \|Z - Q\|^2_F,$$

(4)

the global optimal solution is $Q = \mathcal{U} \ast \Sigma \ast \mathcal{V}^H$. Then, the adaptive singular value threshold processing method is adopted for $\mathcal{Z}$, which is expressed as follows

$$\tilde{Z} = \mathcal{U} \ast D_{\mathcal{V}_2} \ast \mathcal{V}^H,$$

(5)

where $D_{\mathcal{V}_2} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ denotes f-diagonal tensor. In Fourier domain, each frontal slice of $D_{\mathcal{V}_2}$ is $D_{\mathcal{V}_2} \in \mathbb{R}^{n_1 \times n_2 \times n_3} = \left(\tilde{S}(i,j,k) - \nabla \phi(\sigma_i(\cdot))/\rho\right)$. The gradient of $\phi$ in $\sigma_i(\cdot)$ is $\nabla \phi(\sigma_i(\cdot)) = 1/2 \exp(-\sigma_i(\cdot)/\rho)$. And the $i$th singular value of the $k$th frontal slice of $\Sigma$ at the $l$th previous iteration.

Algorithm1 briefly describes each iterative solution of the optimization problem in Eq. (4).

B. Asymmetric spatial-temporal total variation regularization

In recent years, TV regularization is widely exploited in infrared small target detection, such as [20], [44], [45] because of its good performance in preserving the spatial piecewise smoothness, edge structure and spatial sparsity of the images. However, existing methods are based on matrix framework and can only describe the spatial continuity of small targets, but ignore their temporal continuity. In addition, it is time-consuming to calculate SVD and TV regularization. Since target is temporally consistent among successive frames and spatially smooth in local area. Considering the importance of spatial-temporal information, Sun et al. [34] introduced STTV into the existing IPT model. The remarkable performance of STTV-WNIPT demonstrate the effectiveness of simultaneously using spatial-temporal information. To model spatial and temporal continuity, we propose an asymmetric spatial-temporal total variation (ASTTV) regularization approach in the tensor framework [46]. The temporal coherence and the spatial-temporal smoothness of small targets are explored. There are two reasons for choosing the ASTTV regularization term. First, imposing ASTTV constraints to the background helps preserve the details of the image and remove noise [47]. Second, compared with STTV-WNIPT method, the ASTTV-NTLA method introduces the parameter $\delta$ to give different weights to temporal TV and spatial TV. Therefore, ASTTV regularization can detect targets more flexibly. The formulation of ASTTV regularization can be expressed as follows:

$$\|\mathcal{X}\|_{\text{ASTTV}} = \|D_h\mathcal{X}\|_1 + \|D_v\mathcal{X}\|_1 + \delta\|D_z\mathcal{X}\|_1,$$

(6)

where $D_h$, $D_v$ and $D_z$ represent the horizontal, vertical and temporal difference operators, respectively. $\delta$ denotes a positive constant, which is used to control the temporal dimension contribution. The ASTTV in Eq. (6) encourages both spatial and temporal smoothness. The three operators of ASTTV regularization are defined as:

$$D_h\mathcal{X}(i,j,k) = \mathcal{X}(i+1,j,k) - \mathcal{X}(i,j,k)$$

(7)

$$D_v\mathcal{X}(i,j,k) = \mathcal{X}(i,j+1,k) - \mathcal{X}(i,j,k)$$

(8)

$$D_z\mathcal{X}(i,j,k) = \mathcal{X}(i,j,k+1) - \mathcal{X}(i,j,k).$$

(9)
The image shows a page of a document discussing a method for processing infrared images. The content includes mathematical equations, algorithms, and figures illustrating the process. The document describes the construction of a spatial-temporal patch tensor model and its application in background subtraction. It mentions the use of alternating direction method of multipliers (ADMM) for solving the equations. The proposed model aims to improve target detection by considering spatial and temporal relationships. The document includes a figure showing the overall framework of the proposed method, with steps outlined in an algorithm. Key points include the exploitation of data from multiple views and the use of low-rank tensor theory to separate background from target images.
regularization. The main reasons are that the NTLa regularization ensures automatic weight assignment to the singular values, which helps more accurate background evaluation. Furthermore, ASTTV is more flexible as compared with STTV for target detection.

C. Optimization Procedure

The optimization Eq. (13) can be solved effectively by using the ADMM [50] approach. By introducing four auxiliary variables $Z, V_1, V_2, V_3, V_3 = D_2 B$. Eq. (13) is rewritten as:

$$
B, T, N = \arg \min_{B,T,N} \frac{1}{2} \|Z\|_{\gamma,*} + \lambda_s \|T\|_1 + \lambda_3 \|N\|_F^2
+ \lambda_{tv} (\|V_1\|_1 + \|V_2\|_1 + \delta \|V_3\|_1)
+ \langle y_1, B - T - N \rangle + \langle y_2, Z - B \rangle
+ \langle y_3, V_1 - D_1 B \rangle + \langle y_4, V_2 - D_2 B \rangle
+ \langle y_5, V_3 - D_3 B \rangle
+ \frac{\mu}{2} \left( \|D - B - T - N\|_F^2 + \|Z - B\|_F^2 + \|V_1 - D_1 B\|_F^2
+ \|V_2 - D_2 B\|_F^2 + \|V_3 - D_3 B\|_F^2 \right)
+ \lambda_3 \|N\|_F^2,
$$

where $\mu$ represents a positive penalty scalar and $y_1, y_2, y_3, y_4, y_5$ represent the Lagrangian multiplier. Eq. (15) is decomposed into five optimization subproblems by ADMM algorithm, including $Z$, $B$, $T$, $V_1$, $V_2$, $V_3$. Since it is difficult to optimize all of these variables in Eq. (15) at the same time, we can alternately update the variables as:

1) Updating $Z$ with other variables being fixed:

$$
Z^{k+1} = \arg \min_{Z} \frac{1}{2} \|Z\|_{\gamma,*} + \frac{\mu}{2} \|Z - B^{k} + \frac{y_k}{\mu k}\|_F^2.
$$

Let $B^{k} - \frac{y_k}{\mu k} = U \ast S \ast V^H$, then the optimal solution can be obtained by Eq. (5). Therefore, the solution of Eq. (16) is:

$$
Z^{k+1} = U \ast D_{S^*} \ast V^H,
$$

where $D_{S^*} = \left( \mathcal{S}(i,j,k) - \sum_{\rho} \sigma(i^\rho) \right)$. The detailed solving process of Eq. (16) is shown in Algorithm 1.

2) Updating $B$ with other variables being fixed:

$$
B^{k+1} = \frac{1}{T} \left( \|D - B - T^{k} - N^{k} + \frac{y_k}{\mu k}\|_F^2
+ \|Z^{k+1} - B + \frac{y_k}{\mu k}\|_F^2 + \|V_1 - D_1 B^{k+1} + \frac{y_k}{\mu k}\|_F^2
+ \|V_2 - D_2 B^{k+1} + \frac{y_k}{\mu k}\|_F^2 + \|V_3 - D_3 B^{k+1} + \frac{y_k}{\mu k}\|_F^2 \right).
$$

The solution to Eq. (18) is equivalent to the following linear system of equations:

$$
(2I + \Delta) B^{k+1} = L^k + \theta_1 + \theta_2 + \theta_3,
$$

where $\Delta = D_h D_h + D_T D_T + D_N D_N, L^k = D - T^{k} - N^{k} + \frac{y_k}{\mu k} + \frac{y_k}{\mu k} + \frac{y_k}{\mu k}, \quad \theta_1 = D_h \left( V_1 + \frac{y_k}{\mu k} \right), \quad \theta_2 = D_T \left( V_2 + \frac{y_k}{\mu k} \right), \quad \theta_3 = D_N \left( V_3 + \frac{y_k}{\mu k} \right)$, and $T$ is the matrix transpose. By considering $D_h B$, $D_T B$, and $D_N B$ as convolutions along two spatial directions and one temporal direction, the closed form solution of Eq. (19) is obtained by nFFT as follows:

$$
B^{k+1} = F^{-1} \left( \frac{F \left( L^k + \theta_1 + \theta_2 + \theta_3 \right)}{2 + \sum_{i \in \{h, v, z\}} F (D_i) F (D_i)} \right),
$$

where $H, F$ and $F^{-1}$ denote the complex conjugate, the fast nFFT operator and the inverse nFFT operator, respectively.

3) Updating $T$ with other variables being fixed:

$$
T^{k+1} = \arg \min_{T} \lambda_s \|T\|_1 + \frac{\mu}{2} \left( \|D - B^{k+1} - T - N^{k} + \frac{y_k}{\mu k}\|_F^2 \right)
$$

The similar element-wise shrinkage operation approach in [52] is used to solve Eq. (21):

$$
T^{k+1} = T h_{\lambda_s (\mu_k)^{-1}} \left( \frac{D - B^{k+1} - N^{k} + \frac{y_k}{\mu k}}{2} \right),
$$

where $Th(\cdot)$ denotes the element-wise shrinkage operator.

4) Updating $V_1, V_2, V_3$ with other variables being fixed:

$$
\begin{align*}
V_1^{k+1} &= \arg \min_{V_1} \lambda_{tv} \|V_1\|_1 + \frac{\mu}{2} \left( \|V_1 - D_1 B^{k+1} + \frac{y_k}{\mu k}\|_F^2 \right)
V_2^{k+1} &= \arg \min_{V_2} \lambda_{tv} \|V_2\|_1 + \frac{\mu}{2} \left( \|V_2 - D_2 B^{k+1} + \frac{y_k}{\mu k}\|_F^2 \right)
V_3^{k+1} &= \arg \min_{V_3} \lambda_{tv} \|V_3\|_1 + \frac{\mu}{2} \left( \|V_3 - D_3 B^{k+1} + \frac{y_k}{\mu k}\|_F^2 \right).
\end{align*}
$$

Figure 3. Singular values of unfolding matrices.
The above problem can also be solved by element-wise shrinkage operator:
\[
\begin{aligned}
V_1^{k+1} &= T_{h_{\lambda_3(y_k)}}^{-1} \left(D_n B^{k+1} - \frac{y_k}{\mu_k^2}\right) \\
V_2^{k+1} &= T_{h_{\lambda_3(y_k)}}^{-1} \left(D_n B^{k+1} - \frac{y_k}{\mu_k^2}\right) \\
V_3^{k+1} &= T_{h_{\lambda_3(y_k)}}^{-1} \left(D_n B^{k+1} - \frac{y_k}{\mu_k^2}\right).
\end{aligned}
\]

(24)

5) Updating $\mathcal{N}^{k+1}$ with other variables being fixed:
\[
\begin{aligned}
\mathcal{N}^{k+1} &= \arg\min_{\mathcal{N}} \lambda_3 ||\mathcal{N}||^2_F \\
&+ \frac{\lambda_3}{2} \|D - B^{k+1} - T^{k+1} - \mathcal{N}^k + \frac{y_k}{\mu_k^2}\|^2_F.
\end{aligned}
\]

(25)

The solution of Eq.(25) is expressed as follows:
\[
\mathcal{N}^{k+1} = \mu_k \left(D - B^{k+1} - T^{k+1}\right) + \frac{\lambda_3}{\mu_k^2} + 2 \lambda_3
\]

(26)

6) Updating multipliers $y_1, y_2, y_3, y_4, y_5$ with other variables being fixed:
\[
\begin{aligned}
y_1^{k+1} &= y_1^k + \mu_k \left(D - B^{k+1} - T^{k+1} - \mathcal{N}^{k+1}\right) \\
y_2^{k+1} &= y_2^k + \mu_k \left(2Z^{k+1} - B^{k+1}\right) \\
y_3^{k+1} &= y_3^k + \mu_k \left(V_1^{k+1} - D_n B^{k+1}\right) \\
y_4^{k+1} &= y_4^k + \mu_k \left(V_2^{k+1} - D_n B^{k+1}\right) \\
y_5^{k+1} &= y_5^k + \mu_k \left(V_3^{k+1} - D_n B^{k+1}\right).
\end{aligned}
\]

(27)

7) Updating $\mu_k^{k+1}$ by $\mu_k^{k+1} = \min \left(\mu_k^k, \mu_{\max}\right)$.

Finally, the proposed ASTTV-NTLA method is summarized in Algorithm 2.

Algorithm 2: ASTTV-NTLA Algorithm

Input: infrared image sequence $d_1, \ldots, d_p \in \mathbb{R}^{n_1 \times n_2}$, number of frames $L$, parameters $\lambda_1, \lambda_2, \lambda_3, \mu > 0$

Initialize: Transform the image sequence into the original tensor $D, B^0 = T^0 = N^0 = V^0 = 0, i = 1, 2, 3, y_i^0 = 0, i = 1, \ldots, 5, \mu_0 = 1e-2, \mu_{\max} = 1e7, k = 0, \rho = 1.5, \zeta = 1e-6$

While: not converged do
1: Update $Z^{k+1}$ by Algorithm 1
2: Update $B^{k+1}$ by Eq.(20)
3: Update $T^{k+1}$ by Eq.(22)
4: Update $V_1^{k+1}, V_2^{k+1}, V_3^{k+1}$ by Eq.(24)
5: Update $\mathcal{N}^{k+1}$ by Eq.(26)
6: Update multipliers $y_i^{k+1}, i = 1, \ldots, 5$ by Eq.(27)
7: Update $\mu_k^{k+1}$ by $\mu_k^{k+1} = \min \left(\mu_k^k, \mu_{\max}\right)$
8: Check the convergence conditions $\frac{\|D - B^{k+1} - T^{k+1} - \mathcal{N}^{k+1}\|^2_F}{\|D - B^{k+1}\|^2_F} \leq \zeta$
9: Update $k = k + 1$
end While

Output: $B^{k+1}, T^{k+1}, \mathcal{N}^{k+1}$

D. Target Detection procedure

The flow chart of the proposed ASTTV-NTLA method is shown in Fig. 2. Next, we will describe each step in detail.

1) Patch-tensor construction. By stacking $n_3$ adjacent frames in chronological order, the original infrared image sequence $d_1, d_2, \ldots, d_p \in \mathbb{R}^{n_1 \times n_2}$ transform into several patch-tensors $D \in \mathbb{R}^{n_1 \times n_2 \times n_3}$.

2) Background and target separation. The original patch-tensor $D$ is decomposed into three parts by Algorithm 2, which are target patch-tensor $T$, noise patch-tensor $\mathcal{N}$ and background patch-tensor $B$.

3) Image reconstruction. The target image $f_T$ and background image $f_B$ can be reconstructed by simple inverse operation.

4) Target detection. Due to the high pixel value of the true target in the reconstructed target image, we exploit the adaptive threshold segmentation approach in [13] to extract small target.

\[
t_{th} = \max \left(v_{\min}, \mu + k\sigma\right),
\]

(28)

where $\sigma$ and $\mu$ are the standard deviation and mean of the $f_T$, respectively. $k$ is a constant determined experimentally. $v_{\min} = 0.85$ is an adaptive value. If $f_T(x, y) > t_{th}$, then the pixel at $(x, y)$ is considered as the target.

E. Complexity analyses

In this subsection, we concisely analyze the computational complexity of the ASTTV-NTLA method. For the input image sequence $d_1, d_2, \ldots, d_p \in \mathbb{R}^{n_1 \times n_2}$, we can obtain $s = p/n_3$, and the dimension of each tensor is $D \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. In ASTTV-NTLA algorithm, the main cost is to update $Z$ and $B$. Updating $Z$ requires performing FFT and $\lceil n_3/2 \rceil$ SVDs of $n_1 \times n_2$ matrices in each iteration by t-SVT, which cost $O\left(ksn_1n_2n_3 \log n_3 + (n_2 \lfloor n_3/2 \rfloor / n_3)\right)$.

Updating $B$ requires performing FFT operation, which cost...
Figure 6. Ablation experiments of each regularization of ASTTV-NTLA method. Row 1 and 3: the noisy images. Row 2 and 4: the corresponding results. Results in (a) and (b) are achieved by our model without Frobenius norm. Results in (c) and (d) are achieved by our model without ASTTV regularization. Results in (e) and (f) are achieved by our model. The blue ellipses denote noise and background residuals.

\[ O(ksn_1n_2n_3 \log (n_3)) \]  The \( k \) denotes the iteration times. In summary, the computational complexity of each iteration is

\[ O(ksn_1n_2n_3 \left(2 \log n_3 + n_2 \left[\frac{n_3+1}{2}\right]/n_3\right)) \]

F. Convergence analysis

In this section, we analyze the convergence of the proposed ASTTV-NTLA method. Due to the existence of NTLA regularization, the solving process of Eq.(12) is actually a nonconvex optimization problem. In order to solve this problem, we use Theorem 3 in [53]. In our method, we use an empirical convergence condition

\[ \left|P - \left(T^{k+1} - \Lambda^{k+1}\right)\right|_F^2 \leq \zeta \]  to analyze the convergence. Fig. 7 shows the convergence curve on Sequence 2. It can be seen from Fig. 7 that the value of the objective function converges to zero when \( k \geq 60 \).

IV. EXPERIMENTAL RESULTS AND ANALYSES

This section is mainly composed of the following parts. First, it introduces four relevant evaluation metrics and ten baseline methods, and then analyzes several important parameters in the ASTTV-NTLA method. Finally, the advantages of the ASTTV-NTLA method and the baseline methods are compared in various scenes.

A. Evaluation Metrics and Baseline Methods

To quantitatively validate the detection performance of the ASTTV-NTLA method, we use four evaluation metrics including the background suppression factor (BSF), local signal to noise ratio gain (LSNRG), contrast gain (CG) and signal to clutter ratio gain (SCRG). They are used to evaluate detection performance and background suppression ability. LSNRG is usually used to describe the local signal to noise ratio (LSNR) gain as follows

\[ \text{LSNRG} = \frac{\text{LSNR}_{\text{out}}}{\text{LSNR}_{\text{in}}} \]  (29)

where \( \text{LSNR}_{\text{out}} \) and \( \text{LSNR}_{\text{in}} \) represent LSNR values before and after processing respectively. LSNR = \( P_T/P_B \). The maximum pixel values of the neighborhood and the target are

\[ P_T = \max \{T(x,y) \} \]  and

\[ P_B = \max \{B(x,y) \} \]
represented by $P_{3}$ and $P_{1}$, respectively. Then, BSF compares the background suppression ability, which is expressed as:

$$\text{BSF} = \frac{\sigma_{\text{in}}}{\sigma_{\text{out}}},$$  \hspace{1cm} (30)$$

where $\sigma_{\text{out}}$ and $\sigma_{\text{in}}$ are the standard variance of the neighboring background region of target image and original image, respectively. The SCRG denotes the signal-to-clutter ratio (SCR) before and after processing, which is defined as:

$$\text{SCRG} = \frac{\text{SCR}_{\text{out}}}{\text{SCR}_{\text{in}}},$$ \hspace{1cm} (31)$$

where SCR uses the same expression as [54]:

$$\text{SCR} = \frac{|\mu_{t} - \mu_{b}|}{\sigma_{b}},$$ \hspace{1cm} (32)$$

where $\mu_{t}$, $\mu_{b}$ and $\sigma_{b}$ represent the average value of the target area, the pixel average value and standard deviation of the surrounding local neighborhood region, respectively. It can be seen from Fig. 9 that $d$ and $a \times b$ denote the size of the width of the adjacent area and the target region, respectively. Meanwhile, $(a + 2d) \times (b + 2d)$ represents the size of the neighboring area. In our paper, $d = 40$, $a = b = 9$. To compare the ability of gray difference between expanded target and background, we introduce CG evaluation metric [48], which is expressed as:

$$\text{CG} = \frac{\text{CON}_{\text{out}}}{\text{CON}_{\text{in}}},$$ \hspace{1cm} (33)$$

where $\text{CON}_{\text{out}}$ and $\text{CON}_{\text{in}}$ denote the contrast (CON) of the target and original images, respectively, and CON is defined as:

$$\text{CON} = |\mu_{t} - \mu_{b}|.$$

(34)$$

Generally speaking, the higher values of the above four evaluation metrics indicate that the method has better background suppression ability. It is worth noting that LSNRG, BSF and SCRG are the evaluation metrics to describe the local neighborhood suppression ability, not the global suppression ability. In addition, detection probability $P_{d}$ and false-alarm rate $F_{a}$ are also important evaluation indicators, which are defined as [13]:

$$P_{d} = \frac{\text{number of true detections}}{\text{number of actual targets}},$$ \hspace{1cm} (35)$$

$$F_{a} = \frac{\text{number of false detections}}{\text{number of image pixels}}.$$ \hspace{1cm} (36)$$

The above two indicators range between 0 and 1.

In order to estimate the advantages of the ASTTV-NTLA method, we compare it with ten other methods in the above metrics and various scenes. These methods are mainly divided into the following categories: BS-based methods (Top-Hat [5]), HVS-based methods (WSLCM [11]), and recently developed LRSD-based methods (NRAM [16], TV-PCP [20], RIPT [21], PSTNN [23], STTV-WNIPT [34], IVSTTM [32], MSLSTIPT [33], NTFRA [26]). Among the above methods, the single-frame method using only spatial information includes Top-Hat, WSLCM, NRAM, TV-PCP, RIPT, PSTNN and NTFRA. Because the proposed ASTTV-NTLA method uses spatial-temporal information, we compare our method with three
Figure 10. ROC curves with respect to different $H$.

Table I
DETAILED PARAMETER SETTING FOR TESTED METHODS.

| Methods                                      | Acronyms                        | Parameter settings       |
|----------------------------------------------|----------------------------------|--------------------------|
| Top-Hat method                               | Top-Hat                          | Structure size: $3 \times 3$, structure shape: square |
| Weighted strengthened local contrast measure | WSLCM                            | Sliding step: 10, $\lambda = \frac{1}{\sqrt{M \times N}}$, patch size: $50 \times 50$, $\gamma = 0.002$, $L = 10$ |
| Non-Convex Rank Approximation Minimization   | NRAM                             | $C = \sqrt{\min (M, N)} \times 2.5$, $\mu = 3\sqrt{\min (M, N)}$, $\epsilon = 1e - 7$ |
| Total Variation Regularization and Principal Component Pursuit | TV-PCP                           | $\lambda_1 = 0.005$, $\lambda_2 = \frac{H}{\sqrt{\min (M, N)^2}}$, $\beta = 0.025$, $\gamma = 1.5$ |
| Reweighted Infrared Patch-Tensor Model       | RIPT                             | Sliding step: 10, $\lambda = \frac{1}{\sqrt{\min (n_1, n_2, n_3)}}$, patch size: $50 \times 50$, $h = 10$, $e = 1 \times 10^{-7}$, $L = 1$ |
| Partial Sum of the Tensor Nuclear Norm       | PSTNN                            | Sliding step: 40, $\lambda = \frac{1}{\sqrt{\min (n_1, n_2, n_3)}}$, patch size: $40 \times 40$, $e = 1 \times 10^{-7}$ |
| Spatial-temporal Total Variation Regularization and weighted Tensor Nuclear Norm | STTV-WNIPT                       | $L = 3$, $H = 8$, $\lambda_1 = 0.005$, $\lambda_2 = \frac{H}{\sqrt{\min (M, N)^2}}$, $\lambda_3 = 100$ |
| Infrared Videos Based on Spatio-Temporal Tensor Model | IVSTTM                           | Sliding step: 15, patch size: $80 \times 80$, spatial patch cube: $n_x = 9$ |
| Multiple Subspace Learning and Spatial-temporal Patch-Tensor Model | MLSSTIPT                         | $L = 6$, $p = 0.8$, $\lambda = 1/\sqrt{\min (n_1, n_2)}$, patch size: $30 \times 30$ |
| Nonconvex tensor fibered rank approximation | NTFRA                            | Sliding step: 40, patch size: $40 \times 40$, $H = 1/\sqrt{\min (n_1, n_2)}$, $\beta = 0.01$ |
| Non-Convex Tensor Low-Rank Approximation     | ASTTV-NTLA                       | $L = 3$, $H = 6$, $\lambda_{tv} = 0.005$, $\lambda_3 = \frac{H}{\sqrt{\min (M, N)^2}}$, $\lambda_3 = 100$ |

methods using spatio-temporal information, namely STTV-WNIPT, IVSTTM, MLSSTIPT. Table I shows the detailed parameter settings of the comparison method in this paper. All experiments were conducted on a computer with a 16 GB RAM and an Inter Core i7-10870H CPU (2.20 GHz). Our method was implemented in MATLAB 2014a, and the codes are available at https://github.com/LiuTing20a/ASTTV-NTLA.

**B. Parameter Setting and Datasets**

In our model, we test on a synthetic data to determine the values of several important parameters. The regularized parameter $\lambda_{tv}$ could balance the tradeoff between the non-convex tensor low-rank approximation and ASTTV regularization, and it is empirically set to 0.005 following [46]. In most cases, the range of $\delta$ is $[0, 1]$, we follow [47] to set $\delta = 0.5$. Following [55], we set $\lambda_s = \frac{H}{\sqrt{\max (m, n) \times L}}$. $H$ represents a tuning parameter. According to [49], we set $\lambda_3 = 100$. In the following experiments, we further analyzed the influence of the parameters $L$, $H$ and $\delta$ on the detection performance. Section IV-D explains more details.

As can be seen from Table II, we simulated six image sequences come from diverse scenes to demonstrate the stability and detection ability of our method. In our datasets, the sky background with some clouds in Sequences 1 and 2 are relatively simple. In Sequence 3, the small target is moving in a blurred sealand background. The main challenges of Sequence 3 are the bright artificial buildings in the background. In Sequence 4, an airplane is flying towards the mountains with strong reflections at the bottom of the mountains. The main challenges of Sequence 4 are the high reflection intensity, which easily leads to miss detection and false alarm. The main challenges of Sequences 5 and 6 are the reflective road, roof and forests in the background. Therefore, our datasets consists of both simple and complex background. The diverse datasets
with various scenes can help to comprehensively evaluate the performance of each algorithm. To obtain synthetic data, we use the approach in [13] to add the targets on six real background data. Figs. 12-13 shows the 2D gray distribution of the representative image.

C. Validation of the proposed ASTTV-NTLA method

In this subsection, we validate the robustness of the ASTTV-NTLA method in various scenes.

1) Robustness to single targets scene: First, we test the ASTTV-NTLA method on three real single target scenes. The first and second row of Fig. 4 show the representative images and the related target images, respectively. For better visualization, we adopt red rectangular box to mark the target in the result image. The results in Fig. 4 demonstrate that each target is detected successfully and the background clutters are suppressed perfectly.

2) Robustness to multiple targets scene: In a variety of real scenes, the number of targets of interest is different. Therefore, we test the robustness of ASTTV-NTLA method in multi-targets scenario (actually 3). It is worth noting that we adopt the approach in [13] to synthesize multi-targets scene. The experimental results in the second row of Fig. 5 show that the background noise and clutter are well suppressed.

3) Robustness to noisy scene: In real scenes, noise is another crucial factor that impacts the performance of background suppression and target detection. Therefore, we tested the ASTTV-NTLA method in different noise scenes. In our method, ASTTV regularization can not only make full use of spatial-temporal information, but also remove noise and preserve image details. The robustness of ASTTV-NTLA method to noisy scenes is tested on scenes with noise of $\sigma = 15$ and $\sigma = 25$. To validate the effectiveness of ASTTV regularization, we compare our model with the model without ASTTV. Fig. 6 (c)-(d) show the experimental results of the model without ASTTV regularization. Fig. 6 (e)-(f) show the experimental results of our model. As can be seen from the second and fourth rows of Fig. 6 (c)-(d), the detection results in non-smooth and non-uniform scenes have noise and background edge residue. It can be seen from Fig. 6 (e)-(f) that the proposed ASTTV-NTLA method can obtain good detection results in non-smooth and non-uniform scenes and remove noise at the same time. Therefore, introducing ASTTV regularization into the model can better solve the problem of small target detection in complex noisy scenes. Meanwhile, to demonstrate the effectiveness of Frobenius norm, we compare the performance of our model and the model without Frobenius norm. From the second row of Fig. 6 (a)-(b), it can be seen that in the noisy scene with $\sigma = 15$, the
Figure 12. Comparative results achieved by different methods on Sequence 1-3 (without segmentation). The blue ellipses denote noise and background residuals. For more intuitive visualization, the demarcated target area by the red rectangle is zoomed in the left bottom corner.

result image has only a little noise residue. As can be seen from the fourth row of Fig. 6 (a)-(b), the detection results in heavy noisy scenes have many noise residues. As can be seen from Fig. 6 (e)-(f), the proposed ASTTV-NTLA method can obtain good detection results in heavy noisy scenes. It demonstrates that the introduction of Frobenius norm into the model is helpful to better suppress noise, especially in heavy noisy scenes. In summary, the experimental results in Fig. 6 show that the proposed ASTTV-NTLA method can suppress noise well, which validates the robustness of the proposed ASTTV-NTLA method to different noisy scenes.

D. Parameter Analysis

In this subsection, we analyze the impact of number of frame $L$, tuning parameter $H$, and parameter $\delta$ on the performance of the method.

1) Number of frames: We use the temporal information via introducing ASTTV regularization. Note that the $L$ is an important parameter. We change $L$ from 2 to 6 with a step of 1. Fig. 8 shows the relevant ROC curve. The experimental results in Fig. 8 show that $L = 3$ has the best performance. It is worth noting that if the $L$ value is set smaller, the detection probability will be decrease. At the same time, Figs. 8 (d) and (f) show that an over-large $L$ will decrease the
detection probability. The main reason is that too large $L$ leads to the failure of low-rank assumption. In the following experiments, we set $L = 3$, which can maintain a balance between effectiveness and performance, so as to achieve good experimental results.

2) Tuning parameter: $H$ plays a key role in the optimization of the ASTTV-NTLA model. We change $H$ from 2 to 10 with a step of 2, and Fig. 10 shows the relevant ROC curve. The ROC curves of $H = 10$ in Figs. 10 (a) and (b) demonstrate that an over-large $H$ will decrease the detection probability. Meanwhile, from the results of $H = 2$ and $H = 4$ in Figs. 10 (a), (d) and (f), we can conclude that an over-small $H$ will increase the false alarm rate. Therefore, in the following experiment, we set $H = 6$.

3) Parameter $\delta$: $\delta$ is a crucial parameter related to temporal information. It indicates that the temporal difference in the ASTTV regularization helps to improve the performance of the ASTTV-NTLA method. We vary $\delta$ from 0 to 1 with a step of 0.2. Fig. 11 shows the relevant ROC curve. The conclusion that can be drawn from the ROC curves of $\delta = 0$ in Fig. 11 is that if there is no temporal information, the detection probability will decrease. $\delta = 1$ is STTV regularization. As can be seen from Fig. 11, selecting the appropriate $\delta$ value will get better detection performance. Therefore, in the following
experiment, we set $\delta = 0.5$.

E. Comparison to state-of-the-art Methods

To demonstrate the advantages of the ASTTV-NTLA method, we compare it with other ten methods on six different real infrared image scenes. Figs. 12 and 13 show the comparative results of Sequences 1-6. As can be seen from Figs. 12 and 13, Top-hat method produces very rough detection results. Experimental results show that it not only enhances the target, but also enhances the noises and clutters. From the results of sequence 4 and sequence 5 in Fig. 13, it can be seen that the residuals of the reflective mountains and roads still remain in the target image. The main reason is that the size of the Top-hat filter is not suitable for the scenes with strong reflection clutter. As a top-performing HVS method, WSLCM can detect the target more accurately in simple background, but there are still clutter or missed detection in complex background. Compared with the BS and HVS method, matrix-based LRSD methods have less background residual or missed clutter in complex background, such as NRAM, TV-PCP methods. From the highlight scene Sequence 4 and the complex ground scene Sequences 5-6, it can be seen that NRAM method still has a little residual and clutter, but TV-PCP method achieves poor performance on these complex scenes. To handle these problems, tensor-based method is proposed. The results in Figs. 12 and 13 show that RIPT method is more effective in clutter suppression than matrix-based methods. Therefore, many improved methods are proposed, such as PSTNN and NTFRA methods. As can be seen from Fig. 12, these tensor-based methods can suppress clutter in complex background, but some non-target pixels still remain in their target image. The main reason is that there are many clutter in a single-frame complex image background, such as highlight background and ground background, which may seriously affect the detection

![Figure 14. ROC curves achieved by different methods.](image)

### Table III

| Method         | $60^{th}$ frame of Sequence 1 | $100^{th}$ frame of Sequence 2 | $90^{th}$ frame of Sequence 3 |
|----------------|-------------------------------|--------------------------------|-------------------------------|
|                | LSNRG BSF SCRG                 | LSNRG BSF SCRG                  | LSNRG BSF SCRG                |
| Top-hat [5]    | 0.92 0.63 1.42                 | 0.54 0.99 0.43                  | 0.48 0.72 0.24                |
| WSLCM [11]     | 0.90 0.86 4.76                 | 0.93 1.80 4.47                  | 0.77 1.47 6.18                |
| NRAM [16]      | 2.70 7.68 10.71                | 1.19 4.18 2.95                  | 1.28 4.74 4.36                |
| TV-PCP [20]    | 1.41 4.42 6.57                 | 0.97 4.70 3.67                  | 1.47 7.46 8.27                |
| RIPT [21]      | 2.55 6.06 8.25                 | 1.22 6.21 3.53                  | 3.16 16.00 14.73              |
| PSTNN [23]     | 2.74 7.10 77.98                | 0.87 3.76 4.55                  | 1.52 4.47 6.54                |
| STTV-WNIP [34] | 1.62 2.58 57.45                | 1.47 4.27 4.53                  | 1.12 2.86 6.38                |
| IVSTTM [32]    | 2.52 7.18 10.01                | 1.60 8.18 3.58                  | 2.76 19.17 13.73              |
| MSLSTIPT [33]  | 1.86 2.17 72.39                | 1.09 1.44 1.29                  | 1.05 1.52 2.50                |
| NTFRA [26]     | – – –                         | 1.05 1.14 1.13                  | – – –                         |
| ASTTV-NTLA (ours) | 3.05 8.96 97.57              | 3.86 18.16 14.24                | 3.70 21.30 17.19              |
of real targets. The information in a single-frame image is not enough to distinguish small targets. Based on this, many scholars have introduced temporal information and use spatial-temporal information to solve small target detection. As can be seen from Fig. 12-13, compared with those tensor methods that only use spatial information, the spatial-temporal information method (STTV-WNIPT, IVSTTM and MSLSTIPT) can better detect the target and suppress the background. It can be seen from Sequence 3 and Sequence 4 that IVSTTM method has background residue on the target image. The main reason is that NNM treats all singular values equally, which will lead to over-shrinkage problem. To solve this problem, the STTV-WNIPT method introduces WNNM and STTV regularization. It can be seen from Sequence 4 and Sequence 5 that there is background residue in the target image obtained by STTV-WNIPT method, which indicates that WNNM can only alleviate over-shrinkage problem. Further, MSLSTIPT method introduces WSNM to obtain more accurate background estimation. However, it can be seen from Sequence 2 and Sequence 3 that the background suppression effect of MSLSTIPT method is not good. In contrast, ASTTV-NTLA method can detect targets accurately under the premise of better suppression of background and noise. The results demonstrate the advantages of the ASTTV and non-convex tenor low-rank approximately property. Note that the dataset contains a variety of scenes, which shows the superiority and robustness of the ASTTV-NTLA method.

In this paper, we adopt LSNRG, BSF and SCRG as metric for quantitative evaluation, which are widely used by the exiting methods. Table III and Table IV show the results. The best results of the test method are highlighted in bold. The results in the Table show that ASTTV-NTLA method obtains the best results on the indicator values. It shows that the ASTTV-NTLA method not only can effectively suppress noise and clutters, but also can better highlight the targets. It is worth noting that we do not show the indicators values of NTFRA on the 60th frame of Sequence 1 and the 90th frame of Sequence 3 and WSLCM on the 50th frame of Sequence 4. The main reason is that the target is lost in these scenes, so it is meaningless to calculate the indicator value. To deal with the above problem, we use the approach in [48] and introduce the CG metric to ensure the ability of expanding the gray value difference between the background and the target. Table V shows the average CG value of the entire sequence. The results in Table V show that ASTTV-NTLA method obtains the highest average CG value on all test Sequences. This metric demonstrate that the ASTTV-NTLA method has satisfactory background suppression capacity.

Fig. 14 shows the ROC curves of all comparison methods in this paper. The results in Fig. 14 show that ASTTV-NTLA method has best performance. Top-hat method is greatly affected by bright clutters and noise. WSLCM and TLLCM method achieve better detection performance on simple background scenes such as Sequence 1-3, and worse detection performance on highlighted and complex ground background scenes such as Sequence 4-6. The ROC curves in Fig. 14 (d)-(f) show that the performance of NRAM method and TV-PCP method is slightly better than WSLCM and Top-hat method in complex scenes. However, it can be seen from Fig. 14 (b) and Figs. 14 (d)-(f) that the detection results of NRAM and TV-PCP methods are relatively poor. Therefore, tensor-based method (RIPT, PSTNN, NTFRA) are proposed. As can be seen from (d)-(e) in Fig. 14, the detection results of PSTNN in complex scenes are relatively poor. The main reason is that
Table IV
QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON SEQUENCES 4-6.

| Method                          | 50\(^{th}\) frame of Sequence 4 | 10\(^{th}\) frame of Sequence 5 | 90\(^{th}\) frame of Sequence 6 |
|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
|                                | LSNRG  | BSF    | SCRG   | LSNRG  | BSF    | SCRG   | LSNRG  | BSF    | SCRG   |
| Top-hat [5]                    | 0.51   | 0.72   | 0.42   | 0.76   | 1.76   | 0.50   | 0.45   | 1.88   | 0.37   |
| WSLCM [11]                     | –      | –      | –      | 0.12   | 4.47   | 24.87  | 0.77   | 2.29   | 5.84   |
| NRAM [16]                      | 9.56   | 23.74  | 32.79  | 1.33   | 7.08   | 21.12  | 1.34   | 9.31   | 9.54   |
| TV-PCP [20]                    | 1.22   | 3.98   | 6.98   | 1.16   | 9.23   | 19.51  | 1.14   | 6.70   | 6.61   |
| RRIPT [21]                     | 1.27   | 3.28   | 5.74   | 1.40   | 7.90   | 25.19  | 1.53   | 10.25  | 8.70   |
| PSTNN [23]                     | 1.41   | 3.86   | 8.49   | 1.25   | 4.14   | 22.34  | 1.04   | 4.14   | 6.92   |
| STTV-WNIPT [34]                | 1.17   | 2.42   | 6.63   | 1.34   | 4.68   | 21.15  | 1.03   | 4.00   | 7.42   |
| IVSTTM [32]                    | 1.38   | 4.09   | 7.40   | 0.78   | 6.32   | 6.11   | 1.69   | 12.23  | 14.00  |
| MSLSTIPT [33]                  | 1.20   | 2.07   | 6.55   | 1.33   | 7.08   | 21.12  | 1.34   | 9.31   | 9.54   |
| NTFRA [26]                     | 2.84   | 24.69  | 12.15  | 1.10   | 7.01   | 22.28  | 1.00   | 3.46   | 7.21   |
| ASTTV-NTLA (ours)              | 14.32  | 33.84  | 253.34 | 26.70  | 110.82 | 97.21  | 5.12   | 53.07  | 34.04  |

Table V
AVERAGE CG VALUES ACHIEVED BY DIFFERENT METHODS ON SEQUENCES 1-6

| Methods                       | Sequence 1 | Sequence 2 | Sequence 3 | Sequence 4 | Sequence 5 | Sequence 6 |
|-------------------------------|------------|------------|------------|------------|------------|------------|
| Top-hat [5]                   | 2.25       | 1.41       | 1.12       | 3.16       | 1.29       | 1.15       |
| WSLCM [11]                    | 13.03      | 3.16       | 4.20       | 5.85       | 2.36       | 3.46       |
| NRAM [16]                     | 13.16      | 1.03       | 1.80       | 6.38       | 1.57       | 1.92       |
| TV-PCP [20]                   | 10.55      | 1.70       | 1.11       | 7.75       | 1.82       | 1.66       |
| RRIPT [21]                    | 15.43      | 1.69       | 1.92       | 6.72       | 1.58       | 3.11       |
| PSTNN [23]                    | 19.22      | 1.80       | 5.46       | 8.20       | 1.16       | 2.99       |
| STTV-WNIPT [34]               | 22.24      | 1.94       | 7.23       | 8.74       | 1.18       | 5.30       |
| IVSTTM [32]                   | 13.29      | 2.44       | 1.81       | 7.90       | 1.70       | 1.72       |
| MSLSTIPT [33]                 | 33.33      | 2.89       | 2.64       | 22.38      | 2.14       | 4.18       |
| NTFRA [26]                    | 1.38       | 1.02       | 1.52       | 2.56       | 1.42       | 1.64       |
| ASTTV-NTLA (ours)             | 35.54      | 3.57       | 5.64       | 27.49      | 5.88       | 8.30       |

Table VI
RUNNING TIME OF DIFFERENT METHODS

| Methods                       | Sequence 1 | Sequence 2 | Sequence 3 | Sequence 4 | Sequence 5 | Sequence 6 |
|-------------------------------|------------|------------|------------|------------|------------|------------|
| Top-hat [5]                   | 32.52s     | 40.46s     | 33.79s     | 32.73s     | 32.42s     | 31.34s     |
| WSLCM [11]                    | 1087.39s   | 1701.05s   | 1411.60s   | 1450s      | 1451.99s   | 1400.41s   |
| NRAM [16]                     | 80.07s     | 223.71s    | 185.28s    | 153.01s    | 190.28s    | 157.94s    |
| TV-PCP [20]                   | 7976.50s   | 13314.75s  | 11144.09s  | 11339.91s  | 11164.87s  | 11233.05s  |
| RRIPT [21]                    | 91.32s     | 153.95s    | 138.49s    | 112.31s    | 168.66s    | 131.04s    |
| PSTNN [23]                    | 29.97s     | 38.95s     | 34.86s     | 46.16s     | 52.04s     | 45.42s     |
| STTV-WNIPT [34]               | 174.70s    | 293.81s    | 235.94s    | 243.41s    | 245.79s    | 245.77s    |
| IVSTTM [32]                   | 80.37s     | 117.61s    | 106.35s    | 139.16s    | 102.02s    | 115.34s    |
| MSLSTIPT [33]                 | 176.13s    | 275.39s    | 208.98s    | 225.32s    | 181.82s    | 181.63s    |
| NTFRA [26]                    | 129.19s    | 171.97s    | 156.08s    | 351.07s    | 320.56s    | 281.38s    |
| ASTTV-NTLA (ours)             | 206.64s    | 377.35s    | 287.89s    | 284.78s    | 280.17s    | 283.33s    |

PSTNN method assigns same weights for all singular values. As can be seen from Fig. 14 (d)-(f), the detection results of tensor-based method in complex scenes are alleviated to a certain extent. Compared with the methods (NRAM, TV-PCP, RRIPT, PSTNN, NTFRA) that only uses spatial information, the STTV-WNIPT, IVSTTM and MSLSTIPT methods achieve better performance in various scenes because they incorporate spatial-temporal information. It indicates the effectiveness of spatial-temporal information. It can be seen from Fig. 14 that among all the test methods, the Pd of the ASTTV-NTLA method can reach 1 the fastest. This shows the necessity of considering temporal correlation and the effectiveness of integrating NTLA regularization and ASTTV regularization. In summary, all the above evaluation metrics show that ASTTV-NTLA method can achieve satisfactory robust performance in different real scenes, particularly for complex ground scenes and highly heterogeneous scenes.

F. Ablation Experiments
To show the advantages of the ASTTV-NTLA method that integrating the ASTTV regularization and NTLA regularization, we evaluate the contributions of the ASTTV and NTLA, respectively. The ASTTV-NTLA method mainly composes of three parts: the spatial-temporal tensor structure, ASTTV regularization and NTLA regularization. Therefore, the performance of the three versions of ASTTV-NTLA method is compared on Sequences 1-6, including: 1) Adopting the STTV regularization constraint weighted IPT model (STTV-WNIPT); 2) Adopting the ASTTV regularization constraint the weighted IPT model (ASTTV-WNIPT); 3) Adopting the ASTTV regularization constraint the NTLA regularization (ASTTV-NTLA). Fig. 15 shows the ROC curves of STTV-WNIPT, ASTTV-WNIPT and ASTTV-NTLA methods on Sequences 1-6. Moreover, the performance of ASTTV-WNIPT is better than that of STTV-WNIPT, which demonstrates that...
ASTTV regularization can improve the detection ability of the model to a certain extent by using different smoothness strength for temporal TV and spatial TV. The conclusion drawn from Fig. 15 is that the detection results of ASTTV-NTLA is better than that of ASTTV-WNIPT. Compared with WNIPT regularization, NTLA regularization is a better substitute for tensor rank, which can obtain more accurate background estimation and further improve the ability of target detection. In summary, the above experimental results demonstrate that integrating Laplace norm with ASTTV regularization can achieve better target detection performance.

G. Running time

In addition to the above four evaluation metrics, running time is also a crucial factor. However, it is difficult to balance these two factors. Therefore, we compare the efficiency of the comparison method on Sequences 1-6. Table VI shows the running time of all comparison methods in various scenes. Top-hat method is the fastest among all comparison methods, but its detection performance is relatively rough. In constrast, the speed of LRSD methods are relatively slow. Because the matrix-based LRSD method needs many SVD operations. Among them, TV-PCP needs relatively long running time. The main reason is that they adopt accelerated proximal gradient (APG) method for optimization. The methods optimized by ADMM are more efficient, such as NRAM, RAPT, PSTNN, STTV-WNIPT, IVSTTM, MSSTIFT, NTFRA and the proposed method. Among these methods, the speed of PSTNN is only slower than Top-hat. The main reason is that with the help of the additional stop criterion and t-SVD, the computing time and complexity of the algorithm are greatly reduced. The three experimental results based on the TV regularization method in Table VI show that the running time of the ASTTV-NTLA method is slightly slower than the STTV-WNIPT method, but much faster than TV-PCP method. Considering that the ASTTV-NTLA method achieves good performance in various complex scenes, the running time can be slightly sacrificed. As can be seen from Table VI, our method is faster than some matrix-based SVD methods (e.g., TV-PCP), but slower than some t-SVD based methods (e.g., PSTNN, IVSTTM). That is because, ASTTV regularization has high computational complexity. In the future work, we can refer to the strategies in recent works [23], [56], [57] to enhance the real-time performance of our method.

V. CONCLUSION

To improve the capacity of target detection and background suppression in complex noise and strong clutters scenes, an asymmetric spatial-temporal total variation regularized non-convex low-rank tensor approximation method is proposed. The NTLA regularization is used to adaptively assign different weights to all singular values, which helps reconstruct the background image more accurately. In addition, ASTTV regularization can fully utilize the structure prior information to detect targets. Furthermore, ASTTV regularization can effectively exploit spatial-temporal information to suppress the background noise and detect the targets in non-uniform and non-smooth scenes. Extensive experimental results show the promising detection performance of the ASTTV-NTLA method.

However, in our model, we only consider white Gaussian noise, which deviates from the complicated real-world noise model. Therefore, the performance of the ASTTV-NTLA method in complex noisy scenes is limited. Recently, deep convolutional neural networks (CNNs) have achieved impressive success in image denoising and can handle more realistic noise. Therefore, in the future, we will consider to use a learning-based CNN denoiser to handle the denoising subproblem for better detection performance.

VI. ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61972435, Grant 61401474, and Grant 61921001.

REFERENCES

[1] N. T. Thanh, H. Sahli, and D. N. Hao, “Infrared thermography for buried landmine detection: Inverse problem setting,” IEEE Transactions on Geoscience and Remote Sensing, vol. 46, no. 12, pp. 3987–4004, 2008.
[2] S. Bouroumiane, C. Fossati, and A. Caillly, “Improvement of target-detection algorithms based on adaptive three-dimensional filtering,” IEEE Transactions on Geoscience and Remote Sensing, vol. 49, no. 4, pp. 1383–1395, 2010.
[3] X. Bai, Z. Chen, Y. Zhang, Z. Liu, and Y. Lu, “Infrared ship target segmentation based on spatial information improved fcm,” IEEE Transactions on Cybernetics, vol. 46, no. 12, pp. 3259–3271, 2015.
[4] M. M. Hadhoud and D. W. Thomas, “The two-dimensional adaptive lms (tdlms) algorithm,” IEEE Transactions on Circuits and Systems, vol. 35, no. 5, pp. 485–494, 1988.
[5] J.-F. Rivest and R. Fortin, “Detection of dim targets in digital infrared imagery by morphological image processing,” Optical Engineering, vol. 35, no. 7, pp. 1886–1893, 1996.
[6] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, “Max-mean and max-median filters for detection of small targets,” in Signal and Data Processing of Small Targets 1999, vol. 3809. International Society for Optics and Photonics, 1999, pp. 74–83.
[7] S. Kim, Y. Yang, J. Lee, and Y. Park, “Small target detection utilizing robust methods of the human visual system for irsi,” Journal of Infrared, Millimeter, and Terahertz Waves, vol. 30, no. 9, pp. 994–1011, 2009.
[8] C. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, “A local contrast method for small infrared target detection,” IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 1, pp. 574–581, 2013.
[9] J. Han, Y. Ma, B. Zhou, F. Fan, K. Liang, and Y. Fang, “A robust infrared small target detection algorithm based on human visual system,” IEEE Geoscience and Remote Sensing Letters, vol. 11, no. 12, pp. 2168–2172, 2014.
[10] J. Han, K. Liang, B. Zhou, X. Zhu, J. Zhao, and L. Zhao, “Infrared small target detection utilizing the multiscale relative local contrast measure,” IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 4, pp. 612–616, 2018.
[11] J. Han, S. Moradi, I. Faramarzi, H. Zhang, Q. Zhao, X. Zhang, and N. Li, “Infrared small target detection based on the weighted strengthened local contrast measure,” IEEE Geoscience and Remote Sensing Letters, 2020.
[12] J. Han, S. Moradi, I. Faramarzi, C. Liu, H. Zhang, and Q. Zhao, “A local contrast method for infrared small-target detection utilizing a tri-layer window,” IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 6, pp. 1822–1826, 2019.
[13] C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, “Infrared patch-image model for small target detection in a single image,” IEEE Transactions on Image Processing, vol. 22, no. 12, pp. 4996–5009, 2013.
[14] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis,” Journal of the ACM (JACM), vol. 58, no. 3, pp. 1–37, 2011.
[15] J. Guo, Y. Wu, and Y. Dai, “Small target detection based on reweighted infrared patch-image model,” IET Image Processing, vol. 12, no. 1, pp. 70–79, 2017.
[9] L. Chen, Y. Liu, and C. Zhu, “Iterative block tensor singular value thresholding for extraction of low-rank component of image data,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 1862–1866.

[10] C. D. Martin, R. Shafer, and B. LaRue, “An order-p tensor factorization with applications in imaging,” SIAM Journal on Scientific Computing, vol. 35, no. 1, pp. A474–A490, 2013.

[11] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover, “Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging,” SIAM Journal on Matrix Analysis and Applications, vol. 34, no. 1, pp. 148–172, 2013.

[12] C. Lu, J. Feng, Y. Chen, W. Liu, Z. h. Lin, and S. Yan, “Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5249–5257.

[13] W.-H. Xu, X.-L. Zhao, T.-Y. Ji, J.-Q. Miao, T.-H. Ma, S. Wang, and T.-Z. Huang, “Laplace function based nonconvex surrogate for low-rank tensor completion,” Signal Processing: Image Communication, vol. 73, pp. 62–69, 2019.

[14] Y. Sun, J. Yang, and W. An, “Infrared small target detection based on reweighted infrared patch-image model and total variation regularization,” Mathematical Problems in Engineering, 2020, 2020.

[15] L. Sun, T. Zhan, Z. Wu, and B. Jeon, “A novel 3d anisotropic total variation regularized low rank method for hyperspectral image mixed denoising,” ISPRS International Journal of Geo-Information, vol. 7, no. 10, p. 412, 2018.

[16] A. J. Tom and S. N. George, “Simultaneous reconstruction and moving object detection from compressed sampled surveillance videos,” IEEE Transactions on Image Processing, vol. 29, pp. 7590–7602, 2020.

[17] C. Gao, L. Wang, Y. Xiao, Q. Zhao, and D. Meng, “Infrared small-dim target detection based on markov random field guided noise modeling,” Pattern Recognition, vol. 76, pp. 463–475, 2018.

[18] Y. Wang, J. Peng, Q. Zhao, Y. Leung, X.-L. Zhao, and D. Meng, “Hyperspectral image restoration via total variation regularized low-rank tensor decomposition,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, no. 4, pp. 1227–1243, 2017.

[19] S. Boyd, N. Parikh, and E. Chu, Distributed optimization and statistical learning via the alternating direction method of multipliers. Now Publishers Inc, 2011.

[20] Z. Lin, M. Chen, and Y. Ma, “The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices,” arXiv preprint arXiv:1009.5055, 2010.

[21] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” SIAM journal on imaging sciences, vol. 2, no. 1, pp. 183–202, 2009.

[22] Y. Xie, S. Gu, Y. Liu, W. Zuo, W. Zhang, and L. Zhang, “Weighted schatten p-norm minimization for image denoising and background subtraction,” IEEE Transactions on Image Processing, vol. 25, no. 10, pp. 4842–4857, 2016.

[23] C. Gao, T. Zhang, and Q. Li, “Small infrared target detection using sparse ring representation,” IEEE Aerospace and Electronic Systems Magazine, vol. 27, no. 3, pp. 21–30, 2012.

[24] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, “Tensor robust principal component analysis with a new tensor nuclear norm,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 925–938, 2019.

[25] D. Pang, T. Shan, W. Li, P. Ma, S. Liu, and R. Tao, “Infrared dim and small target detection based on greedy bilateral factorization in image sequences,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 3394–3408, 2020.

[26] D. Pang, T. Shan, W. Li, P. Ma, R. Tao, and Y. Ma, “Facet derivative-based multidirectional edge awareness and spatial-temporal tensor model for infrared small target detection,” IEEE Transactions on Geoscience and Remote Sensing, 2021.
Ting Liu received the B.E. degree in electrical engineering and automation from Hunan Institute of Engineering, Xiangtan, China, in 2017, and the M.E. degree in control engineering from Xiangtan University (XTU), Xiangtan, China, in 2020. She is currently pursuing the Ph.D. degree with the College of Electronic Science in NUDT, Changsha, China. Her research interests focus on signal processing, target detection and image processing.

Jungang Yang received the B.E. and Ph.D. degrees from National University of Defense Technology (NUDT), in 2007 and 2013 respectively. He was a visiting Ph.D. student with the University of Edinburgh, Edinburgh from 2011 to 2012. He is currently an associate professor with the College of Electronic Science, NUDT. His research interests include computational imaging, image processing, compressive sensing and sparse representation. Dr. Yang received the New Scholar Award of Chinese Ministry of Education in 2012, the Youth Innovation Award and the Youth Outstanding Talent of NUDT in 2016.

Boyang Li received the B.E. degree in Mechanical Design manufacture and Automation from the Tianjin University, China, in 2017 and M.S. degree in biomedical engineering from National Innovation Institute of Defense Technology, Academy of Military Sciences, Beijing, China, in 2020. He is currently working toward the Ph.D degree in information and communication engineering from National University of Defense Technology (NUDT), Changsha, China. His research interests focus on VHR remote sensing image classification, infrared small target detection and deep learning.

Chao Xiao received the BE degree in the communication engineering and the ME degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China in 2016 and 2018, respectively. He is currently working toward the Ph.D. degree with the College of Electronic Science in NUDT, Changsha, China. His research interests include deep learning, small object detection and object tracking.

Yang Sun received the Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2020. He is currently a postdoctoral with the College of Electronic Science, NUDT. His research interests are image processing and infrared target detection.

Yingqian Wang received the B.E. degree in electrical engineering from Shandong University (SDU), Jinan, China, in 2016, and the M.E. degree in information and communication engineering from National University of Defense Technology (NUDT), Changsha, China, in 2018. He is currently pursuing the Ph.D. degree with the College of Electronic Science and Technology, NUDT. His research interests focus on low-level vision, particularly on light field imaging and image super-resolution.

Wei An received the Ph.D. degree from the National University of Defense Technology (NUDT), Changsha, China, in 1999. She was a Senior Visiting Scholar with the University of Southampton, Southampton, U.K., in 2016. She is currently a Professor with the College of Electronic Science and Technology, NUDT. She has authored or co-authored over 100 journal and conference publications. Her current research interests include signal processing and image processing.