RPCANet
Deep Unfolding RPCA Based Infrared Small Target Detection
Wu, Fengyi; Zhang, Tianfang; Li, Lei; Huang, Yian; Peng, Zhenming

Published in:
2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

Publication date:
2024

Document version
Peer reviewed version

Document license:
Unspecified

Citation for published version (APA):
RPCANet: Deep Unfolding RPCA Based Infrared Small Target Detection

Fengyi Wu¹,* Tianfang Zhang¹,* Lei Li², Yian Huang¹, Zhenming Peng¹†

¹University of Electronic Science and Technology of China, Chengdu, China
²University of Copenhagen, Denmark

fengyiwu@std.uestc.edu.cn; sparkcarleton@gmail.com; lilei@di.ku.dk
huangyian@std.uestc.edu.cn; zmeng@uestc.edu.cn

Abstract

Deep learning (DL) networks have achieved remarkable performance in infrared small target detection (ISTD). However, these structures exhibit a deficiency in interpretability and are widely regarded as black boxes, as they disregard domain knowledge in ISTD. To alleviate this issue, this work proposes an interpretable deep network for detecting infrared dim targets, dubbed RPCANet. Specifically, our approach formulates the ISTD task as sparse target extraction, low-rank background estimation, and image reconstruction in a relaxed Robust Principle Component Analysis (RPCA) model. By unfolding the iterative optimization updating steps into a deep-learning framework, time-consuming and complex matrix calculations are replaced by theory-guided neural networks. RPCANet detects targets with clear interpretability and preserves the intrinsic image feature, instead of directly transforming the detection task into a matrix decomposition problem. Extensive experiments substantiate the effectiveness of our deep unfolding framework and demonstrate its trustworthy results, surpassing baseline methods in both qualitative and quantitative evaluations. Our source code is available at https://github.com/fengyiwu98/RPCANet.

1. Introduction

Research on infrared small target detection (ISTD) plays a crucial role in both civilian and military applications, such as maritime rescue [23], early-warning systems [6], and reconnaissance activities [51]. However, targets usually appear as point targets in just a few pixels due to the inherent extensive imaging distance in infrared detection, leading to limited shape and texture information. Further, background clutter and imaging system noise interfere with distinguishing infrared targets from complex noise, making ISTD a challenging academic topic.

In recent years, ISTD techniques have been advanced, falling into model-driven and data-driven categories. The model-driven framework includes filter-based [2,9,14], human vision system (HVS)-based [5,17,36], and optimization-based methods [6,10,35]. These methods provide theoretical and physical foundations and interpretable detection results, but their efficacy often hinges on finely tuned parameters, which may not suffice when the environment changes.

Data-driven methods for detecting small infrared targets have been enabled by the evolution of Artificial Neural Networks (ANNs). However, general deep-learning-based object detection techniques [11,20,25] have shown limited
performance on ISTD due to the significant difference in shape and texture information between visible light and infrared. Therefore, recent studies have formulated ISTD as a segmentation problem [7, 8, 19, 47], which can better capture the discriminative and salient features of small targets. However, common segmentation ISTD methods [7, 8] that adopt ImageNet [27] backbones may limit the infrared data representation [38]. Moreover, the multiple downsampling operations in mainstream networks can cause target loss due to the small target size. The reliability of these networks is questionable as they are designed without ISTD domain knowledge.

Taking into account the benefits and drawbacks of both model and data-driven ISTD techniques, a pertinent question arises: Can we devise a balanced ISTD method that integrates the strengths of data-driven and model-driven methods to yield more reliable ISTD results? Deep unfolding network (DUN), also known as algorithm unrolling, is an emerging technique that bridges the gap between iterative algorithms and neural networks. This technique has attracted a great deal of attention across multiple fields [26, 29, 33, 42]. DUN constructs a network by unrolling the iterative solving algorithm of an existing model at the iteration level. Its hyperparameters are subsequently updated in a network-based fashion. By establishing systematic and precise connections between iterative algorithms and neural networks, DUNs exhibit considerable potential for building more interpretable networks [1].

Interestingly, optimization-based ISTD schemes, as an extended application of image completion, involve optimization steps such as model formulation (as shown in Fig. 1 (a), low-rank background and sparse target) and iterative solving methods [10, 37], which align with the requirement of DUNs. However, complex matrix operations hinder scholars from delving into DUN-based ISTD. Some studies in medical image restoration [16] and video separation [4] have modeled images using robust principle component analysis (RPCA). But directly learning parameters in singular value thresholding (SVT) and soft thresholding (ST) overlooks the inherent correlation of images and complicates the empirical selection of regularization parameters.

Thus, to design an ISTD network that addresses the above issues and balances efficiency and interpretability, we develop a deep architecture called Robust Principle Component Analysis Network (RPCANet), as shown in Fig. 1 (d). Rather than applying ST on deep features, this framework uses neural layers to approximate the sparsity constraint function within a target extraction module (TEM). In addition, we design a background extraction module (BEM) to mimic the proximal function with convolution layers, thus approximating the background and eliminating steps such as SVT. An image reconstruction module (IRM) is also introduced to iteratively combine the target and background outputs. In summary, our work contributes to the field of ISTD by:

- RPCANet is a learnable deep network architecture derived from a relaxed RPCA model in Fig. 1 (b) and (c), combining the accuracy of data-driven networks with the interpretability of model-driven ISTD approaches.
- Our TEM and BEM approximate the target and background for segmentation. Nonlinear proximal mapping problems for targets are handled effectively, while complex matrix calculations in background estimation are replaced with neural layers. Moreover, the proposed IRM combines the background and targets, aiming to reconstruct the image.
- Experimental results on multiple datasets demonstrate the effectiveness of RPCANet compared to state-of-the-art baselines. Visualization of learned characteristics provides comprehensive and reliable results.

2. Related Work

2.1. Optimization-based ISTD

Numerous ISTD methods have been developed over the past decades. Optimization-based methods are dominant in the model-driven category, compared to traditional filter-based [2, 9, 14] and HVS-based [5, 17, 36] schemes. The infrared patch image (IPI) model by Gao et al. [10] introduces the method of low-rank and sparse decomposition within the stable RPCA model [52]. This matrix-based category has been enriched by works in single [44, 49] and multi [35] subspaces. Based on the hypothesis that tensors leverage correlation information better than matrices [12], Dai et al. proposed a reweighted infrared patch tensor (RIPT) model for ISTD [6]. Various strategies [18, 37, 45] based on tensors have been applied to both single and multi-frame infrared sequences. The optimization-based methods model targets and backgrounds in a mathematically interpretable manner within RPCA. However, they have limitations in robustness, parameter tuning, and efficiency [46, 48]. Therefore, our goal is to develop a model that preserves interpretability while improving robustness and efficiency.

2.2. Deep Learning-based ISTD

In contrast to model-driven ones, convolutional neural networks (CNNs) are data-driven to learn the non-linear mappings between the original images and masks and are flexible to complicated scenarios. Wang et al. proposed a conditional GAN (MDvsFA-CGAN) [34] to reduce false alarms and missed detections in ISTD. Dai et al. presented an asymmetric contextual modulation (ACM) method [7] to enhance feature fusion. Zhang et al. [47] extracted contextual information of the target in the deep layer, based
on ACM. Li et al. [19] designed the DNA Netanyahu with an enhanced receptive field to prevent small targets from disappearing. Wu et al. [38] integrated residual blocks into the network, which preserves feature resolution. Zhang et al. [46] merged an edge block in a U-Net framework to enhance edge features, considering target shapes vary. Ying et al. [40] updated labels with single-point supervision, based on effective models [7, 8, 19]. However, most data-driven methods tailor the segmentation network, which is a black box, instead of incorporating classic ISTD algorithms with field knowledge. Only a few studies have tried this approach [8, 15]. Therefore, we aim to find a solution that balances interpretability and effectiveness.

### 2.3. Deep Unfolding Networks

Deep unfolding networks (DUNs) solve iterative optimization problems via neural networks and have been widely used in image processing. Gregor and LeCun [13] first proposed the Learned IST A (LISTA) model, which learns the parameters of the iterative shrinkage-thresholding algorithm (ISTA). Zhang et al. [41, 42] improved it by introducing ISTA-Nets, which incorporate neural layers into the update steps. Yang et al. [39] developed ADMM-CSI Net, which is based on unfolding the alternating direction multiplier method (ADMM) for MRI-oriented image reconstruction. Borgerding et al. [3] applied the approximate message passing algorithm (AMP) to learn sparse linear inverse problems, and Zhang et al. [50] extended it to compressive sensing. Other works also use DUNs for optimization models, such as low-rank representation [48] and RPCA [4, 28]. These DUN frameworks blend the strengths of model and data-driven methods, achieving high and robust performance. Given the solid knowledge of optimization and deep learning-based methods, a progression towards DUN-based ISTD is a logical next step.

### 3. Deep Unfolding RPCA Network

#### 3.1. Problem Formulation

For an infrared image $\mathbf{D}$, the physical model separates it into low-rank background $\mathbf{B}$ and sparse target $\mathbf{T}$ as:

$$
\mathbf{D} = \mathbf{B} + \mathbf{T},
$$

where $\mathbf{D}, \mathbf{B}, \mathbf{T} \in \mathbb{R}^{m \times n}$. In a general RPCA [52] manner, we aim to recover the low-rank background $\mathbf{B}$ that could pair the given $\mathbf{D}$ by constraining sparse target $\mathbf{T}$ and usually transform the detection problem into:

$$
\min_{\mathbf{B, T}} \text{rank} (\mathbf{B}) + \lambda \| \mathbf{T} \|_0 \quad \text{s.t.} \quad \mathbf{D} = \mathbf{B} + \mathbf{T},
$$

where $\lambda$ indicates a positive trade-off parameter, and $\| \cdot \|_0$ denotes the $l_0$-norm as the number of nonzero entries.

However, solving (2) is NP-hard since the rank function and $l_0$-norm are both non-convex and discontinuous. Considering this, models like IPI [16] individually replace them with the nuclear norm ($\| \cdot \|_*$, the sum of singular values in a matrix) and $l_1$-norm ($\| \cdot \|_1$) through principal component pursuit (PCP) and reformulate (2) as:

$$
\min_{\mathbf{B, T}} \| \mathbf{B} \|_* + \lambda \| \mathbf{T} \|_1 \quad \text{s.t.} \quad \mathbf{D} = \mathbf{B} + \mathbf{T}.
$$

In complex infrared scenarios, the background and target may vary in complexity and sparsity, and a single norm or rank function may not capture the practical constraints [49]. Thus, we use $\mathcal{R} (\mathbf{B})$ and $\mathcal{S} (\mathbf{T})$ to constrain the prior knowledge of the background and target images, respectively:

$$
\min_{\mathbf{B, T}} \mathcal{R} (\mathbf{B}) + \lambda \mathcal{S} (\mathbf{T}) \quad \text{s.t.} \quad \mathbf{D} = \mathbf{B} + \mathbf{T}.
$$

Moreover, to simplify the complexity of updating variables due to the augmented Lagrange multipliers [6], we adopt a simpler and more intuitive $l_2$-norm to transform the constrained problem into an unconstrained one [30] as:

$$
\mathcal{L} (\mathbf{B}, \mathbf{T}) = \mathcal{R} (\mathbf{B}) + \lambda \mathcal{S} (\mathbf{T}) + \frac{\mu}{2} \| \mathbf{D} - \mathbf{B} - \mathbf{T} \|_F^2,
$$

where $\mu$ is a penalty coefficient, and $\| \cdot \|_F$ indicates the Frobenius norm (F-norm). For a matrix $\mathbf{X}$, its F-norm equals $\sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |X_{ij}|^2}$. Based on (5), we can optimize the background and target individually in an iterative scheme.

### 3.2. Model Iterative Solving

**Updating $\mathbf{B}^\star$:** To update the background, we draw the sub-problem as:

$$
\mathbf{B}^\star = \arg \min_{\mathbf{B}} \mathcal{R} (\mathbf{B}) + \frac{\mu}{2} \| \mathbf{B} + \mathbf{T} - \mathbf{D} \|_F^2.
$$

As (3) illustrates, former optimization-based ISTD approaches usually set $\mathcal{R} (\mathbf{B})$ as $\| \mathbf{B} \|_*$, then degenerates (6) into the sum of the nuclear norm and $l_2$ norm, and this problem has an analytical solution as:

$$
\mathbf{B}^\star = \mathcal{D}_\mu (\mathbf{D} - \mathbf{T}),
$$

where $\mathcal{D}_\mu (\cdot)$ denotes the SVT operator [21] with the threshold of $\mu$. However, solving (7) involves SVD and functions on its singular values. In DUNs, we emulate this using neural networks, which means we must perform SVD on each neural tensor in each forward propagation. This poses challenges for time consumption and accuracy. [4] proposes an initialization method based on the best rank-r approximation SVD, but still faces precision issues. [48] introduces a technique with different rank constraints for the submatrices, which improves interpretability but depends on manual tuning and ignores the intrinsic image properties when simulating matrix computation in neural layers.

Thus, instead of adopting the nuclear norm and solving it with complex SVDs, we degrade it to a constrain function $\mathcal{R} (\mathbf{B})$ in this study, and introduce a proximal operator $\text{prox}_\mu (\cdot)$ to approximate the closed-form solution for the background, which is formulated as:

$$
\mathbf{B}^\star = \text{prox}_\mu (\mathbf{D} - \mathbf{T}).
$$
We use convolutional layers to approximate proximal functions and solve the optimization problem. This method eliminates the need for complex matrix operations and leverages the nonlinear capabilities of neural networks to extract deep features from images in a data-driven way. And we will illustrate the detailed construction in Section 3.3.

Updating $T^*$: Similar with (6), the sub-problem of optimizing the target is written as:

$$T^* = \arg \min_{T} \lambda S(T) + \frac{\mu}{2} \|T + B - D\|_F^2.$$  \hspace{1cm} (9)

As discussed in Section 3.1, common optimization approaches impose an $l_1$ norm constraint on the sparse target image. However, this poses the challenge of mapping the soft thresholding into the neural network [42]. Moreover, sparse constraints often vary depending on the detection scenarios change [49]. Thus, we aim to derive a simpler and more intuitive representation for the closed-form solution of (9). To handle the above issues, we consider conducting Taylor expansion of $S(T)$. For a function $f(t)$, with its Lipschitz continuous gradient function $\nabla f(t)$, then $f(t)$ can be Taylor approximated at the a fix point $t_0$ by:

$$\hat{f}(t, t_0) \approx \frac{L}{2} \|t - t_0 + \frac{1}{L} \nabla f(t_0)\|^2 + C,$$  \hspace{1cm} (10)

where $L$ is a constant, $C = \frac{1}{2L} \|\nabla f(t_0)\|^2 + f(t_0)$ (please refer to the supplementary material for detail deductions). Based on this, we approximate $S(T)$ at the last $T^{k-1}$ as:

$$\hat{S}(T, T^{k-1}) \approx \frac{L_S}{2} \|T - T^{k-1} + \frac{1}{L_S} \nabla S(T^{k-1})\|^2_2 + C_S,$$  \hspace{1cm} (11)

where $L_S$ is the Lipschitz constant of $S(T)$ and $C_S = \frac{1}{2L_S} \|\nabla S(T^{k-1})\|^2_2 + S(T^{k-1})$ represents as a constant. And the updating formulation of the target matrix can be substituted as:

$$T^* = \arg \min_{T} \lambda \hat{S}(T, T^{k-1}) + \frac{\mu}{2} \|T + B - D\|_F^2.$$  \hspace{1cm} (12)

which simply involves the sum of two $l_2$ norms, rather than a traditional $l_1$ norm constraint optimization problem, and therefore does not require the use of conventional algorithms or simulating soft thresholding [42]. By taking the derivative of the equation and equating it to zero, a closed-form solution for updating $T$ at the $k$-th step can be derived:

$$T^k = \frac{\lambda L_S}{\lambda L_S + \mu} T^{k-1} + \frac{\mu}{\lambda L_S + \mu} (D^{k-1} - B^k) - \frac{\lambda}{\lambda L_S + \mu} \nabla S(T^{k-1}) ,$$  \hspace{1cm} (13)

Figure 2. Overall structure of RPCANet. The network is composed of $K$ stages, each following the unfolding steps detailed in Fig. 3.

$3.3$. RPCANet Framework

This section describes the overall architecture of the network and module design of RPCANet, based on the optimization equations in Section 3.2. As shown in Fig. 2, the input to the network is an infrared image $X \in \mathbb{R}^{H \times W}$ with targets, where $H$ and $W$ are the image height and width, and the update parameters are initialized as $D^0 = X$ and $T^0 = 0$. These parameters are then passed through $K$ decomposition stages, each corresponding to an iterative matrix low-rank sparse decomposition process, to simulate the update operation of multiple iterations in model-driven approaches.

In detail, the updated parameters $D^{k-1}$ and $T^{k-1}$ are fed into the $k$-th decomposition stage, where $k \in \{1, \ldots, K\}$. The background $B^k$, target $T^k$, and reconstructed result $D^k$ for the current stage are estimated by BEM, TEM, and IRM, respectively. Typically, $B^k$ represents the latent variable of the current decomposition stage and is not involved in the parameter transfer between stages.
32 in RPCANet: background extraction module (BEM), target extraction module (TEM), and image reconstruction module (IRM).

**Background Estimation Module (BEM):** As shown in Fig. 3, we adopt BEM to estimate the background. In (8), proximal operator \( \text{prox}_\mu(\cdot) \) is to-be-decided. Here, inspired by former DUN-based works [29, 32, 33, 50] and in the spirits of flexibility [41], we adopt a residual structure \( \text{proxNet}(\cdot) \) to simulate this operator. As (16) formulates:

\[
B^k = \text{proxNet}(D^{k-1} - T^{k-1}) = D^{k-1} - T^{k-1} + F^k(D^{k-1} - T^{k-1}),
\]

where \( F^k(\cdot) \) consists of \( l_B \) middle layers [Conv + BN + ReLU], and two convolution layers: feature extraction layer [Conv + BN + ReLU] and image reconstruction layer [Conv], respectively. Here, BN stands for batch normalization and ReLU for rectified linear unit [22]. All of them are in the stride of 1 and the padding of 1, and we set \( C = 32 \) and \( l_B = 3 \) in this study.

**Target Extraction Module (TEM):** This module takes the updated background \( B^k \), target \( T^{k-1} \), and reconstruction result \( D^{k-1} \) as inputs, as shown in Fig. 3. Moreover, we set \( \gamma = 0.5 \) to evenly treat the three inputs and rewrite (14) as:

\[
T^k = T^{k-1} + D^{k-1} - B^k - \varepsilon \nabla S(T^{k-1}).
\]

We assign the parameter learning task to \( \epsilon \) as \( \epsilon^k \), which is a learnable scalar independent of each reconstruction stage and does not share parameters. As to the Lipschitz continuous gradient function \( \nabla S \), [31] finds that a single-layer CNN consisting of a convolution layer and a ReLU activation function is Lipschitz continuous, which also holds for multiple stacked layers. Therefore, to avoid complex network design, we adopt simple convolution layers and ReLU to simulate the function \( \nabla S \) as shown in the orange box in Fig. 3, in accordance with the assumption in Section 3.2. We also introduce the difference of last \( D^{k-1} \) and updated \( B^k \) to enhance successive target feature, and the update function of \( T^k \) is written as:

\[
T^k = T^{k-1} + D^{k-1} - B^k - \epsilon^k G^k(T^{k-1} + D^{k-1} - B^k). \quad (18)
\]

Specifically, \( G^k(\cdot) \) comprises an initial convolution layer, \( l_T \) of middle layers, and a reconstruction layer. Since the Lipschitz continuous property of the statistical operation BN is not clear [31], we omit the BN from the convolution block.

**Image Reconstruction Module (IRM):** To align with the RPCA process that maps the target and background into a restored image, we devise an IRM that converts the decomposition task into an image reconstruction task with a neural network \( M^k(\cdot) \), as presented in Fig. 3:

\[
D^k = M^k(B^k + T^k). \quad (19)
\]

Instead of applying residual blocks or other complex networks in the reconstruction module [32, 33], we employ a simple and decent CNN architecture [43] in learning image features and mapping the decomposed background and target effectively. Similar to \( F^k(\cdot) \), \( M^k(\cdot) \) has three types of convolution layer with \( l_D \) middle layer, where \( l_D = 3 \).

The corresponding updating operations in optimization-based and our DUN-based frameworks are shown in Fig. 4. To sum up, we propose an end-to-end training framework for the ISTD task named RPCANet. Its trainable parameters set \( \Theta \) includes the convolutional network parameters in BEM, TEM, and IRM in each decomposition stage, and the independent trainable scalar \( \epsilon \) in TEM, which collects as \( \Theta = \{ \Theta^\text{BEM}, \Theta^\text{TEM}, \Theta^\text{IRM}, \epsilon^k \}_{k=1}^K \), where \( K \) is the number of overall decomposition stages.

4. Network Training
4.1. Training Loss

Since we separate an ISTD task into the target segmentation and infrared image reconstruction, thus the loss func-
Figure 5. Heatmaps of variables $B^k$ and $T^k$ in different decomposition stages of RPCA on three representative scenes when $K = 6$.

Table 1. Effect of stage number $K$ on the detection performance in mIoU ($\%$), $F_1$ ($\%$), $P_d$ ($\%$), and $F_a$ ($10^{-5}$) on SIRST-Aug [47].

<table>
<thead>
<tr>
<th>Stages ($K$)</th>
<th>Params</th>
<th>mIoU</th>
<th>$F_1$</th>
<th>$P_d$</th>
<th>$F_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.113M</td>
<td>60.10</td>
<td>75.08</td>
<td>98.21</td>
<td>62.20</td>
</tr>
<tr>
<td>2</td>
<td>0.227M</td>
<td>69.26</td>
<td>81.84</td>
<td>98.49</td>
<td>46.12</td>
</tr>
<tr>
<td>3</td>
<td>0.340M</td>
<td>69.82</td>
<td>82.23</td>
<td>98.07</td>
<td>42.63</td>
</tr>
<tr>
<td>4</td>
<td>0.453M</td>
<td>69.50</td>
<td>82.01</td>
<td>98.07</td>
<td>40.36</td>
</tr>
<tr>
<td>5</td>
<td>0.567M</td>
<td>70.24</td>
<td>82.52</td>
<td>96.01</td>
<td>36.02</td>
</tr>
<tr>
<td>6</td>
<td>0.680M</td>
<td>72.54</td>
<td>84.08</td>
<td>98.21</td>
<td>34.14</td>
</tr>
<tr>
<td>7</td>
<td>0.793M</td>
<td>70.98</td>
<td>83.02</td>
<td>96.15</td>
<td>35.85</td>
</tr>
</tbody>
</table>

Table 2. Effect of layer number $l_T$ on the detection performance in mIoU ($\%$), $F_1$ ($\%$), $P_d$ ($\%$), and $F_a$ ($10^{-5}$) on SIRST-Aug [47].

<table>
<thead>
<tr>
<th>$l_T$</th>
<th>Params</th>
<th>mIoU</th>
<th>$F_1$</th>
<th>$P_d$</th>
<th>$F_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.402M</td>
<td>67.07</td>
<td>80.29</td>
<td>99.17</td>
<td>40.53</td>
</tr>
<tr>
<td>3</td>
<td>0.513M</td>
<td>70.63</td>
<td>82.79</td>
<td>97.80</td>
<td>41.36</td>
</tr>
<tr>
<td>9</td>
<td>0.846M</td>
<td>67.99</td>
<td>80.95</td>
<td>94.36</td>
<td>32.75</td>
</tr>
<tr>
<td>12</td>
<td>1.013M</td>
<td>69.98</td>
<td>82.34</td>
<td>97.25</td>
<td>38.46</td>
</tr>
<tr>
<td>6</td>
<td>0.680M</td>
<td>72.54</td>
<td>84.08</td>
<td>98.21</td>
<td>34.14</td>
</tr>
</tbody>
</table>

Table 3. Studies on different proximal networks and w/wo IRM on the detection performance in mIoU ($\%$), $F_1$ ($\%$), $P_d$ ($\%$), and $F_a$ ($10^{-5}$) on SIRST-Aug [47].

<table>
<thead>
<tr>
<th>BEM IRM</th>
<th>Parameters</th>
<th>mIoU</th>
<th>$F_1$</th>
<th>$P_d$</th>
<th>$F_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RB ✓</td>
<td>0.624M</td>
<td>66.33</td>
<td>79.76</td>
<td>98.07</td>
<td>61.39</td>
</tr>
<tr>
<td>CNN ✓</td>
<td>0.679M</td>
<td>71.43</td>
<td>83.33</td>
<td>97.77</td>
<td>32.45</td>
</tr>
<tr>
<td>Ours X</td>
<td>0.507M</td>
<td>67.07</td>
<td>80.29</td>
<td>96.56</td>
<td>36.89</td>
</tr>
<tr>
<td>Ours ✓</td>
<td>0.680M</td>
<td>72.54</td>
<td>84.08</td>
<td>98.21</td>
<td>34.14</td>
</tr>
</tbody>
</table>

acquires more comprehensive information, demonstrating its ability to learn from formula-guided training. Likewise, $T^k$ progressively learns the target’s location from disorganized high-frequency information, resulting in a sparse matrix as the segmented output. Fig. 5 (c) shows that false alarms in the initial stages are eliminated under mask supervision, showing the data-driven way corrects the detection result. These successive heatmaps verify the effectiveness of RPCAnet in learning infrared variables according to training data while preserving the interpretability of RPCA model that facilitates the ISTD task.

5.2. Ablation Studies

Effects of Parameter $K$ and $l_T$: Table 1 compares
the effects of different reconstruction stage counts. Our method only reaches significant detection performance in two stages, validating the essential attributes of the suggested RPCANet. Additionally, we see that $K = 7$ has a worse ISTD performance than $K = 6$, which makes sense given that a higher K could hinder gradient propagation, based on this, we set $K$ to 6. Similarly, as shown in Table 2, the number of the convolutional layer $l_T$ faces the same manner, the performance of the network can be improved as $l_T$ raises, but the gain effect of too many reconstruction stages on the network performance is limited, thus, we take $l_T = 6$ in all our experiments.

**Studies of proxNet($\cdot$) and IRM:** The emulation of proximal operators acts an essential role in background estimation [33, 41], we investigate two proxNet($\cdot$) constructions (residual block (RB) [32] and plain CNN) for proximal operators. Table 3 demonstrates that RB can yield some effects and a small parameter count, our network in Section 3.3 achieves better results with a minimal parameter increase. We also examine the influence of the IRM. As depicted in the third row of Table 3, RPCANet, guided by the reconstruction module, exhibits substantial improvements in four metrics, proving the module’s effectiveness.

**Studies of $\nabla S$ Simulation Network:** We investigate the impact of incorporating $D^{k-1}$ and $B^k$ subtraction within $G^k(\cdot)$ for target feature enhancement. Table 4 compares two configurations. While using $T^{k-1}$ alone yields satisfactory results, incorporating additional information in (18) significantly improves $mIoU$ and $F_1$. We can conclude that appropriate modifications to the network structure, guided by the model’s prior, can enhance detection performance.
Table 5. mIoU (%), \(F_1\) (%), \(P_d\) (%), \(F_a\) (10^{-5}), and runtime values of different methods performance on NUDT-SIRST [19], IRSTD-1K [46], and SIRST-Aug [47]. The second column records the parameter statistics for data-driven algorithms.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(mIoU)</td>
<td>(P_d)</td>
<td>(F_1)</td>
<td>(F_a)</td>
<td>(mIoU)</td>
</tr>
<tr>
<td>Tophat [2]</td>
<td>-</td>
<td>22.23</td>
<td>36.37</td>
<td>96.19</td>
<td>76.27</td>
<td>5.35</td>
</tr>
<tr>
<td>MPCM [36]</td>
<td>-</td>
<td>9.26</td>
<td>16.95</td>
<td>70.58</td>
<td>32.74</td>
<td>14.87</td>
</tr>
<tr>
<td>IPI [10]</td>
<td>-</td>
<td>34.62</td>
<td>51.43</td>
<td>92.38</td>
<td>7.54</td>
<td>18.64</td>
</tr>
<tr>
<td>PSTNN [45]</td>
<td>-</td>
<td>25.46</td>
<td>40.58</td>
<td>78.52</td>
<td>7.95</td>
<td>16.38</td>
</tr>
<tr>
<td>ACM [7]</td>
<td>0.398M</td>
<td>69.00</td>
<td>81.66</td>
<td>95.98</td>
<td>13.34</td>
<td>61.56</td>
</tr>
<tr>
<td>ALCNet [8]</td>
<td>0.427M</td>
<td>71.48</td>
<td>83.37</td>
<td>96.30</td>
<td>11.45</td>
<td>58.23</td>
</tr>
<tr>
<td>ISNet [46]</td>
<td>0.967M</td>
<td>87.51</td>
<td>93.34</td>
<td>97.35</td>
<td>3.37</td>
<td>55.29</td>
</tr>
<tr>
<td>AGPCNet [47]</td>
<td>12.360M</td>
<td>85.40</td>
<td>92.13</td>
<td>98.10</td>
<td>4.72</td>
<td>61.00</td>
</tr>
<tr>
<td>DNANet [19]</td>
<td>4.697M</td>
<td>83.94</td>
<td>91.27</td>
<td>98.52</td>
<td>6.21</td>
<td>60.51</td>
</tr>
<tr>
<td>UIUNet [38]</td>
<td>50.540M</td>
<td>88.71</td>
<td>94.01</td>
<td>91.43</td>
<td>1.89</td>
<td>63.06</td>
</tr>
<tr>
<td>Ours</td>
<td>0.680M</td>
<td>89.31</td>
<td>94.35</td>
<td>97.14</td>
<td>2.87</td>
<td>63.21</td>
</tr>
</tbody>
</table>

5.3. Comparison to State-of-the-art Methods

For model-driven methods, we select: filter-based Tophat [2], HVS-based MPCM [36], matrix optimization-based IPI [10], and tensor optimization-based PSTNN [45]. For data-driven algorithms, we conduct experiments on ACM [7], ALCNet [8], ISNet [46], AGPCNet [47], DNANet [19], and UIUNet [38].

Qualitative Results: Fig. 6 depicts visual results obtained from various algorithms applied to three datasets. RPCANet effectively generalizes complex scenarios, producing outputs with accurate target shapes and low false alarm rates. Compared to model-driven algorithms, our network excels at suppressing false alarms while preserving the sparse shape of infrared targets. Although some DL-based models can effectively reduce false alarms, they may struggle in small and complex situations, leading to missed detections. Algorithms like ISNet largely maintain the target shape; however, they occasionally encounter difficulties in multi-target scenarios. The visualization results clearly illustrate that our network combines the advantages of optimization-based and DL-based methods.

Quantitative Results: To showcase the effectiveness of our network, we compare its detection performance with the SOTA baseline. Table 5 demonstrates that RPCANet outperforms most model-driven and data-driven ISTD frameworks in four indicators using fewer parameters. Model-driven methods like IPI excel in target-level metrics but lack pixel-level accuracy. Data-driven networks improve mIoU and \(F_1\) scores while preserving the target’s shape. However, DL-based algorithms like AGPCNet and UIUNet suffer from overfitting and require more parameters. In contrast, RPCANet combines model-driven priors with accurate object extraction and segmentation guided by data, achieving superior performance with fewer parameters. ROC curves in Fig. 7 highlight that our model rapidly reaches the upper-left corner and exhibits competitive performance in terms of AUC: RPCANet (0.9857), ACM (0.9830), and UIUNet (0.9477). Moreover, the last column in Table 5 shows that our framework obtains great computational efficiency.

6. Conclusion

In this paper, we propose an interpretable framework for infrared small target detection based on deep unfolding networks. We model the ISTD task as a relaxed RPCA problem and solve the optimization steps via network emulations, including the proximal network and sparsity-constrained neural layers. Our scheme produces trustworthy visualizations and outstanding detection results in extensive experiments on various public datasets. The architecture of RPCANet effectively guides neural layers to learn low-rank backgrounds and sparse targets, facilitating the detection tasks in an almost “white box” manner. We hope our findings cloud inspire researchers to explore more interpretable solutions for ISTD problems in the future.

Acknowledgments

This work was supported by the Natural Science Foundation of Sichuan Province of China (No.2022NSFSC40574) and National Natural Science Foundation of China (No.61775030, No.61571096).


