Local Patch Network with Global Attention for Infrared Small Target Detection

Fang Chen, Chenqiang Gao*, Fangcen Liu, Yue Zhao, Yuxi Zhou, Deyu Meng, Member, IEEE, Wangmeng Zuo, Senior Member, IEEE

Abstract—Infrared small target detection plays an important role in the infrared search and tracking applications. In recent years, deep learning techniques were introduced to this task and achieved noteworthy effects. Following general object segmentation methods, existing deep learning methods usually processed the image from the global view. However, the imaging locality of small targets and extreme class-imbalance between the target and background pixels were not well-considered by these deep learning methods, which causes the low-efficiency on training and high-dependence on numerous data. A local patch network (LPNet) with global attention is proposed in this paper to detect small targets by jointly considering the global and local properties of infrared small target images. From the global view, a supervised attention module trained by the small target spread map is proposed to suppress most background pixels irrelevant with small target features. From the local view, local patches are split from global features and share the same convolution weights with each other in a patch net. By leveraging both the global and local properties, the data-driven framework proposed in this paper has fused multi-scale features for small target detection. Extensive synthetic and real data experiments show that the proposed method achieves the state-of-the-art performance compared with existing both conventional and deep learning methods.

Index Terms—Infrared image, small target detection, patch network, attention mechanism.

I. INTRODUCTION

The infrared search and tracking (IRST) system plays an important role in kinds of applications, including early warning and maritime surveillance, etc. As one of critical techniques of IRST, infrared small target detection determines the performance of the IRST system and is still a challenging task since the target lacks obvious shape and texture characteristics. Besides, complex background clutters usually seriously interfere the detector, as shown in Fig. 1.

In early stages, some methods based on background estimation [1], [2] were proposed through subtracting the estimated background from the original infrared image to detect small targets. These methods were fast and had low computational complexity. Some other methods based on local contrast measure [3]–[7] focused on target enhancement and background suppression through capturing local contrast and saliency. These methods could effectively suppress clutters, especially when facing to complicated backgrounds. Some methods [8]–[12] based on nonlocal self-correlation property achieved favorable performance by transforming the task into a low-rank and sparse matrix separation problem.

Totally, above methods fully rely on prior knowledge to design filters or models for small target detection. These methods are not feasible when applying to kinds of specific applications. When the prior knowledge of designing methods does not match that of a specific application scene, the new prior knowledge of the scene is not easily or fast embedded into these methods. In contrast, learning based methods with training samples from the scene. Recently, some deep learning frameworks [13]–[15] are proposed to detect infrared small targets by training models on large amounts of data. These methods are completely driven by data instead of expert knowledge so that better robustness and generalization have been verified. However, existing deep learning methods are based on the global view to extract features from infrared images and ignore the imaging locality of small targets. Only focusing on the global view causes numerous parameters of

Fig. 1. Representative examples of infrared small targets. The target is indicated by a red bounding box and a close-up is shown in the bottom right corner of each example. Left: an infrared small target image which has heavily cloudy clutters interfering detectors. Right: a target so small to lose its shape characteristics due to the long imaging distance.
models and low-efficiency on training because the designed models have to exclude the disturbance of background which took up almost all pixels of an infrared image. Besides, the complexity and disparity of the background lead to these global view based methods depending on numerous training data with heterogeneous scenes in order to learn sufficient features for suppressing the background.

When just considering the global property of small target images, the class-imbalance of target and background pixels is the main obstacle for training a favorable and robust detector. Taking up only a few pixels, small targets are generally too small to be distinguished from the background. For example, in a $120 \times 120$ image containing a $2 \times 2$ small target, these four target pixels have to share the same convolution weights with the rest 14396 background pixels. This will cause the extreme class-imbalance: the amount of small target pixels is much less than background pixels, which leads to performance degradation of deep learning frameworks.

In this paper, we propose a local patch network (LPNet) with global attention for infrared small target detection from the global view and the local view. From the global view, we utilizes a small target spread map to train an attention module. This spread map and attention mechanism are designed to suppress distal background pixels irrelevant with the small target. By this way, most background pixels could be suppressed while the local saliency between small target pixels and adjacent background pixels is retained, so that the class-balance between small target pixels and background pixels is achieved. Besides, with most background pixels being suppressed, the complexity of background is greatly reduced and thus the proposed method could still effectively extract small target features even being trained on limited data with heterogeneous scenes.

From the local view, the global feature maps are divided to a series of patches by using a sliding window to extract sub-images and a neural network is utilized to extract small target features and allocate a patch-likelihood map for each local patch. The ratio of small target pixels to background pixels in each patch is much higher than in the original image, so that the class-balance of small target and background is achieved. Besides, with sharing the same convolution weights for all local patches, the amount of parameters of the neural network is limited.

To leverage both the global view to the local view, a well-designed framework is proposed to combine above modules for the global and local feature fusion. The proposed method is performed on two widely used public infrared small target datasets and a private dataset collected by ourselves to compare with state-of-the-art methods on comprehensive scenes. Besides, a series of ablation studies are performed for illustrating the remarkable effect of focusing on locality and class balance for small target detection.

In summary, contributions of this paper can be summarized as follows:

- We provide a paradigm of deep learning methods to leverage both global and local properties of the infrared image for detecting small targets.
- We propose a supervised attention module trained on a small target spread map. This module suppresses background likely pixels and enhances small target likely pixels to force the whole network to focus on some local regions, which is effective to address the class-imbalance and reduce the complexity of the background. Besides, this attention module enables the proposed method to still maintain its effect even being trained on limited data with heterogeneous scenes.
- We propose a patch net to effectively extract local features correlated with small targets and fusing all these local features to provide the final predict. By patch splitting, the class-imbalance of target pixels and background pixels is further addressed. Through sharing the same convolution weights for all local patches, the task of small target detection is greatly simplified and the amount of parameters in patch net is limited.
- Extensive experiments on widely used and our private infrared small target datasets illustrate that the proposed method has achieved the best performance compared with state-of-the-art methods. Besides, our method has the ability of effectively extracting small target features even being trained on limited data with heterogeneous scenes.

The remainder of this paper is organized as follows: In Section [II], related works are briefly reviewed. In Section [III], we present the proposed method in detail. In Section [IV], the experimental results are given and discussed. Conclusions are drawn in Section [V].

II. Related work

Up to now, many methods for infrared small target detection have been proposed, which can be roughly categorized into two groups: conventional methods and deep learning methods.

Conventional methods were instructed by prior knowledge to design filters or modules [1], [16]–[23]. For example, some methods focused on the background estimation in early stages, such as the top-hat algorithm [2] and max-mean/max-median algorithm [24], which directly subtracted the estimated background from the original image to acquire detected small targets. The average absolute gray difference (AAGD) algorithm [3], absolute directional mean difference (ADMD) algorithm [4] and generalised structure tensor (GST) based algorithm [25] detected small targets through suppressing the estimated background as much as possible. Unlike these background estimation methods, some noise estimation methods regarded the dim-small target as noise and addressed the small target detection issue by estimating the noise. Gao et al. [9] proposed a Mixture of Gaussians (MoG) to model the small target as a special sparse noise component of the background noise by MoG with Markov random field (MRF), so that the targets could be separated from background by variational Bayesian.

Some local contrast methods designed special local filters to construct pixel-contrast in local regions. For example, local contrast measure (LCM) [26] and the tri-layer local contrast method (TLLCM) [5] applied an internal sliding window over the entire image to capture the local contrast and used an
adapting threshold to segment small targets from the image. Inspired by LCM and human visual system, some local based methods have been proposed [27]–[32]. Han et al. [33] proposed an improved local contrast measure (ILCM) to enhance the local saliency of small targets. Wei et al. [34] proposed a multi-scale patch-based contrast measure (MPCM), which divided each image patch into nine cells and calculated the dissimilarities between the surrounding cells and the central one. By multiplying directional dissimilarities and minimum selection among the results, the final output was obtained. Based on MPCM, Xia et al. [6] conceived a local dissimilarity descriptor encoded by the local energy factor (LEF) and applied the LEF in MPCM. Gao et al. [35] applied a temporal variance filter after MPCM to remove small broken cloud regions and suppress noise. Zhang et al. [7] proposed a method based on local intensity and gradient (LIG) properties of small targets to suppress clutter and enhance small targets.

Nonlocal self-correlating methods utilized the nonlocal self-correlating property of infrared background images and achieved favorable performance by transforming the task of detecting the small target into a low-rank and sparse matrix separation problem [7], [36]–[39]. Gao et al. [8] proposed the patch-image in infrared small target detection and constructed the low-rank based infrared patch-image (IPI) model. Dai et al. [40] proposed a column weighted IPI model (WIPI), and then proposed a re-weighted infrared patch-tensor model (RIPT) [41]. Wang et al. [10] proposed a patch image model with local and global analysis (PILGA) to constrain the separability of noise patches. Based on the IPI model, Wang et al. [42] mapped the background to multiple sub-spaces and proposed a low-rank and sparse decomposition method based on greedy bilateral factorization. There are also some methods constrained sparse term to suppress background or enhance target. Zhang et al. [11] proposed a novel infrared small target detection method based on non-convex optimization with $l_p$-norm constraint. Besides, Zhang et al. [12] employed a novel non-convex low-rank constraint named partial sum of tensor nuclear norm (PSTNN) joint weighted $l_1$-norm to efficiently suppress the background and preserve the target.

In conclusion, the performance of these conventional methods were based on and limited by the expert prior on the characteristics of infrared small targets. However, due to the prior knowledge that these methods highly relied on, the conventional methods lacked generalization for detecting infrared small targets beyond the prior. A conventional method usually works well for a specific application, with limited generalization to other applications. Data driven is a potential technique to address this issue, which enables to learn features of infrared small targets from numerous samples, so that the generalization and effectiveness of detecting unknown infrared small targets beyond the prior are ensured.

Deep learning methods recently have attracted more and more attention on infrared small target detection [43]–[46]. Some works proposed deep learning frameworks decoupled with prior knowledge and experts analysis [15], [15], [47]–[53]. Specifically, these methods used global features of the infrared image extracted by designed neural networks to allocate a small target probability for each pixel. Wang et al. [14] focused on balancing the Miss Detection (MD) and False Alarm (FA), using two sub-tasks handled by two-stream models trained adversarially, with each stream focusing on reducing either MD or FA. Ju et al. [49] proposed an end-to-end network named ISTDet to detect the dim and small target. Motivated by the fact that the infrared small target has its unique distribution characteristics, Shi et al. [48] treated infrared small target as “noise” and transformed small target detection task into denoising problem. Zhao et al. [15] constructed a generative adversarial networks (GAN) model to learn the features of targets and directly predicted the intensity of targets, automatically. However, due to focusing on global features, these methods had to be assembled with numerous parameters for effectively extracting small target features from global features and highly relied on large amounts of data with discrepant scenes for learning features as much as possible.

Compared with these deep learning methods, the proposed method additionally explores local property of the infrared small target, which is conductive to simplify the task on extracting small target features and could effectively extract small target features from limited data with various scenes.

## III. The Proposed Method

The overall architecture of the proposed method is shown as Fig. 2. It mainly consists of three modules: global feature extractor, supervised attention module and patch net. The global feature extractor is used to extract basic features of the input image $I$ by viewing the whole image. Capturing these basic features is useful for reducing the complexity of the background. An attention-density (Atn-D) map $\hat{y}^a$ is calculated by the supervised attention module for the background suppression and the small target enhancement. The Atn-D map is applied on the global feature maps to get attention-global (Attn-G) maps. Based on a series of patches $p_n$, split from the Attn-G maps, local features correlated with small targets are extracted by the patch net to calculate patch-likelihood maps $\hat{y}^{p_n}$. By sharing the same convolution weights for all local patches, the task of small target detection is greatly simplified and the amount of parameters in the patch net can be limited. In each patch-likelihood map $\hat{y}^{p_n}$, the value of a pixel represents the probability that the pixel belongs to the background or the small target. Finally, a patch fusion strategy is applied to fuse these patch-likelihood maps and to get the final detection result by applying threshold segmentation on the fused likelihood map, i.e., confidence map $\hat{y}$.

### A. Global Feature Extractor

Due to the complexity of infrared images under different scenes, it is necessary to enable the network to capture basic features of the infrared image for simplifying the small target detection task. To avoid omitting small target features that usually take up only a few pixels, all of pooling layers are thrown away in the global feature extractor. The architecture of the global feature extractor is shown in Fig. 3. In this paper, the number $L_y$ of residual blocks used in this module is set to 4, and the number of convolution kernels for each block is 64, 128, 64 and 8, respectively. While a single-channel infrared
Fig. 2. The architecture of the proposed method. The global feature extractor aims to extract basic features of the input infrared image. The supervised attention module suppresses inconsequential distant background pixels to force the network to focus on the local saliency of small targets. The patch net utilizes a designed convolutional neural network (CNN) to extract multi-scale features from each patch.

Fig. 3. The architecture of the global feature extractor and the residual block. The input image as input $I$ is fed to the global feature extractor, the output data are several global feature maps as the same size as the input $I$.

B. Supervised Attention Module

Fig. 4. The architecture of the supervised attention module. The element-wise soft-max layer is a softmax active function applied on all elements to enhance regions of interest.

For the case of the complicated background such as bright clutters, most conventional methods could not perform well and state-of-the-art methods heavily depended on large amounts of data with heterogeneous scenes. In this paper, this issue is addressed through applying an attention module to suppress the background and enhance small targets. Through suppressing most irrelevant background pixels, the complexity of extracting small target features from local patches for the patch net can be greatly reduced, which is conducive to train the network on a small dataset with a large number of complex scenes. This is very important in practical applications, because it could be very difficult or expensive to collect infrared data for some specific scenes.

Although the attention module can be trained by the non-supervised way, it will take too much time and iterations to achieve the convergence. Thus, we use the supervised way to train the attention module by transformed the ground-truth. Due to the sparsity of small targets and the extreme class-imbalance between small targets and the background, it is difficult to directly train the attention module by the ground-truth. A feasible way is transforming the ground-truth by a low pass filter into a target spread map which focuses on the adjacent region of small targets. By this way, most background suppression and rough target enhancement can be achieved in the attention module, while the precise segmentation will be assigned to the patch-net.

Specifically, we utilize a 2D-Gaussian Low Pass Filter described as:

$$G = e^{-\frac{1}{2}(\frac{u^2+v^2}{\sigma^2})},$$

(1)

where the frequency filtering range is defined by extent parameter $\sigma$ and $u, v$ represent the 2-D components in the frequency domain. This filter is applied on the ground-truth in the frequency domain and the filtered result $f'$ is transformed back to the spatial domain by Inverse Discrete Fourier Transform (IDFT). Finally, the target spread map is calculated by:

$$m_t = \frac{f'}{||f'||},$$

(2)

where $m_t$ represents the target spread map which indicates some regions that the model should pay attention to and the pixel value of $m_t$ represents how much attention that the detector should pay. The architecture of the supervised attention module is shown in Fig. 4. Specifically, the number $L_a$ of residual blocks used in this module is set to 6, and the number of convolution kernels for each residual block is 32, 64, 128, 64, 32 and 1, respectively. The $1 \times 1$ convolution layer aims to allocate an probability for each element to get an attention-probability map. The element-wise soft-max layer is applied on attention-probability map to get the Atn-D map $\hat{y}^a$. This element-wise soft-max layer is helpful to enhance the regions of interest for accelerating the convergence in training.

After that, through multiplying the Atn-D map $\hat{y}^a$ with each
global feature map by element-wise, most background will be suppressed and thus small targets will be enhanced. The result of the element-wise multiplication between the Attn-D map \( \hat{y}^n \) and a global feature map is an Attn-G feature map.

### C. Patch Net

In the Attn-D map from the supervised attention module, there are many rough target possible regions which are not enough for acquiring the final segmentation due to the high FA rate in the Attn-D map. The patch-net is designed to reduce the FA rate and acquire the precise segmentation results by effectively extracting local features correlated with small targets. The architecture of the patch net is shown in Fig. 5.

In the patch net, each Attn-G feature map is split into a series of patches by using a sliding window. There are two circumstances in these patches: 1) A patch totally belongs to the background. Under this circumstance, the patch net is prone to transform this patch to a black block that all pixel values equal to zeros. This operation does not demand any computational complexity because these background pixels have been suppressed in the supervised attention module and already similar to zeros. 2) A patch includes features correlated with small targets. Under this circumstance, the patch net will focus on the local saliency of small targets because the sliding window has restricted the receptive field of the patch net on this patch, so that the precise result can be acquired. Besides, by sharing the same convolution weights for all patches, the number of parameters of the patch net can also be greatly reduced.

In this paper, we utilize a sliding window to extract patches from each Attn-G feature map. Inspired by [54], an inception module is used to extract multi-scale small target features from each patch. This inception module consists of several k-conv layers with different kernel sizes and each k-conv is composed of a convolution kernel, a batch norm layer and a ReLU function. Specifically, the number \( L_k \) of k-conv used in this module is set to 3, and the convolution kernel size of each k-conv is \( 1 \times 1 \), \( 3 \times 3 \) and \( 5 \times 5 \), respectively. Different kernel sizes could be conducive to extract multi-scale features of small targets with various sizes. These multi-scale features from each patch are fed to a sub-net for extracting a local feature map named patch-feature map in this paper. In the sub-net, each input is firstly scaled to double times through a deconvolution layer which consists of a \( 1 \times 1 \) deconvolution kernel, a batch norm layer and a ReLU function. The deconvolution layer aims to increase the size of small targets in each patch for expanding the perception range of the patch net on small targets. Then several residual blocks are used to extract local features from deconvoluted patches. The number \( L_p \) of these residual blocks used in this module is set to 3, and the number of convolution kernels in each residual block is 32, 32 and 1, respectively. A max pooling layer is utilized to resize each patch-feature map to the original size as the same as the input patch. Finally, a single-channel \( 1 \times 1 \) convolution layer and the sigmoid function are applied to convert each resized feature map to a patch-likelihood map, in which the pixel value represents the probability that the pixel belongs to a small target.

To get the final confidence map, a patch fusion is proposed to fuse all patch-likelihood maps. The patch fusion used in this paper is directly to calculate the mean value of overlapped regions of different patches, as shown in Fig. 5. The pixel values in the region \( C \) are the mean value of overlapped pixels from \( A \) and \( B \). Through this strategy, those pixels regarded by most patch-likelihood maps as small target pixels would be enhanced, so that the FA rate will be reduced.

An adaptive threshold \( t_{\text{adpt}} \) is used to segment the confidence map \( \hat{y} \) to get the final prediction result:

\[
t_{\text{adpt}} = \max \{ v_{\min}, \mu + k\sigma \},
\]

where \( k \) and \( v_{\min} \) are constants determined experientially. \( v_{\min} \) is to process the case with no target. If the max value of the confidence map is lower than \( v_{\min} \), it means that there is no target in the predicted result. \( \mu \) and \( \sigma \) are the mean value and standard deviation of the confidence map \( \hat{y} \). Therefore, a pixel at \((i, j)\) can be segmented as the target pixel if \( \hat{y}(i, j) \geq t_{\text{adpt}} \), otherwise it is a background pixel.
D. Loss Function

The proposed method can be trained in an end-to-end fashion. We use the infrared image $I$, its ground-truth image $y$ and its target spread map $m_t$ calculated by Eq. (2) to train the proposed method.

The attention loss is calculated by the sum-square error (SSE) between the target spread map $m_t$ and the Atm-D map $\hat{y}^a$ of the supervised attention module, defined as follows:

$$loss_a = \|\hat{y}^a - m_t\|^2_2.$$  

The total loss function is defined as:

$$loss = loss_a + \frac{1}{N}\sum_{n=1}^{N}loss_{pn},$$

where $N$ is the number of patches.

IV. EXPERIMENTS

In this section, we firstly introduce evaluation metrics, state-of-the-art methods used for comparison and experimental settings of all tested methods. Then the quantitative comparisons with state-of-the-art methods are performed on two public widely used datasets and one private dataset collected by ourselves, respectively. Qualitative results are also given for analysis.

The effect of each module, operation and hyperparameter of the proposed method are discussed through a series of ablation studies.

A. Experimental Setup

1) Evaluation Metrics: In early stages, some evaluation metrics widely used in the small target detection community [8], [25], [55] are based on target-level. In these metrics, a connected domain in the detection result will be regarded as a true-positive result if (1) the connected domain has overlapped pixels with the ground-truth and (2) the center pixel distance of the connected domain and the ground-truth is within a threshold (usually 4 pixels) [8]. Additionally, some pixel-level evaluation metrics are also widely used in deep learning based methods [13]–[15], [48]. Generally, these pixel-level metrics are more rigorous than target-level metrics when evaluating the performance of a method. However, due to the inevitable artificial errors when labeling the ground-truth images, the pixel-level metrics are not enough to provide comprehensive evaluations as well. Thus, we use both target-level metrics and pixel-level metrics to comprehensively evaluate all methods.

Specifically, the metrics used for comparison between the proposed method and state-of-the-art methods are as follows:

- The probability of detection ($P_d$) and false alarm rate ($P_a$): These two target level metrics are widely used to evaluate the method on infrared small target detection, which are defined as [8]:

$$P_d = \frac{# \text{number of true detections}}{# \text{number of real targets}},$$

$$P_a = \frac{# \text{number of false detections}}{# \text{number of images}}.$$  

We use $P_d$ with $P_a = 0.2$/image as [8] to evaluate the performance of each method and the area under $P_d$ curve (AUC) with $P_a \leq 2.0$/image to evaluate average performance of each method.

- Target-level precision, recall and f1 measure ($F_1T$): We also use the target-level precision, recall and f1 measure ($F_1T$) to evaluate the proposed method and state-of-the-art methods. The highest $F_1T$ of each method is provided to display the best performance, and corresponding precision and recall are provided for analysis.

- Pixel-level precision, recall and f1 measure ($F_1P$): For providing comprehensive comparison, we utilize pixel-level metrics to evaluate each method as well. As the same as deep learning methods [13]–[15], [48], the highest $F_1P$ and corresponding precision and recall are provided.

2) State-of-the-Art Methods: We compare the proposed method with two groups of related methods:

- Conventional methods: Top-Hat [1], [2], [22], Max-Mean/Max-Median [23], [24], AAGD [3], ADM [4], GST [25], LIG [7], ILCM [33], MPCM [34], TLCCM [5], LEF [6], IPI [8].

- Deep learning method: MDvsFA-cGAN [14], ACM [56]. These two methods achieve excellent performance to the best of our knowledge so that are used to represent the state-of-the-art deep learning methods.

3) Experimental Setting: The parameter settings for all tested methods are listed in Tab. 1. The experiment is conducted on a computer with one 3.40GHz CPU, 32GB RAM and two NVIDIA TITAN V GPUs. All trainable methods are trained from the scratch, and the batch size is set to 20 for training. The proposed method is implemented by Python and PyTorch. Specifically, the input image is resized to $120 \times 120$, the patch size and sliding step are set to $30 \times 30$ and 10, respectively. Adam algorithm [57] is used as optimizer and learning rate is set to 0.001 for all batches.

B. Quantitative Analysis

To evaluate the performance of the proposed method, we use three datasets with differences, including a synthetic dataset namely MFIRST used in [14], a widely used SIRST dataset [56] and a SeqIRST dataset which consists of seven infrared sequences collected by ourselves.

Some representative images of these three datasets are shown in Fig. 6, while the training and testing splittings of these three datasets are listed in Tab. 1.
TABLE I
PARAMETER SETTING FOR TESTED METHODS.

| Methods          | Key parameter settings                  |
|------------------|-----------------------------------------|
| Top-Hat          | structure size: 12 × 12                  |
| Max-Mean/Max-Median | filter size: 15 × 15                   |
| AAGD             | scale size: 3, 5, 7, 9                  |
| ADMD             | scale size: 3, 5, 7                     |
| LIG              | k = 0.2, N = 11                          |
| IPI              | patch size: 50 × 50, sliding step: 10,  |
|                  | λ = (1/√max(m, n)), ε = 10^{-6}         |
| ILCM             | subblock size: 8 × 8, moving step: 4    |
| MPCM             | scale size: 3, 5, 7                     |
| TLLCM            | gaussian kernel size: 3 × 3,            |
|                  | scale size: 3, 5, 7                    |
| LEF              | α = 0.5, h = 0.2                         |
| GST              | σ_1 = 0.6, σ_2 = 1.1, boundary width: 5,|
|                  | filter size: 5 × 5                      |
| MDvsFA-cGAN      | image size: 128 × 128, λ_1 = 100, λ_2 =1,|
|                  | trainable parameters: 391998            |
| ACM              | image size: 256 × 256, backbone: fpn,  |
|                  | fuse: asymbi, trainable parameters: 387187|
| Ours             | image size: 120 × 120,                  |
|                  | patch size: 30 × 30, sliding step: 10,  |
|                  | trainable parameters: 925108            |

Fig. 6. Representative images from (a) MFIRST, (b) SIRST and (c) SeqIRST. Small targets to segment are indicated by red circles. Single-frame infrared images in the SIRST and infrared sequences in the SeqIRST are collected under real scenes, while single-frame infrared images are synthesized through coding in the MFIRST.

1) MFIRST dataset: The MFIRST is a dataset containing a large number of real and synthetic infrared images. The real infrared images come from two bespoke datasets containing 11 real infrared sequences with 2098 frames in total and 100 real individual infrared images with different small objects, respectively [14]. In the synthetic infrared images, the backgrounds are generated by cropping different regions from infrared high-resolution natural scene images and the small targets are separated from the real infrared images or synthesized by using the 2-D Gaussian function [14]. We use this dataset to evaluate the performance of each method on augmented data with a large number of different scenes. The comparison results on the MFIRST dataset are listed in Tab. III.

2) SIRST dataset: The SIRST dataset is a widely used public dataset for single-frame infrared small target detection [14]. It contains 427 representative images and 480 instances of different scenes from hundreds of real-world videos. We use this dataset to evaluate the performance of each method on limited but authentic data with a large number of different real scenes. The comparison results on the SIRST dataset are listed in Tab. IV.

On the MFIRST dataset, due to the large amount of data, it is hard for prior based conventional methods to be generalized enough for discrepant scenes so that these methods perform much worse than data-driven methods. Compared with the deep learning method, the proposed method outperforms MDvsFA-cGAN approximately 2% on AUC, and outperforms ACM approximately 4% on both metrics. Although the P_d, F1_T and F1_P metrics of the proposed method are similar with MDvsFA-cGAN, the higher AUC illustrates the better average performance of the proposed method. By the attention mechanism and the patch splitting, the proposed method addresses the class-imbalance of small target pixels and background pixels. Focusing on the local saliency of small targets enables the proposed method to capture more effective features of small targets for achieving robust detection results.
Compared with these conventional methods, even if no-prior methods also perform much worse than no-prior methods. The reason might be that when small targets in a sequence are too dim, a conventional method will fail to detect targets in all frames if it fails in a single frame. The performance of each method on detecting infrared small targets under two scenes. We use this dataset to evaluate the effectiveness of the attention mechanism and the patch splitting on effectively extracting small target features when being trained on limited data for deep learning methods, as described in Section I.

3) SeqIRST dataset: The SeqIRST dataset contains seven infrared sequences which are collected by ourselves and the details are listed in Tab. VI. It contains three types of small targets under two scenes. We use this dataset to evaluate the performance of each method on detecting infrared small targets from continuous frames. The comparison results are listed in Tab. VI.

On the SeqIRST dataset with small amount of data, prior based conventional methods also perform much worse than no-prior methods. The reason might be that when small targets in a sequence are too dim, a conventional method will fail to detect targets in all frames if it fails in a single frame. Compared with these conventional methods, even if no-prior methods fail to detect a small target in one frame, they could still detect it in other frames since decoupling with the prior correlated with the specific scene. Compared with the deep learning method, the proposed method outperforms MDvsFA-cGAN approximately 40% on AUC, and outperforms ACM approximately 14% on AUC. These similar Pd, F1T, F1P and much better AUC on the SeqIRST demonstrate the superiority of the proposed method on average performance again.

C. Qualitative Analysis

We give some representative processed results of the proposed method and six methods with top performance, as shown in Fig. 7.

It can be seen from Fig. 7 that: (1) When detecting the dim target at the first row, most existing methods fail to detect the target and erroneously regard several cloudy clutters of the background as targets. (2) When detecting the small target with low signal-to-clutter ratio (SCR), most tested methods have a large number of FA and IPI module fails to detect the target. (3) When detecting the small target from the infrared image with bright clutters, most existing methods are interfered by the bright clutters and have a lot of FA in the clutter regions. Under above three mainly representative obstacles of infrared small target detection, the proposed method has has the least FA and MD. Besides, when facing to other general obstacles of the object detection, such as the high noise level at the fourth row, high boundary contrast of background at the fifth row, gaussian blurring at the sixth row and detecting multiple targets at the last row, the proposed method also possesses detection results with the least FA and MD among tested methods.

D. Ablation Study

1) Module Discussion: To investigate the effect of each modul in the proposed method, we conduct an ablation study on the MFIRST dataset. This MFIRST dataset is the largest dataset with numerous heterogeneous scenes among three datasets, which is crucial for validating the comprehensive performance of each ablation block. As listed in Table VII the AttnNet consists of the global feature extractor, the supervised attention module and a single convolution kernel used to approximately 3% on Pd, 7% on AUC and F1T, and 3% on F1P. The state-of-the-art performance of the proposed method demonstrates the effectiveness of the attention mechanism and the patch splitting on effectively extracting small target features when being trained on limited data for deep learning methods, as described in Section I.

| Methods         | SIRST | SeqIRST |
|-----------------|-------|---------|
| Top-Hat [2]     | 85.34 | -       |
| Max-Mean/Max-Median [24] | 78.46 | -       |
| AAGD [3]        | 89.09 | -       |
| ADMD [4]        | 94.13 | -       |
| LIG [5]         | 90.19 | -       |
| IPI [6]         | 86.87 | -       |
| ILCM [33]       | -     | -       |
| MPCM [54]       | 93.56 | -       |
| TLLCM [75]      | 61.61 | -       |
| LEF [85]        | -     | -       |
| GST [95]        | 77.01 | -       |
| ACM [96]        | 93.67 | -       |
| Ours            | 97.40 | 97.86   |

1" means that the method can not get reasonable values under fixed Fa = 0.2/image for Pd or under Fa ≤ 2.0/image for AUC. MDvsFA-cGAN [14] does not get reasonable values in our experiment due to the failure of convergence on such small dataset.

| No.          | Target | Background | Frames |
|--------------|--------|------------|--------|
| Seq. 1       | plane  | sky (cloudy)| 30     |
| Seq. 2       | plane  | sky (cloudy)| 39     |
| Seq. 3       | plane  | sky        | 39     |
| Seq. 4       | bird   | sky (cloudy)| 123    |
| Seq. 5       | vessel | sea        | 100    |
| Seq. 6       | plane  | sea        | 109    |
| Seq. 7       | vessel | sea        | 100    |

| Methods         | Pd (%) | AUC (%) | F1T (%) | F1P (%) |
|-----------------|--------|---------|---------|---------|
| Top-Hat [2]     | 85.34  | 82.38   | 82.52   | 44.13   |
| Max-Mean/Max-Median [24] | 78.46  | 77.45   | 73.49   | 23.97   |
| AAGD [3]        | 89.09  | 88.14   | 84.69   | 50.27   |
| ADMD [4]        | 94.13  | 90.46   | 88.50   | 56.69   |
| LIG [5]         | 90.19  | 90.00   | 89.72   | 59.15   |
| IPI [6]         | 86.87  | 84.45   | 85.32   | 56.97   |
| ILCM [33]       | -      | 47.26   | 0.71    |
| MPCM [54]       | 93.56  | 90.40   | 86.96   | 58.59   |
| TLLCM [75]      | 61.61  | 79.14   | 79.66   | 7.60    |
| LEF [85]        | -      | -       | 57.60   | 2.45    |
| GST [95]        | 77.01  | 76.81   | 80.40   | 35.32   |
| ACM [96]        | 93.67  | 86.93   | 95.85   | 69.61   |
| Ours            | 97.40  | 97.86   | 96.40   | 71.66   |

| Methods         | Pd (%) | AUC (%) | F1T (%) | F1P (%) |
|-----------------|--------|---------|---------|---------|
| Top-Hat [2]     | -      | 58.90   | 61.20   | 10.35   |
| Max-Mean/Max-Median [24] | -      | -       | 23.95   | 5.33    |
| AAGD [3]        | -      | 56.51   | 60.47   | 17.72   |
| ADMD [4]        | 47.19  | 67.08   | 69.01   | 17.27   |
| LIG [5]         | -      | 57.59   | 60.67   | 20.99   |
| IPI [6]         | 76.42  | 81.54   | 78.45   | 30.95   |
| ILCM [33]       | -      | -       | 24.72   | 1.05    |
| MPCM [54]       | 74.43  | 77.49   | 78.89   | 17.04   |
| TLLCM [75]      | 60.44  | 63.33   | 66.26   | 10.67   |
| LEF [85]        | -      | 55.68   | 56.65   | 11.23   |
| GST [95]        | -      | -       | 47.06   | 14.06   |
| ACM [96]        | 96.33  | 57.00   | 98.88   | 89.68   |
| Ours            | 97.03  | 98.50   | 99.99   | 89.72   |

"-' means that the method can not get reasonable values under fixed Fa = 0.2/image for Pd or under Fa ≤ 2.0/image for AUC.
allocate probability for each pixel. It is trained by the target spread map with the SSE loss. The PatchNet consists of the global feature extractor and the patch net, which is trained by the ground-truth with the BCE loss.

According to the definition, precision is negatively correlated with FA and recall is negatively correlated with MD, which means that the higher precision represents the lower FA and the higher recall represents the lower MD. In AttnNet, the precision is higher than recall, which demonstrates that the attention module has better performance on reducing FA. In PatchNet, the recall is higher than precision, which demonstrates that the patch net module has better performance on reducing MD. Compared with the AttnNet and PatchNet, our method has both highest precision and recall, which demonstrates that the combination of the attention mechanism and the patch net module is conductive for reducing both FA and MD. Besides, the increment of recall from the AttnNet (without the patch net module) to our method is much more than precision, which illustrates that the patch net module has more positive effects on capturing those small targets that might be omitted. The increment of precision from the PatchNet (without the attention module) to our method is much more than recall, which illustrates that the attention module has more positive effects on ensuring the final accuracy of detecting.

**Fig. 7.** The representative processed results of different methods. The first row shows results of detecting a dim small target. The second row shows results of detecting a small target with low signal-to-clutter ratio (SCR), i.e. the small target buried in clutters of the background. The third row shows results of detecting a small target from a image with bright clutters. Compared with the second row, the small target in the third row is not buried in clutters but these clutters consist of a series of discrete areas with bright pixels. The fourth row shows results of detecting the small target image with high noise level. The fifth row shows results of detecting the small target image with high background boundary contrast. The sixth row shows results of detecting the small target image with gaussian blurring. The last row is results of detecting the small target image with multiple targets.

**TABLE VII**

| Module Discussion on MFIRST Dataset. |
|--------------------------------------|
| Methods | Target Level | Pixel Level |
|         | Prec. (%) | Rec. (%) | F1 (%) | Prec. (%) | Rec. (%) | F1 (%) |
| AttnNet1 | 62.96 | 48.57 | 54.84 | 39.40 | 38.89 | 39.14 |
| PatchNet2 | 38.25 | 78.42 | 51.42 | 31.27 | 46.27 | 37.32 |
| Ours3 | 86.11 | 88.57 | 87.32 | 52.06 | 71.00 | 60.07 |

1 AttnNet combines a global feature extractor, a supervised attention module and a single convolution kernel for allocating probability for each pixel. The supervised attention module is trained by the target spread map calculated by Eq. 2 with the SSE loss.

2 PatchNet contains a global feature extractor and a patch net. It is trained by the ground-truth with the BCE loss.

3 Our method is trained by the ground-truth with the BCE loss and the target spread map with the SSE loss jointly.
to capture basic features by viewing the whole image. The supervised attention module enables the network to suppress background irrelevant with small target features for effectively simplify the task for subsequent modules. The patch net enables the network to focus on the local saliency of small target features based on decomposed local patches. Extensive synthetic and real data experiments show that compared with state-of-the-art methods, the proposed method performs better on detecting infrared small targets with complicated scenes, such as dim targets, bright clutters and low SCR. Besides, the proposed method has higher robustness and comprehensive performance than the compared deep learning method. In the future, we will generalize the proposed method from single-frame detection to multi-frame detection by synthesizing spatial and temporal features of infrared small target videos.

### References

[1] Xiangzi Bai and Fugen Zhou. Analysis of new top-hat transformation and the application for infrared dim small target detection. Pattern Recognition, 43(6):2145–2150, 2010.

[2] Ming Zeng, Jianxun Li, and Zhang Peng. The design of top-hat morphological filter and application to infrared target detection. Infrared Physics & Technology, 48(1):67–76, 2006.

[3] Saeed Aghaeei, Saed Moradi, and Hasan Talebi. Small infrared target detection using absolute average difference weighted by cumulative directional derivatives. Infrared Physics & Technology, 101:78–87, 2019.

[4] Saed Moradi, Payman Moallem, and Mohamad Farzani Sabahi. Fast and robust small infrared target detection using absolute directional mean difference algorithm. Signal Processing, 177:107727, 2020.

[5] Jinhui Han, Saed Moradi, Iman Faramarzi, Chengyin Liu, Honghui Zhang, and Qian Zhao. A local contrast method for infrared small-target detection utilizing a tri-layer window. IEEE Geoscience and Remote Sensing Letters, 17(10):1822–1826, 2020.

[6] Chaofan Xia, Xiaorun Li, Liaoying Zhao, and Rui Shu. Infrared small target detection based on multiscale local contrast measure using local energy factor. IEEE Geoscience and Remote Sensing Letters, 17(1):157–161, 2019.

[7] Hong Zhang, Lei Zhang, Ding Yuan, and Hao Chen. Infrared small target detection based on local intensity and gradient properties. Infrared Physics & Technology, 89:88–96, 2018.

[8] Chengqiang Gao, Deyu Meng, Yi Yang, Yongtao Wang, Xiaofang Zhou, and Alexander G Hauptmann. Infrared patch-image model for small target detection in a single image. IEEE Transactions on Image Processing, 22(12):4996–5009, 2013.

[9] Chengqiang Gao, Lan Wang, Yongxing Xiao, Qian Zhao, and Deyu Meng. Infrared small-dim target detection based on markov random field guided noise modeling. Pattern Recognition, 76(Supplement C):465–475, 2018/04/01/ 2018.

[10] Huan Wang, Fei Yang, Congcong Zhang, and Mingwu Ren. Infrared small target detection based on patch image model with local and global analysis. International Journal of Image and Graphics, 18(01):1850002, 2018.

[11] Tianfeng Zhang, Hao Wu, Yuhua Liu, Lingbing Peng, Chunting Yang, and Zhenming Peng. Infrared small target detection based on non-convex optimization with lp-norm constraint. Remote Sensing, 11(5):559, 2019.

[12] Lanban Zhang and Zhenming Peng. Infrared small target detection based on partial sum of the tensor nuclear norm. Remote Sensing, 11(4):382, 2019.

[13] Bin Zhao, Chunting Wang, Qiang Fu, and Zishuo Han. A novel pattern for infrared small target detection with generative adversarial network. IEEE Transactions on Geoscience and Remote Sensing, 2020.

[14] Huan Wang, Luping Zhou, and Lei Wang. Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 8508–8517. IEEE, 2019.

[15] Mingxin Zhao, Li Cheng, Xa Yang, Peng Feng, Liyuan Liu, and Nanjian Wu. Tbc-net: A real-time detector for infrared small target detection using semantic constraint. arXiv preprint arXiv:2001.05852, 2019.

### Table VIII

| Hyperparameters | Target Level | Pixel Level |
|-----------------|--------------|-------------|
| Patch Step      | Prec. (%)    | Rec. (%)    | F1 (%)    |
| 20 x 20 10      | 85.93        | 82.86       | 84.36     | 42.81 | 60.72 | 50.22 |
| 30 x 30 10      | 86.11        | 85.57       | 87.32     | 52.06 | 71.00 | 60.07 |
| 30 x 30 15      | 84.92        | 78.10       | 81.37     | 45.98 | 64.59 | 53.72 |
| 40 x 40 20      | 72.80        | 66.91       | 69.73     | 37.45 | 54.89 | 44.52 |

### Table IX

| Methods        | Time (s/100 images) | GPU Memory (MB) |
|----------------|---------------------|-----------------|
| Top-Hat [2]    | 1.78                | -               |
| Max-Mean/Max-Median [24] | 1.50            | -               |
| AAGD [3]       | 3.52                | -               |
| ADM [4]        | 2.02                | -               |
| LIG [7]        | 70.44               | -               |
| IPI [8]        | 424.60              | -               |
| ILCM [33]      | 1.92                | -               |
| MPCMT [74]     | 4.60                | -               |
| TLCCM [5]      | 321.91              | -               |
| LEF [6]        | 430.22              | -               |
| GST [5]        | 1.05                | -               |
| MDvs-FC-C-GAN [14] | 10.62            | 432.98         |
| ACM [56]       | 1.61                | 82.15           |
| Ours           | 37.43               | 363.11          |

‘-‘ means that the method does not possess trainable parameters. The smallest and the second smallest values are shown in boldface.

2) Hyperparameter Discussion: Besides, we conduct a hyperparameter experiment to illustrate the impact of the patch size and the sliding step on the proposed method, as listed in Tab. VIII. With the increasing of the patch size and the sliding step, the performance of the proposed method will be worse and worse because the local property of small targets will lose under a larger and larger patch size, and the patch fusion will also lose effect under a too large sliding step. However, a too small patch size also causes the performance degradation. This might be because the inception module of the patch net could not capture enough multi-scale features on patches with too small size.

3) Consumption Discussion: A time and memory consumption experiment is also conducted for evaluating the practicability of each method, as shown in Tab. IX.

Although the time consumption of the proposed method does not perform best among all tested methods, it could be further reduced by parallelizing the patch splitting that the most time consuming module of the proposed method. However, it can be seen from Tab. IX that the proposed method still has a receivable time and memory consumption compared with most state-of-the-art methods, which ensures the practicability.

V. Conclusion

A new deep learning framework decoupled with prior for infrared small target detection is presented in this paper. The global feature extractor enables the proposed method
