# Mask-Guided Infrared Small Multi-Target Detection via Coarse-to-Fine Candidate Selection

**Tianyi Lu** ( lutianyi@shu.edu.cn )
Shanghai University

**Junjie Zhang**
Shanghai University https://orcid.org/0000-0002-0033-0494

**Yi Lin**
Shanghai University

**Dan Zeng**
Shanghai University

**Research Article**

**Keywords:** Infrared target, Multi-target detection, Mask-Guided feature extraction, Candidate selection

**Posted Date:** June 16th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1554001/v1

# Mask-Guided Infrared Small Multi-Target Detection via Coarse-to-Fine Candidate Selection

Tianyi Lu[1], Junjie Zhang[1*], Yi Lin[1] and Dan Zeng[1]

[1*]Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute of Advanced Communication and Data Science, Shanghai University, ShangDa road 99, Shanghai, 200444, China.

*Corresponding author(s). E-mail(s): junjie_zhang@shu.edu.cn;
Contributing authors: lutianyi@shu.edu.cn; liny@shu.edu.cn;
dzeng@shu.edu.cn;

**Abstract**

Infrared acquisition technology is often used for target detection in military fields, since it is not easily disturbed by diverse environmental factors. However, the characteristics of low image contrast, coupled with complex imaging backgrounds, long distance and the lack of visual features, making infrared small targets detection a challenging task. The generic deep neural network-based models are struggled to be applied for infrared small targets, since the increased network layers may lead to the gradually loss of target features and positional information. To address above issues, we propose a mask-guided detection model via the coarse-to-fine candidate selection for infrared small multi-target detection. More specifically, to enhance target features and guide the localization process in the neural network, we propose to utilize the foreground mask generated by referring to the non-local self-correlation property of infrared background and the sparse property of target distribution. The obtained mask is treated as the prior to re-weight the convolutional feature maps. Considering the complex background, the multi-target detection is prone to mis-detections. Therefore, we propose a coarse-to-fine candidate selection method on top of the initial detection results. A shallow network is constructed to extract more nuanced visual features from the candidate


**Fig. 1** The examples of infrared small targets. The part of the target areas is enlarged in the top-left corner for better visualization. The SCR is marked next to the bounding box.

positions for a binary classification, in which the false positive candidates are ruled out free from the disruptions of other background features. Moreover, given the lack of multi-target infrared datasets, we propose two synthetic datasets based on the public available and own collected infrared data. Extensive experimental results verify the effectiveness and advantages of our model compared to state-of-the-art methods.

**Keywords:** Infrared target, Multi-target detection, Mask-Guided feature extraction, Candidate selection

# 1 Introduction

Given the infrared acquisition is mainly determined by the temperature of the target, the infrared detection technology has the advantage of not being easily interfered by environmental factors, and has been widely used in the infrared guidance, early warning and other military fields. Apart from its advantages against the generic imaging, the infrared target detection poses following unique challenges. First of all, the visual features of infrared targets are deficient. On the one hand, with the long imaging distance, the radiation energy of infrared target is often smaller than the background, which only occupies a few pixels in the entire image, lacking the texture or shape features [1]. On the other hand, due to the acquisition of the infrared imaging, the most commonly used color features are unavailable for these small-sized targets. Secondly, the Signal-to-Clutter Ratio (SCR) of target is low. Affected by the long imaging distance and diverse backgrounds, small targets often posses similar characteristics as the background clutter and noise, leading to the low intensity and SCR, which are easily submerged in unpredictable signals. As shown in Fig. 1, we can observe that the infrared targets are small sized with the low SCR in diverse application scenarios, and the above phenomena are much more serious with the multi-target cases, making the infrared small multi-target detection a challenging task.

The single-frame infrared target detection, as the foundation of its multi-frame counterpart and various infrared vision applications, has been extensively studied for decades [2, 3], which can be roughly grouped into the traditional and deep neural network-based methods. For traditional ones, filter-based methods [4–7] are one of the earliest methods used for small target detection. They aim at utilizing the frequency difference between the target and background to distinguish the target. For example, Seyed *et. al.* [8] proposed to utilize the structural elements based on the genetic algorithm to suppress background clutter and noise to detect targets. A non-subsampled contourlet transform model combined with SVD was proposed in [9] to adjust coefficients through singular values to make the infrared target protuberant. Different from them, local contrast-based methods [10–15] use the brightness differences of the target and neighboring areas to detect small targets. Chen *et. al.* [16] proposed to enhance the target signal and suppress the background simultaneously by referring to the local contrast. Han *et. al.* [17] designed a matched filter to enhance the true target purposefully before the computation of local contrast to improve the detection. Given infrared targets are often presented in a sparse distribution, and the background has low-rank characteristics, the low-rank sparse recovery-based methods [18, 19] have been proposed, of which the Infrared Patch-Image (IPI) model [20] and its variations [21–23] are the most widely applied models. The IPI model focuses on modeling the characteristic of non-local self-correlation property of the background, and separating the features of sparse targets by the sliding window. Through refactoring the image, the properties of each local patch are enhanced. Motivated by the IPI, the Weighted Infrared Patch-Image model (WIPI) [24], the Re-weighted Infrared Patch-Tensor model (RIPT) [25], and the improved optimization of IPI [26] were consecutively proposed from different perspectives to enhance the original IPI.

Since traditional methods generally rely on the hand-crafted features and have relatively high computation complexity, with the advanced representation learning ability of deep learning, various neural network-based models [27, 28] have been proposed for the infrared target detection. For example, Liu *et. al.* [29] proposed the first CNN-based model using a five-layer Multi-Layer Perception (MLP) for infrared small target detection. McIntosh *et. al.* [30] employed the eigen-vectors as the input, and further fine-tuned multiple generic object detection networks. To address the problem of lacking visual feature and class information during training, Hou *et.al.*[31] proposed a TBC-Net, which contains a semantic constraint module to count the target number as an auxiliary task to assist the detection. Fan *et.al.*[32] proposed to enhance the target intensity by its local intensity characteristics, and combined the corner detection with CNN to locate the infrared target. A Multi-patch Attention Network (MANet) is proposed in [33], where the global and local properties of infrared small targets are jointly considered to suppress the background pixels and capture the target locality. Different from prior works, Shi *et.al.* [34] proposed an end-to-end detection framework based on denoising autoencoder,

where small targets are treated as the noise. Similarly, Ju *et.al.* [35] proposed ISTDet using image filtering modules to enhance the response of target. Dai *et. al.* [36] proposed the segmentation-based detection framework with the asymmetric contextual module to improve the local contrast of target from multiple scales. Li *et. al.* [37] proposed a Dense Nested Attention Network (DNANet) to handle the problem of the loss of targets in deep layers via densely connected interactive modules among low-level and high-level features. Moreover, Generative Adversarial Network (GAN) is also applied for this task. Zhao *et. al.* [38] proposed to focus on the essential features of infrared small target in an adversarial learning manner. Wang *et. al.* [39] proposed a deep adversarial learning framework by training two adversarial models to reduce the miss-detection and false alarm simultaneously.

Though existing infrared target detection models, especially the deep neural network-based ones, have achieved promising results on publicly available infrared datasets, the increased layers of deep neural network may lead to the gradually loss of target features and positional information, causing smaller targets difficult to be detected. Moreover, these methods often focus on detecting single or a few targets in each frame (normally two targets), the more challenging case of multi-target detection are yet to be explored. With the complex backgrounds, the probability of false alarm can occur greatly as the number of targets increases. Therefore, in this paper, we focus on tackling the infrared small multi-target detection, where multiple smaller and dimmer targets presented in each frame compared to existing cases.

To address above issues, we propose a Mask-Guided detection model via Coarse-to-Fine Candidate Selection for infrared small multi-target detection. More specifically, considering the non-local self-correlation property of the infrared background and the sparsity of targets, we obtain the rough foreground mask by recovering the low-rank and sparse matrices of the infrared image. The foreground mask is then used to re-weight the convolutional feature map, which guides the network more focus on the small targets. To further eliminate the increased falsely detected regions, we treat them as candidates, and design a shallow network to extract the nuanced features from them. A binary classifier is applied as a coarse-to-fine candidate selection process to obtain final results. Overall, the main contributions of this paper can be summarized as follows:

1. We propose a mask-guided neural network for infrared small multi-target detection. By recovering the low-rank and sparse matrices of the infrared background and targets, the obtained foreground mask can guide the network more focus on the extraction of small target features;
2. We design a coarse-to-fine candidate selection process to reduces false alarms in multi-target detection, where the initial detected regions are treated as candidates, and nuanced visual features can be extracted from them for further classification;

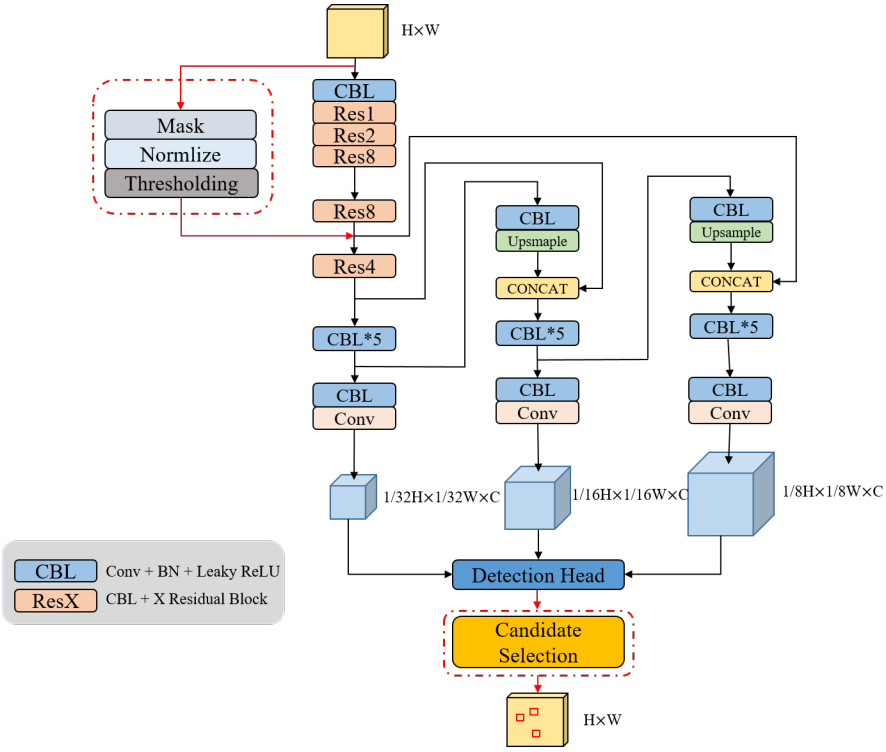



**Fig. 2** The overall architecture of the proposed model. The feature extraction of the infrared image is guided by the generated mask to concentrate more on the nuanced small target areas. The initial candidate regions are further screened by the proposed coarse-to-fine selection process to obtain the final detection results.

3. We propose two synthetic datasets based on the publicly available and own collected infrared data. Extensive experiments on these datasets demonstrate the advantages of our method compared against state-of-the-art models.

The remainder of the paper is organized as follows: in Section 2, we describe the details of our model and conduct thorough analysis. In Section 3, we present the experimental comparisons on the proposed datasets. Last, conclusions are given in Section 4.

# 2 Methods

## 2.1 Overall Architecture

The overall architecture of the proposed model is shown in Fig. 2. We take the infrared image as its input and sequentially performs the feature extraction, the initial detection, and the candidate selection. For the feature extraction, a foreground mask generated by recovering the low-rank and sparse matrices

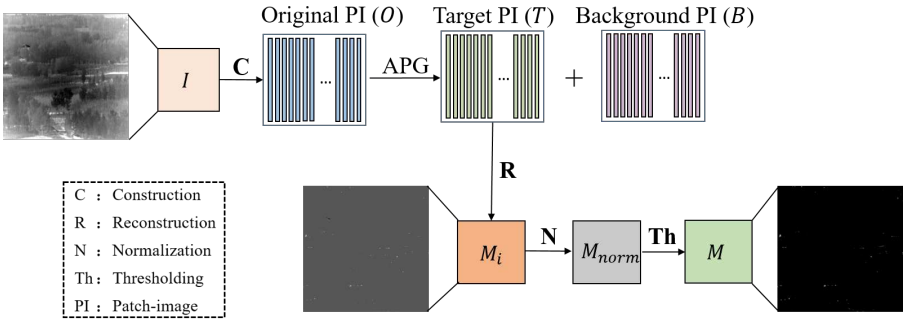



**Fig. 3** The illustration of Mask-Guided feature extraction module. The input infrared image $I$ is constructed and reconstructed to generate the initial mask $M_i$, and the final mask $M$ is obtained by normalization and thresholding sequentially.

of the infrared image is utilized to guide the extraction process to emphasize small targets for further detection. With the initial detection results as candidates, we design a coarse-to-fine candidate selection process, the candidates are screened by the shallow visual features of the corresponding original image regions to reduce the false alarm rate of the network. We employ a light-weighted CNN classifier to determine whether the candidate is the real target or not.

## 2.2 Mask-Guided Feature Extraction

Mask-Guided feature extraction aims to generate a foreground mask that is instructive for the detection of small targets in infrared images by enhancing target features. Considering the sparsity of targets, as they generally occupy only a few pixels with the scattered distribution, we propose to adopt the Infrared Patch-Image (IPI) model [20] to generate the foreground mask. Given the sparsity of the target distribution and the low rank property of infrared backgrounds, the problem of detecting small targets converts to the optimization problem of recovering low rank and sparse matrices of the original image. The Accelerated Proximal Gradient (APG) Algorithm [40] is thus employed for the optimization.

More specifically, as shown in Fig. 3, we obtain image patches through a series of sliding windows from the top-left to the right-bottom, and reformulate them as a new patch-image by vectoring each patch into a column vector. After the construction, we can treat each patch-image as:

$$O = B + T \tag{1}$$

where $O$ is the constructed patch-image, $B$ and $T$ are the background and target patch-image that need to be obtained respectively. Since the proportion of targets is relatively small for the whole image, the target patch-image $T$ thus can be regarded as a sparse matrix. The background pixels of the infrared image are often correlated with each other in a distant position, *i.e.,* possess

the non-local self-correlation property. Based on this characteristic of the background, the backgournd patch-image $B$ can be treated as a low-rank matrix. The target patch-image $T$ and background patch-image $B$ can be simultaneously estimated by APG Algorithm, and the sparse target patch-image $T$ is reconstructed as the initial foreground mask.

Given the initial mask aims at highlighting the sparse component of the original image, to further improve the accuracy of the mask, we obtain the final mask by firstly normalizing the initial mask and multiplying the original input image:

$$M_{norm}(x, y) = N(M_i(x, y)) \times I(x, y) \tag{2}$$

where $x$ and $y$ represent the spatial index of position. $M_i$, $I$ indicate the initial mask and the input image. $N(\cdot)$ stands for the min-max normalization. Then, we threshold the normalized output $M_{norm}$. The threshold $t$ is determined by:

$$t = \mu + k\sigma \tag{3}$$

where $\mu$ and $\sigma$ are the mean value and standard deviation of the target image, respectively, $k$ is the weight to balance two values, by referring to [20], we empirically set it to three. We consider a pixel belongs to the target if its value is above the threshold, otherwise, it is a background pixel. We set the background pixel to zero and keep the target pixel as itself. After the thresholding, we obtain the final mask $M$.

The obtained mask $M$ contains the positions of targets and other target-liked background pixels. We use this mask to guide the feature extraction process, namely enhancing the masked part of the convolutional feature map to guide the network to focus on the positions of potential targets as bellow:

$$F^{'}(x, y) = F(x, y) \times M(x, y) + F(x, y) \tag{4}$$

where $F$ represents the extracted convolutional feature map. The mask is firstly resized to the same size as the feature map $F$. Then, it multiples the feature map and add on it. This is to highlight the features of potential target areas. The processed feature map $F^{'}$ is utilized for subsequent network operations.

## 2.3 Coarse-to-Fine Candidate Selection

Since false alarms are inevitable in the initial detection results for multi-target cases, especially when we utilize the foreground masks to re-weight the convolutional feature maps. To further refine the initial results and make use of the visual features of small targets, we propose a coarse-to-fine candidate selection process to screen out the true predictions.

The initial detected regions are treated as candidates, and we locate the corresponding regions in the original image to crop out the image patches. With these patches, we can utilize a shallow convolutional neural network to extract the visual features of targets without the interference of the extra backgrounds. Therefore, a simple yet effective binary classification network

can be trained to classify the given candidate as the true target or not. The selection process is shown in Fig. 4.

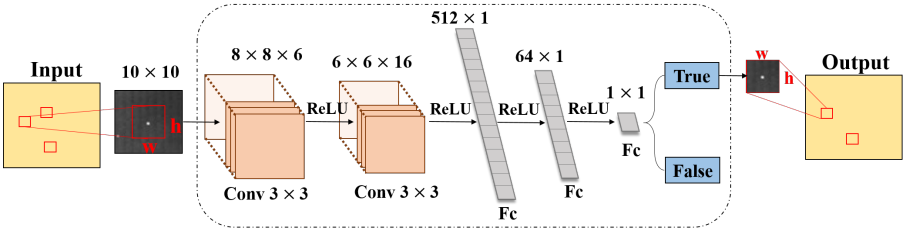



**Fig. 4** The architecture of candidate selection process, where $w$ and $h$ represent the size of the candidate region.

More specifically, the selection network takes the bounding boxes of true targets and surrounding regions as positive and negative samples, and feed them through two 3×3 convolution layers followed by the ReLU layer to extract features, and further pass them through the fully-connected layers to compute the classification confidence. During the inference, we feed initial candidates to the trained classifier, and remove ones that are identified as the background.

## 2.4 Loss Function

The loss function of the proposed model contains two parts. One is the loss of detection network, the other is for the candidate selection network.

The loss function of detection network mainly consists of three different losses, *i.e.,* bounding box regression loss, confidence loss, and category loss [41]. The regression loss includes the loss of the target position $x$ and $y$ and the loss of the target size $w$ and $h$. The loss for horizontal position $x$ is defined as follows, and the losses for $y$, $w$ and $h$ are of the same form:

$$Loss_x = \sum_{i}^{S^2} \sum_{j}^{B} I_{ij}^{obj} (x_i - \hat{x}_i)^2 \tag{5}$$

where $i$ is the index of the image cell, and each image is separated to $S^2$ cells. $B$ indicates the number of predictions from one cell. $I_{ij}^{obj}$ denotes the $j_{th}$ bounding box predictor in cell $i$ that is responsible for that prediction. $x$ and $\hat{x}$ represent the prediction and the ground-truth respectively.

The confidence loss and the category loss are defined as follows, which are the same as original Yolov3 [42] framework:

$$
\begin{aligned}
Loss_{conf} = &- \sum_{i}^{S^2} \sum_{j}^{B} I_{ij}^{obj} (\hat{C}_i^j log(C_i^j) + (1 - \hat{C}_i^j) log(1 - C_i^j)) \\
&- \sum_{i}^{S^2} \sum_{j}^{B} I_{ij}^{noobj} (\hat{C}_i^j log(C_i^j) + (1 - \hat{C}_i^j) log(1 - C_i^j))
\end{aligned} \tag{6}
$$

$$
Loss_{cls} = - \sum_{i}^{S^2} I_{ij}^{obj} \sum_{c \in cls} (\hat{P}_i^j log(P_i^j) + (1 - \hat{P}_i^j) log(1 - P_i^j)) \tag{7}
$$

where $C_i^j$ represents the confidence of each cell, and $\hat{C}_i^j$ is the binary value of zero and one, which is determined by whether the bounding box of the cell is responsible for prediction. $P_i^j$ is the classification confidence. The final loss takes the form of the sum of above losses. Since the task of detection network is a single category detection and the size of the targets is relatively small, we re-weight the bounding box loss to achieve a better balance among losses as:

$$
Loss = \alpha Loss_x + \alpha Loss_y + \beta Loss_w + \beta Loss_h + Loss_{cls} + Loss_{conf} \tag{8}
$$

Moreover, for the loss function of the candidate selection network, we employ the binary cross entropy loss:

$$
Loss_{cs} = -(\hat{y} log(y) + (1 - \hat{y}) log(1 - y)) \tag{9}
$$

### 2.5 Implementation Details

During training, we set the batch size to 8 and the initial learning rate to 0.01. Adam algorithm [43] was used as the optimizer, the momentum and weight decay were set to 0.9 and 0.0005 respectively. The weight $\alpha$ and $\beta$ of the loss were empirically set to 50 and 1. We adopted the anchor-based detection framework and set nine anchor sizes from $2 \times 2$ to $9 \times 9$ with an additional $15 \times 15$ by referring to the k-means algorithm and artificial adjustments.

## 3 Experiments

In this section, we introduce the details of constructing the synthetic multi-target datasets and the evaluation metrics employed for the multi-target detection. Next, we compare the proposed method against state-of-the-art models and conduct the thorough analysis. Finally, we present ablation studies to investigate the effectiveness of proposed components.

**Table 1** Baisc statistic information of each frame in synthetic datasets

| Type | Size | Foreground Pixels | SCR | Number Per Frame |
|---|---|---|---|---|
| 1 | 2×2 ∼ 4×4 | 4 ∼ 5 | 3, 5, 6 | 6 |
| 2 | 4×4 ∼ 6×6 | 6 ∼ 13 | 3, 5, 6 | 6 |
| 3 | 6×6 ∼ 10×10 | 14 ∼ 20 | 3, 5, 6 | 6 |

Foreground pixels means the number of pixels that are considered as the foreground. SCR indicates the Signal-to-Clutter Ratio, represents the contrast between target and background pixels.

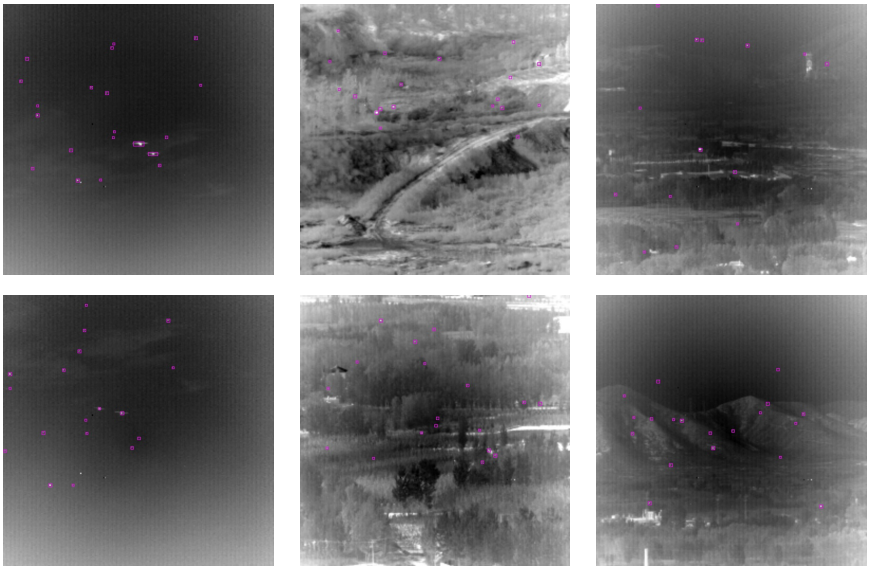
## 3.1 Datasets

For the experimental evaluation, we constructed synthetic datasets based on the publicly available dataset IDST [44] and our own collected infrared data. Given original datasets generally contain only one or two targets in each frame, to mimic the realistic scenarios, we converted them into multi-target counterparts. The following simulation strategy was adopted, mainly consists of two steps:

**Target Template Design:** Firstly, we selected appropriate targets from the original dataset and extracted their corresponding areas on the original image as the initial templates by referring to their ground-truth bounding boxes. Next, considering the actual target often has the diverse brightness in different scenes, we adjusted the brightness of the template to fit the brightness change of the real scene. Specifically, we took the mean value of the template as the threshold to separate the foreground and background. The part of the area above the threshold was considered as the foreground, and its value can be enlarged or reduced with certain proportion to mimic the cases of brightness enhancement and weakening. At last, given the moving characteristics of the infrared targets, we applied the Gaussian blur to fit the blurry phenomenon of moving targets.

**Target Insertion:** After obtaining the target templates, we manually selected a reasonable initial area and randomly selected the coordinates of the start point in this area as the position of the target center. Then, we inferred tracks of inserted targets based on the motion tracks of the original target among consecutive frames, and added certain random displacements in the tracks. Finally, we transformed the sizes and brightness of targets into one of pre-defined settings, and fused them with the original image using Poisson Integration. The pre-defined settings of size and SCR, as well as the number of targets for each mode can be found in Table 1.

For the IDST dataset, we processed it with the above method, and obtained its multi-target counterpart, noted as IDSMT, which includes 16177 images with about 18 targets in each frame and three types of diverse backgrounds. By referring to [45], a small target is defined as possesses less than 0.15% pixels in an entire image. Therefore, the constructed IDSMT meets the standard. The image size of IDSMT is 512×512, and targets varies from 2×2 pixels to 10×10 pixels. Apart from the original targets, the entire dataset contains about 97062



(a) (b) (c)

**Fig. 5** Infrared images of the IDSMT dataset with three types of backgrounds.

objects of each type. The background includes 1397 images of pure sky, 901 images of combined sky and ground filed, and 13879 images of pure ground filed. The total 22 groups of data is divided into the training and test sets with the ratio of 7:3. All types of target size and backgrounds are evenly distributed among training and test sets. The examples of infrared multi-target images of IDSMT are shown in Fig. 5.

Moreover, we also collaborated with Shanghai Institute of Technical Physics of the Chinese Academy of Sciences to collect a second infrared dataset, noted as SITP. These images were captured to record planes from the distance of 10 km, each frame contains one or two planes as targets. In order to model the multi-target scenario, we extended this dataset by following the same synthetic protocol, noted as SITPMT. After removing invalid frames, SITPMT contains 15101 images in total, and the background is mainly composed of the sky and buildings. Each frame includes about 12 targets in average, and there are 53796 Type3 objects, 75784 Type2 objects, and 53985 Type1 objects respectively. The size of each image is 320×256. The example images of SITPMT are shown in Fig. 6.

## 3.2 Evaluation Metrics

To quantitatively evaluate the performance of proposed method for the infrared multi-target detection task, by referring to [35], we select Precision(P), Recall(R), Average Precision(AP) and F1 score to evaluate results. Precision is the ability of a model to identify only the relevant objects, which is the percentage of correct positive predictions and is given by:

**Fig. 6** Infrared images of SITPMT with three consecutive frames

$$P = \frac{TP}{TP + FP} \tag{10}$$

Recall is the ratio of true positive instances against the sum of true positives and false negatives in the detector, based on the ground-truth, and is given by:

$$R = \frac{TP}{TP + FN} \tag{11}$$

F1 score is the measurement that leverages both precision and recall, and is given by:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{12}$$

True Positive (TP) represents the number of correct detections. False Positive (FP) counts the number of incorrect predictions. False Negative (FN) indicates the number of the ground-truth boxes missing in results. We consider a prediction as TP or FP by calculating the IoU and the distance of center points between the ground-truth bounding box and the prediction. An additional metric to compute the area under the curve of the Precision-Recall curve, *i.e.,* Average Precision (AP) is also employed for the evaluation.

## 3.3 Results and Analysis

In this section, we demonstrate the effectiveness of our proposed method by comparing against classic and state-of-the-art methods, including traditional methods IPI [20] and deep neural network-based models. For neural network-based models, we compared against the generic object detection model Yolov3 [42], as well as models proposed tailored for infrared target detection: ACM [46], ALCNet [36], DNANet [37], AGPCNet [47]. For fair comparisons, the compared models were implemented by referring to their publicly released code and re-trained on our synthetic datasets with same experimental settings.

Table 2 shows the results of different methods on the IDSMT. The traditional method IPI behaves worse than deep neural network-based methods, illustrating the advanced learning ability of network models. Among network-based methods, ACM and ALCNet rely on the bottom-up attention mechanism to preserve the spatial information of small targets, while AGPCNet uses the relatively complicated pyramid structure to capture the correlation among adjacent pixels, which achieves the better performance. Compared to these

**Table 2** Detection result of proposed model and compared methods on the IDSMT.

| Method | Precision(%) | Recall(%) | AP(%) | F1(%) |
|---|---|---|---|---|
| IPI[20] | 15.25 | 30.75 | 13.05 | 20.38 |
| Yolov3[42] | 90.83 | 77.47 | 71.20 | 84.14 |
| ACM[46] | 78.47 | 73.27 | 60.14 | 75.78 |
| ALCNet[36] | 80.79 | 73.90 | 59.70 | 77.19 |
| AGPCNet[47] | 75.24 | **89.74** | 67.85 | 81.85 |
| DNANet[37] | 77.44 | 84.97 | 68.78 | 81.03 |
| Ours | **94.13** | 79.83 | **76.27** | **86.36** |

methods, DNANet proposes to utilize the nested network architecture combined with attention mechanism, which obtains a high recall. However, due to the interference of complex backgrounds, the precision of these methods is somewhat limited. It is worth noting that these methods treat the small target detection as a segmentation task, and further ground the foreground pixels with bounding boxes. Different from them, we utilized the Yolov3-based framework and adjusted anchor sizes to suit for identifying the small targets. As we can see, by introducing the mask-guided feature enhancement tailored for detecting small targets, and employing the coarse-to-fine candidate feature selection process, our proposed model obtains significant improvements on AP and F1 scores against the plain Yolov3, as well as outperforms state-of-the-art models.

Fig. 7 shows the visualized results of different methods on IDSMT. We can observe that false alarms generated by IPI are mainly the background areas that possess the similar brightness as targets in (b). A certain number of bright targets are omitted, while a small number of background areas that close to the target are incorrectly detected. As shown in the red boxes of (d) and (e), for targets with the similar high brightness as buildings, ACM and ALCNet also tend to miss them. Besides, ALCNet also misses the dimmer targets. Meanwhile, in the blue boxes of (c) and (d), ACM and ALCNet also have false alarms. DNANet is largely affected by backgrounds with false alarms occurred in complex areas as illustrated in the blue box of (g). In (f), we can see that the results of AGPCNet basically fit the ground-truth, however, the multiple redundant predictions occur for the same target. The results of our model are the closest to the ground-truth, proving the advantage of our model in detecting multi-target under complex backgrounds.

The experimental results in Table 3 and visualized results in Fig. 8 on the SITPMT are consistent with the IDSMT, demonstrating the robustness of the proposed model on different scenes. Specifically, as shown in the boxes in (c), (d) and (e), for targets appeared around buildings, our model performs significant better compared to plain Yolov3, ACM and ALCNet. For dim targets inside the cloud, the proposed model can successfully locate them, while AGPCNet and DNANet often fail to perform detection as illustrated in the boxes of (f) and (g).

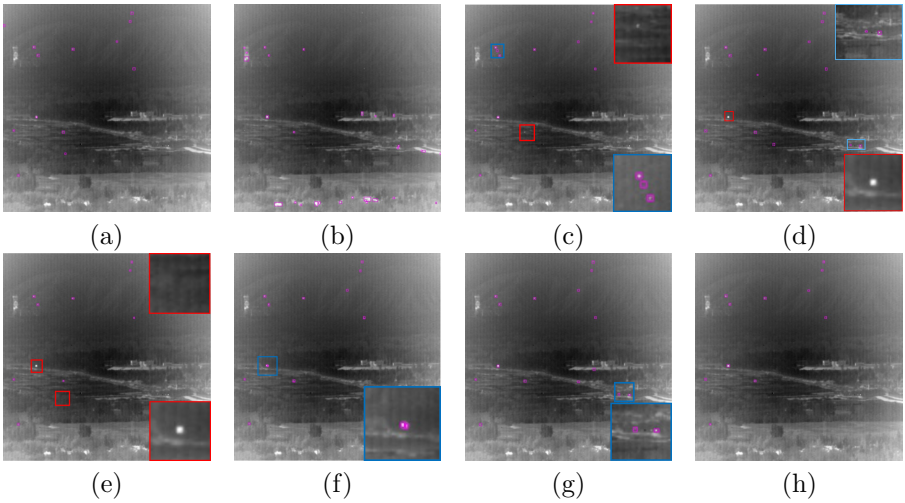



**Fig. 7** The visualized results of different methods on IDSMT. Part of the areas is enlarged in corners for better visualization. The pink boxes are the predicted boxes. The large red boxes indicate the miss-detections and blue ones show the false alarms. (a) Ground-Truth, (b) IPI, (c) Yolov3, (d) ACM, (e) ALCNet, (f) AGPCNet, (g) DNANet, (h) Proposed Model.

**Table 3** Detection results of proposed model and compared methods on the SITPMT.

| Method | Precision(%) | Recall(%) | AP(%) | F1(%) |
|---|---|---|---|---|
| IPI[20] | 50.61 | 17.70 | 11.74 | 26.22 |
| Yolov3[42] | 96.87 | 94.08 | 91.65 | 95.51 |
| ACM[46] | 94.06 | 87.51 | 84.84 | 90.67 |
| ALCNet[36] | 78.23 | 82.60 | 67.02 | 80.36 |
| AGPCNet[47] | 93.57 | 93.11 | 89.68 | 93.34 |
| DNANet[37] | 81.25 | **96.45** | 78.43 | 88.20 |
| Ours | **98.26** | 94.75 | **93.88** | **96.47** |

## 3.4 Ablation Study

To verify the effectiveness of each proposed component, we conducted thorough ablation analysis. The plain Yolov3 was adopted as our backbone network, and we added each component on top of the backbone to carry out experiments as shown in Table 4. It can be seen that the mask-guided feature extraction achieves 3.55% improvement compared to the baseline on AP, while the candidate selection obtains 2.76%, and the final model performs best with two components together. We also observe that the mask-guided feature extraction mainly improves the recall, while the candidate selection process can screen out the false alarms to further increase the precision.

We visualize the predicted results of baseline, its mask-guided version, as well as after the candidate selection. It can be observed from red boxes in the second column in Fig. 9, the dimmer targets can be located by the mask-guided feature extraction. Though introducing the mask increases the false alarms,

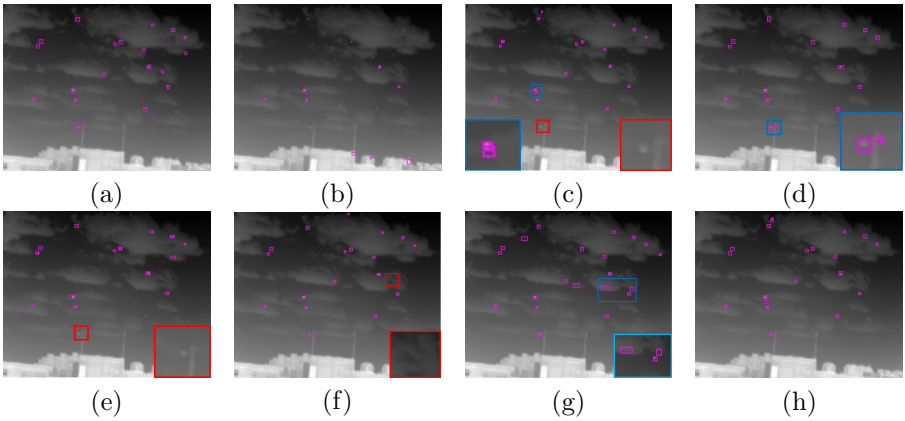



**Fig. 8** The visualized results of different methods on SITPMT. Part of the areas is enlarged in the corner for better visualization. The pink boxes are the predicted boxes. The large red boxes indicate the miss-detections and blue ones show the false alarms. (a) Ground-Truth, (b) IPI, (c) Yolov3, (d) ACM, (e) ALCNet, (f) AGPCNet, (g) DNANet, (h) Proposed Model.

**Table 4** Detection results of ablation studies on IDSMT

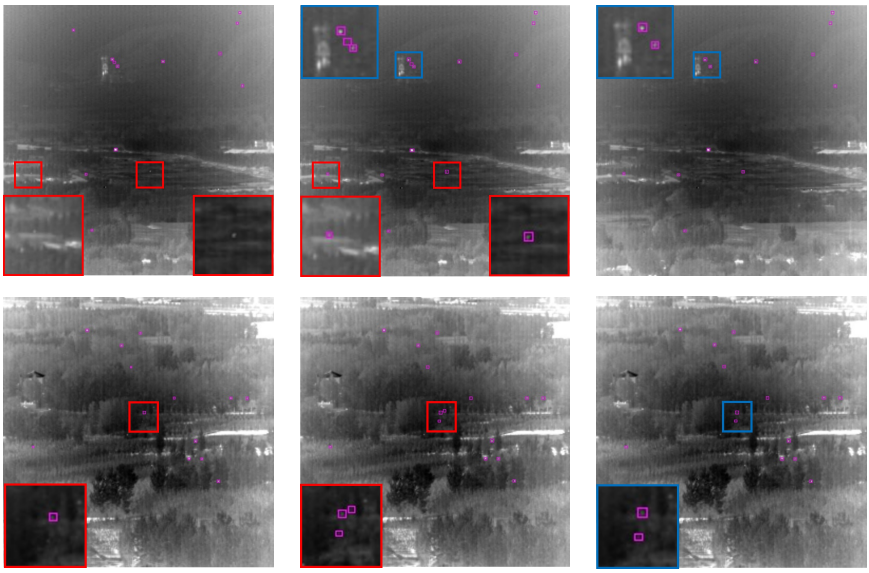
| Method | Precision(%) | Recall(%) | AP(%) | F1(%) |
|---|---|---|---|---|
| Baseline(Yolov3) | 90.83 | 77.47 | 71.20 | 78.96 |
| Baseline + Mask-Guided Feature Extraction | 92.87 | 79.79 | 74.75 | 85.84 |
| Baseline + Candidate Selection | **96.28** | 76.88 | 73.96 | 85.19 |
| Ours | 94.13 | **79.83** | **76.27** | **86.36** |

they can be effectively removed by the candidate selection as shown in blue boxes.

We also investigated the mask-guided position, *i.e.,* which convolotional feature map should be guided in order to obtain the optimal effect. As shown in Table 5, we selected 6 positions for experiments. The positions are noted as follows:

1. Start: the output of the first convolutional layer of Yolov3.
2. Middle: the output of the convolutional layer after two down-sampling layers.
3. Before Branch: the output of the convolutional layer before the first branch.
4. Branch Start: the outputs of the first convloutional layers of three branches.
5. Branch Middle: the outputs of the convloutional layers after four layers in three branches.
6. Branch End: the outputs of the last convloutional layers before detection head in three branches.

The table 5 shows that *Before Branch* is the best position to apply the mask, since the feature map at this position possess both spatial and semantic features, adding the mask-guided information here helps the subsequent network to pay more attention on the features of highlighted areas. It also can be



(a) (b) (c)

**Fig. 9** The visualized results of ablation studies on IDSMT. Part of the areas is enlarged in corners for better visualization. The pink boxes are the detected box. (a) Baseline, (b) Baseline+Mask-Guided Feature Extraction, (c) Proposed Model.

**Table 5** The ablation experimental results of different mask-guided positions.

| Position | Precision(%) | Recall(%) | AP(%) | F1(%) |
|---|---|---|---|---|
| Baseline(Without Mask) | 90.83 | 77.47 | 71.20 | 78.96 |
| Start | 93.61 | 78.28 | 73.67 | 85.26 |
| Middle | 92.64 | 79.55 | 74.22 | 85.59 |
| Before Branch | **92.87** | 79.79 | **74.75** | **85.84** |
| Branch Start | 91.54 | 79.38 | 73.31 | 85.03 |
| Branch Middle | 91.60 | 77.97 | 72.16 | 84.24 |
| Branch End | 88.25 | **81.38** | 72.07 | 84.68 |

seen that the insertion at the *Branch End* greatly improves the recall, yet has a negative impact on the precision. The reason for this phenomenon is that this position is where the feature extraction of targets is already completed. Adding the guided information here increases the influence of the extra interference carried by the mask itself. Since there is no subsequent convolutional layers to further adjust features, the interference may lead to the decrease of precision. Therefore, we adopted position of *Before Branch* in our final model.

Moreover, we also studied the sensitivity of different weight ratios $\alpha$ and $\beta$ among bounding box losses in equation 8. Table 6 shows the results of different weight ratios. As we can see, the $\alpha : \beta = 50 : 1$ reaches the best overall performance. The reason is that the detection errors mainly caused by the inaccurate localization instead of the size of the predicted bounding box in our

**Table 6** The experimental results of the different loss weights.

| $\alpha$ | $\beta$ | Precision(%) | Recall(%) | AP(%) | F1(%) |
|---|---|---|---|---|---|
| 1 | 1 | 78.09 | 77.26 | 61.76 | 77.68 |
| 1 | 10 | 82.34 | 65.87 | 56.41 | 73.19 |
| 10 | 1 | 85.63 | 79.07 | 69.29 | 82.22 |
| 15 | 1 | 87.45 | **80.13** | 71.36 | 83.63 |
| 50 | 1 | **92.93** | 76.88 | **71.93** | **84.14** |
| 100 | 1 | 88.24 | 78.00 | 70.64 | 82.81 |

experiments. Therefore, it is reasonable to set a relatively large weights to the $Loss_x$ and $Loss_y$.

# 4 Conclusions

In this paper, we propose a mask-guided detection model via coarse-to-fine candidate selection for infrared small multi-target detection. The mask-guided feature extraction utilizes the sparse and low-rank properties of small targets and infrared backgrounds to generate the foreground mask, and further re-weights the convolutional feature maps to guide the network focus on small targets, while the candidate selection process is conducted in a coarse-to-fine fashion to rule out false alarms caused by complex backgrounds using nuanced visual features obtained from the original image areas. Extensive experiments demonstrate that the proposed method achieves the best performance against state-of-the-art methods under infrared small multi-target detection scenarios.

# Declarations

- Funding: This work was supported in part by the National Natural Science Foundation of China (No.61572307).
- Conflict of interest: The authors declare that they have no conflict of interest.
- Consent to participate: All authors are agreed to participate in this research work.
- Consent for publication: All authors are agreed to the publication of this research paper.

# References

[1] Wei, Z., Cong, M., Wang, L.: Algorithms for optical weak small targets detection and tracking: Review. In: International Conference on Neural Networks and Signal Processing, 2003. Proceedings of the 2003 (2003)

[2] Li, H.J., Zhang, P., Wang, X.W., Huang, S.Z.: Infrared small-target detection algorithms: a survey. Journal of image and Graphics(JIG) **25**(9), 15 (2020)

[3] Rawat, S.S., Verma, S.K., Kumar, Y.: Review on recent development in infrared small target detection algorithms. Procedia Computer Science **167**, 2496–2505 (2020)

[4] Rivest, J.-F., Fortin, R.: Detection of dim targets in digital infrared imagery by morphological image processing. Optical Engineering **35**(7), 1886–1893 (1996)

[5] Wang, X., Peng, Z., Zhang, P., He, Y.: Infrared small target detection via nonnegativity-constrained variational mode decomposition. IEEE Geoence & Remote Sensing Letters, 1700–1704 (2017)

[6] Zhang, B., Zhu, J., Lu, X., Gu, Y., Li, J., Liu, X., Zhang, M.: An infrared dim target detection algorithm based on density peak search and region consistency. Optical and Quantum Electronics **53**(7), 1–18 (2021)

[7] Ren, K., Song, C., Miao, X., Wan, M., Xiao, J., Gu, G., Chen, Q.: Infrared small target detection based on non-subsampled shearlet transform and phase spectrum of quaternion fourier transform. Optical and Quantum Electronics **52**(3), 1–15 (2020)

[8] Seyed, M.F., Reza, M.M., Mahdi, N.: Flying small target detection in ir images based on adaptive toggle operator. IET Computer Vision **12**(4), 527–534 (2018)

[9] Tianai, W.U., Huang, S., Yuan, Z., Yunrong, W.U., Feng, H.: Nsct combined with svd for infrared dim target complex background suppression. Infrared Technology (2016)

[10] Han, J., Yong, M., Bo, Z., Fan, F., Liang, K., Yu, F.: A robust infrared small target detection algorithm based on human visual system. IEEE Geoscience & Remote Sensing Letters **11**(12), 2168–2172 (2014)

[11] Jinhui, Han, Sibang, Liu, Gang, Qin, Qian, Zhao, Honghui, Zhang: A local contrast method combined with adaptive background estimation for infrared small target detection. Geoscience and Remote Sensing Letters, IEEE **16**(9), 1442–1446 (2019)

[12] Deng, H., Sun, X., Liu, M., Ye, C., Zhou, X.: Small infrared target detection based on weighted local difference measure. IEEE Transactions on Geoscience & Remote Sensing **54**(7), 4204–4214 (2016)

[13] Li, Q., Nie, J., Qu, S.: A small target detection algorithm in infrared image by combining multi-response fusion and local contrast enhancement. Optik - International Journal for Light and Electron Optics **241**(3), 166919 (2021)

[14] Ma, T., Wang, J., Yang, Z., Ren, X., Song, Y., Ku, Y., Liu, Y., Wang, D.: Infrared small target detection based on divergence operator and nonlinear classifier. Optical and Quantum Electronics **53**(7), 1–23 (2021)

[15] Qian, Y., Chen, Q., Zhu, G., Gu, G., Xiao, J., Qian, W., Ren, K., Wan, M., Zhou, X.: Infrared small target detection based on saliency and gradients difference measure. Optical and Quantum Electronics **52**(3), 1–21 (2020)

[16] Chen, C., Li, H., Wei, Y., Xia, T., Tang, Y.Y.: A local contrast method for small infrared target detection. IEEE Transactions on Geoscience & Remote Sensing **52**(1), 574–581 (2013)

[17] Han, J., Moradi, S., Faramarzi, I., Liu, C., Zhao, Q.: A local contrast method for infrared small-target detection utilizing a tri-layer window. IEEE Geoscience and Remote Sensing Letters **PP**(99), 1–5 (2019)

[18] Zhao, J., Tang, Z., Yang, J., Liu, E.: Infrared small target detection using sparse representation. Journal of Systems Engineering and Electronics **22**(6), 8 (2011)

[19] He, Y.J., Li, M., Zhang, J.L., An, Q.: Small infrared target detection based on low-rank and sparse representation. Infrared Physics & Technology **68**, 98–109 (2015)

[20] Gao, C., Meng, D., Yang, Y., Wang, Y., Zhou, X., Hauptmann, A.G.: Infrared patch-image model for small target detection in a single image. IEEE Transactions on Image Processing **22**(12), 4996–5009 (2013)

[21] Kong, X., Yang, C., Cao, S., Li, C., Peng, Z.: Infrared small target detection via nonconvex tensor fibered rank approximation. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–21 (2021)

[22] Xiong, B., Huang, X., Wang, M., Peng, G.: Small target detection for infrared image based on optimal infrared patch-image model by solving modified adaptive rpca problem. International Journal of Pattern Recognition and Artificial Intelligence (2020)

[23] Rawat, S.S., Verma, S.K., Kumar, Y., Kumar, G.: Infrared small target detection based on non-convex lp-norm minimization. Journal| MESA **12**(4), 929–942 (2021)

[24] A, Y.D., D, Y.W.A.B.C., A, Y.S.: Infrared small target and background separation via column-wise weighted robust principal component analysis. Infrared Physics & Technology **77**, 421–430 (2016)

[25] Yimian, Dai, Yiquan, Wu: Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection.

IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **10**(8), 3752–3767 (2017)

[26] Dai, Y., Wu, Y., Song, Y., Guo, J.: Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values. Infrared Physics & Technology **81**, 182–194 (2017)

[27] Du, J., Lu, H., Hu, M., Zhang, L., Shen, X.: Cnn-based infrared dim small target detection algorithm using target-oriented shallow-deep features and effective small anchor. IET Image Processing **15**(1), 1–15 (2021)

[28] Hou, Q., Wang, Z., Tan, F., Zhao, Y., Zhang, W.: Ristdnet: Robust infrared small target detection network. IEEE Geoscience and Remote Sensing Letters **PP**(99), 1–5 (2021)

[29] Liu, M., Du, H.-y., Zhao, Y.-j., Dong, L., Hui, M., Wang, S.: Image small target detection based on deep learning with snr controlled sample generation. Current Trends in Computer Science and Mechanical Automation **1**, 211–220 (2017)

[30] Mcintosh, B., Venkataramanan, S., Mahalanobis, A.: Infrared target detection in cluttered environments by maximization of a target to clutter ratio (tcr) metric using a convolutional neural network. IEEE Transactions on Aerospace and Electronic Systems **PP**(99), 1–1 (2020)

[31] Zhao, M., Cheng, L., Yang, X., Feng, P., Liu, L., Wu, N.: Tbc-net: A real-time detector for infrared small target detection using semantic constraint (2019)

[32] Fan, M., Tian, S., Liu, K., Zhao, J., Li, Y.: Infrared small target detection based on region proposal and cnn classifier. Signal Image and Video Processing (2021)

[33] Chen, F., Gao, C., Liu, F., Zhao, Y., Zhou, Y., Meng, D., Zuo, W.: Infrared small target detection using multi-patch attention network (2021)

[34] Shi, M., Wang, H.: Infrared dim and small target detection based on denoising autoencoder network. Mobile Networks and Applications **25**(4) (2020)

[35] Ju, M., Luo, J., Liu, G., Luo, H.: Istdet: An efficient end-to-end neural network for infrared small target detection. Infrared Physics & Technology **114**(7), 103659 (2021)

[36] Dai, Y., Wu, Y., Zhou, F., Barnard, K.: Attentional local contrast networks for infrared small target detection. IEEE Transactions on

Geoscience and Remote Sensing **PP**(99), 1–12 (2021)

[37] Li, B., Xiao, C., Wang, L., Wang, Y., Lin, Z., Li, M., An, W., Guo, Y.: Dense nested attention network for infrared small target detection (2021)

[38] Zhao, B., Wang, C., Fu, Q., Han, Z.: A novel pattern for infrared small target detection with generative adversarial network. IEEE Transactions on Geoscience and Remote Sensing **59**(5), 4481–4492 (2020)

[39] Wang, H., Zhou, L., Wang, L.: Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8509–8518 (2019)

[40] Toh, K.C., Yun, S.: An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. Pacific Journal of Optimization **6**(3), 615–640 (2010)

[41] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. IEEE (2016)

[42] Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)

[43] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[44] Hui, B.W., Song, Z.Y., Fan, H.Q.: A dataset for infrared detection and tracking of dim-small aircraft targets under ground / air background. China Scientific Data **5**(3), 12 (2020)

[45] Zhang, W., Cong, M., Wang, L.: Algorithms for optical weak small targets detection and tracking. In: International Conference on Neural Networks and Signal Processing, 2003. Proceedings of the 2003, vol. 1, pp. 643–647 (2003). IEEE

[46] Dai, Y., Wu, Y., Zhou, F., Barnard, K.: Asymmetric contextual modulation for infrared small target detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 950–959 (2021)

[47] Zhang, T., Cao, S., Pu, T., Peng, Z.: Agpcnet: Attention-guided pyramid context networks for infrared small target detection. arXiv preprint arXiv:2111.03580 (2021)