Infrared Physics & Technology 114 (2021) 103659

Contents lists available at ScienceDirect

Infrared Physics and Technology

journal homepage: www.elsevier.com/locate/infrared

ISTDet: An efficient end-to-end neural network for infrared small target detection

Moran Ju [a,b,c,d,e], Jiangning Luo [f], Guangqi Liu [a,b,c,d,e], Haibo Luo [a,b,d,e,*]

[a] Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, Liaoning 110016, China
[b] Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, Liaoning 110016, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China
[d] Key Laboratory of Opt-Electronic Information Processing, Chinese Academy of Sciences, Shenyang, Liaoning 110016, China
[e] The Key Laboratory of Image Understanding and Computer Vision, Shenyang, Liaoning 110016, China
[f] McGill University, Montreal, Quebec H3A 0G4, Canada

ARTICLE INFO

Keywords:
Convolutional neural network
Small target detection
End-to-end
Infrared image

ABSTRACT

Infrared small target detection has made many breakthroughs in early warning, guidance and battlefield intelligence. However, infrared small target occupies less pixels and lacks color and texture features, which makes infrared small target detection a challenging subject. To achieve the infrared small target detection, an efficient end-to-end network ISTDet is proposed in this paper. ISTDet mainly consists of two modules, including image filtering module and infrared small target detection module. The image filtering module is proposed to obtain the confidence map, aiming to enhance the response of infrared small targets and suppress the response of background. The infrared small target detection module takes the infrared image activated by the confidence map as input, aiming to speculate the category and position of the infrared small targets. Multi-task loss function is used to train the ISTDet in an end-to-end way. Finally, we do comparative experiments on five infrared small target sequences to demonstrate the detection performance of ISTDet. The results show ISTDet has better performance for infrared small target detection compared with other detectors.

1. Introduction

Computer vision has been widely used in civil and military fields because of the rapid development of image processing technology. As one of the main researches in aerospace defense, early warning and battlefield intelligence, infrared small target detection has become a research hotspot.

Due to the large span of scenes and long-distance imaging, the infrared small targets tend to occupy few pixels and lacks shape and texture features. What is more, acquired infrared images are always with low signal-to-noise ratio. Therefore, infrared small target detection is still a challenge in computer vision [1].

Traditional approaches for infrared small target detection, such as, morphology-based approaches [2,3] and max-median/max-mean filter [4] use the differences between the background and the infrared small targets to perform target detection. However, these approaches are hard to find suitable templates when the targets are in complicated environment. Inspired by human visual system, a lot of infrared small target detection approaches are proposed, such as local contrast measure [5], local difference metric [6] and improved local contrast measure [7]. These approaches reinforce the difference between the background and the targets. [8] proposed an infrared image block model (IPI) to decompose the matrix of the infrared image into sparse matrices. IPI regards the small targets as non-zero values in sparse matrix, which may cause 'false positive".

In recent years, target detection approaches based on Convolutional Neural Network (CNN) [9] have made many breakthroughs. These CNN-based target detection approaches have achieved leading results in target detection tasks. Compared with traditional target detection approaches, CNN-based target detection approaches can integrate different tasks, such as feature extraction, feature fusion and feature classification into the same network. What is more, the network can be trained to learn deep semantic features of the targets in an end-to-end way. CNN-based target detection approaches mainly include two categories, namely two-stage target detection approaches and one-stage target detection approaches. R-CNN [10], Fast R-CNN [11], Faster R-

* Corresponding author.
E-mail address: luohb@sia.cn (H. Luo).

https://doi.org/10.1016/j.infrared.2021.103659
Received 1 September 2020; Received in revised form 8 December 2020; Accepted 19 January 2021
Available online 26 January 2021

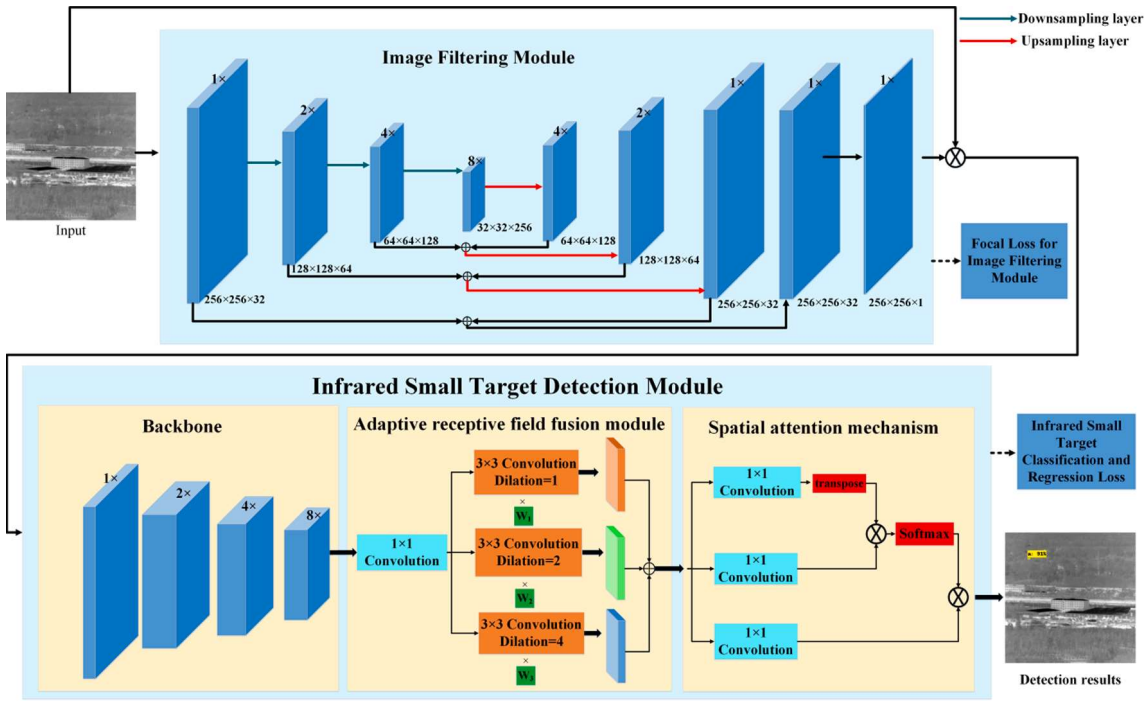

Fig.1. An overview of the architecture of ISTDet. ISTDet consists of two parts, namely, image filtering module and infrared small target detection module. Multi-task loss function is used to train ISTDet in an end-to-end way. For IFM, we use focal loss to regress the confidence map. For ISTDM, we use Generalized Intersection over Union loss and Binary Cross Entropy loss to regress the position and the category of the infrared small targets, respectively.

CNN [12] and Cascade R-CNN [13] are the representative of two-stage target detection approaches. They generate many candidate bounding boxes first and then take these candidates as input to further speculate the position and category of the targets. SSD [14], YOLO [15], YOLO 9000 [16], YOLO V3 [17], RFBnet [18] and RefineDet [19] are the representative of one-stage target detection approaches. They can forecast the category and position of the targets in an end-to-end way.

Although CNN-based target detection approaches have achieved rapid development, the abovementioned approaches are all for generic target detection. Only a few papers on CNN-based infrared small target detectors could be found in the literature. [20] introduces a CNN-based target detector for ship detection in infrared images, which designs a T-Net to generate synthetic targets and separates the detection task into two steps, including candidate targets extraction and candidate identification. The process of the ship detection is relatively complex, which can not be achieved in an end-to-end way. [21] uses CNN-based segmentation approaches and proposes a Nv-Net for infrared target detection. However, Nv-Net focuses on generic target detection, which is not suitable for infrared small target detection. [22] takes the infrared small target detection as the segmentation task and designs a denoising and autoencoder network, called CDAE. CDAE can detect the infrared small targets in an end-to-end way. However, it may cause false alarms.

In this paper, we focus on infrared small target detection. To achieve the end-to-end detection, we design an efficient CNN-based target detector, called ISTDet, which combines the image filtering task and target detection task into one network. ISTDet obtains higher detection accuracy and faster detection speed. There are two modules in ISTDet, including image filtering module (IFM) and infrared small target detection module (ISTDM). The IFM aims to enhance the response of infrared small targets and suppress the response of background, which is helpful to improve the detection performance for infrared small targets. The ISTDM aims to extract the features and predict the position and category of the infrared small targets. Multi-task loss function is used to train the ISTDet in an end-to-end way. Finally, we do comparative experiments on five real infrared small target sequences to demonstrate the detection performance of ISTDet. The results show ISTDet has better performance for real-time small target detection in infrared images compared with other target detection approaches.

The main novelties of this paper are listed in the following.

1) An efficient end-to-end CNN-based target detector ISTDet is proposed, aiming to obtain better speed-accuracy trade-off for infrared small target detection.
2) In ISTDet, we design IFM and ISTDM. The former aims to enhance the response of small targets and suppress the response of background. The latter aims to perform infrared small target detection.
3) We conduct the comparative experiments on five infrared small target sequences to demonstrate the detection performance of ISTDet. The results show that ISTDet outperforms other target detection approaches for infrared small target detection.

The rest of the paper is organized in the following: Section 2 introduces the design of IFM, ISTDM and multi-task loss function. Section 3 conducts comparative experiments on five infrared small target sequences and analyzes the experimental results with quantitative and qualitative evaluation. Section 4 is the conclusion of this paper.

2. Proposed method

The structure of ISTDet will be introduced in detail in this section, as shown in Fig. 1. ISTDet consists of two parts, namely, IFM and ISTDM. With IFM, the confidence map will be acquired. We use the confidence map to activate the input infrared image by element-wise multiplication, aiming to enhance the response of infrared small targets and suppress the response of background. Then, ISTDM takes the activated input infrared image as input and speculates the category and the position of the infrared small targets. To train ISTDet in an end-to-end way, the design of multi-task loss function for ISTDet will be introduced at last.

2.1. Image filtering module

The small targets in infrared images always occupy few pixels

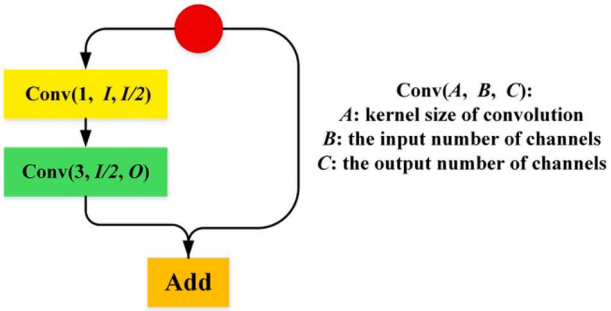

Fig. 2. Residual block: *Residual*(I, O).

Table 1
Configuration of IFM.

Layer	Layer name	Size/Stride	Filters
0	Convolution	3/1	32
1	Convolution	3/2	64
2	*Residual*$(64, 64)$		
3	Convolution	3/2	128
4	*Residual*$(128, 128)$		
5	Convolution	3/2	256
6	*Residual*$(256, 256)$		
7	Upsample		
8	*Residual*$(256, 128)$		
9	Add	Layer8 + Layer4	
10	Upsample		
11	*Residual*$(128, 64)$		
12	Add	Layer11 + Layer2	
13	Upsample		
14	*Residual*$(64, 32)$		
15	Add	Layer14 + Layer0	
16	Convolution	1/1	1

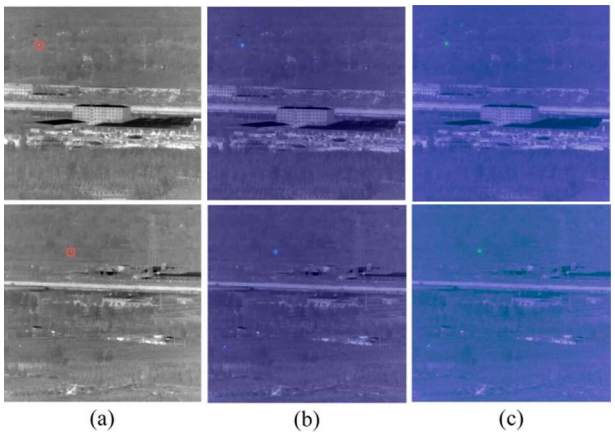

Fig. 3. The visualization of confidence map for infrared small targets. (a) Input infrared image. (b) Confidence map. (c) Activated infrared image.

without distinct texture and shape features, which makes most of the samples belong to negative ones. This extreme imbalance greatly affects the detection performance of small targets. What is more, the noise from the environment and image acquisition system makes the signal-to-noise ratio of the infrared image very low. To alleviate the issues above, we propose IFM to get the confidence map, aiming to enhance the response of infrared small targets and suppress the response of background.

Residual block [23] is used as the basic unit of IFM because it is helpful for the reuse of the features from different layers. As shown in Fig. 2, the Residual block is denoted as *Residual*(I, O), where I and O are the number of input channel and the number of output channel. The convolution layer is denoted as *Conv*(A, B, C), where A, B and C represent kernel size, the input number of channels and the output number of channels.

The detailed configuration of IFM is described in Table 1. There are more finer-grained features in the shallow layer and more semantic features in the deeper layer. The finer-grained features are helpful to locate the targets and semantic features are helpful to categorize the targets. Hence, we perform feature fusion among the features with different resolutions, aiming to improve the detection performance of ISTDet.

To get the confidence map, it is essential to produce the ground-truth confidence by the ground-truth box. The resolution of ground-truth confidence is the same as that of input image. We generate the confidence ground-truth as follows: if a pixel $GT_{h,w}$ locates within the ground-truth box of the infrared small target, we assign the pixel $GT_{h,w}$ as the positive one.

Fig. 3 shows the visualization of confidence map. Fig. 3(a) is the input infrared image and Fig. 3(b) is the confidence map. It is obvious that the response of infrared small target has been enhanced and the response of background has been suppressed by IFM. We use the confidence map to activate the input infrared image by element-wise multiplication. Fig. 3(c) shows the activated infrared image. Finally, we take the activated infrared image as the input of ISTDM to further speculate the category and the position information of the infrared small targets.

2.2. Infrared small target detection module

ISTDM is designed to perform feature extraction and speculate the category and the position of the infrared small targets. ISTDM consists of three parts, namely, the backbone, adaptive receptive field fusion module and spatial attention mechanism, as described in Fig. 1. ISTDM takes the activated input infrared image as input.

2.2.1. Backbone design

We take Darknet53 as the basis design for backbone because it is simple and efficient [19]. Darknet53 makes use of the structure of the Residual Networks [23], which consists of successive 1 × 1 and 3 × 3 convolutions. YOLO V3 uses three scales to detect the targets of different sizes. The feature maps are down-sampled by 8×, 16 × and 32 × respectively at each scale. We focus on infrared small target detection in this paper. Therefore, we truncate the convolutions before the 16 × down-sampled layer in Darknet53 as the backbone and choose 8 × down-sampled feature maps to detect infrared small target.

2.2.2. Adaptive receptive field fusion module

As analyzed in [24], context information is vital to recognize the targets. Expanding receptive field is helpful to increase the context information around the targets. Inspired by [25], we utilize three 3 × 3 convolutions with 1, 2 and 4 dilation to build adaptive receptive field fusion module, aiming to expand the receptive field, as shown in Fig. 1. In addition, we fuse the features with different receptive field to improve the detection performance of ISTDet. The features with different

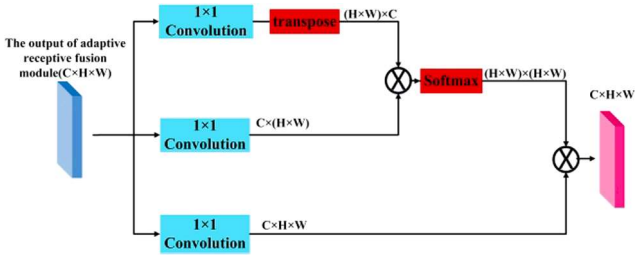

Fig. 4. Spatial attention mechanism.

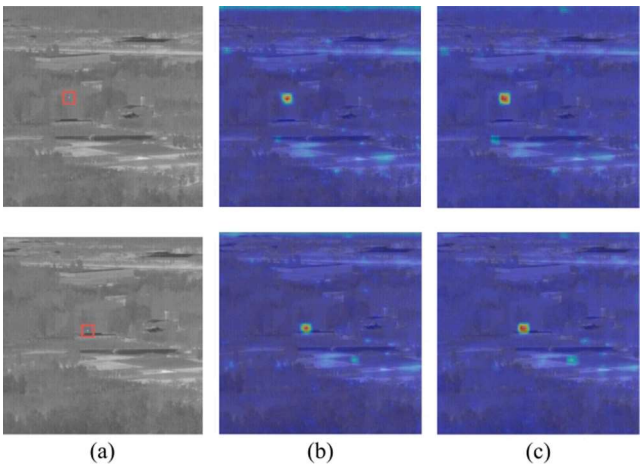

Fig. 5. Comparison of visual results of feature map before and after increasing spatial attention mechanism. (a) Infrared small target image. (b) Without spatial attention mechanism. (c) With spatial attention mechanism.

receptive field usually contribute to the output unequally. Hence, a learnable weight is added to adaptive receptive field fusion module.

2.2.3. Spatial attention mechanism

Due to the large span of scenes and long-distance imaging, the infrared images are always with low signal-to-noise ratio and the infrared small targets are always dim. Inspired by [26], spatial attention mechanism can model the relationships between widely separated spatial regions, which can efficiently enhance the response of infrared small target. Therefore, we add spatial attention mechanism to ISTDM. Fig. 4 shows the structure of spatial attention mechanism.

Suppose that the output of adaptive receptive field fusion module is $F \in \mathbb{R}^{C\times H\times W}$, which is fed into three 1×1 convolutions to generate three feature maps F_1, F_2 and F_3, where $F_1, F_2, F_3 \in \mathbb{R}^{C\times H\times W}$, H, W and C represent the height, width and channels of the feature maps. Then, we reshape F_1, F_2 and F_3 to $\mathbb{R}^{C\times(H\times W)}$. We perform a matrix multiplication between the transpose of F_2 and F_1, and softmax operation to get $H \in \mathbb{R}^{(H\times W)\times(H\times W)}$, which measures the impact among different positions of the feature maps. Finally, we perform a matrix multiplication between F_3 and $H$to get the final output.

To illustrate the impact of spatial attention mechanism, we compare the feature map with spatial attention mechanism and the one without spatial attention mechanism, as shown in Fig. 5. Fig. 5(a) is the original input image, where the red box denotes the position of the infrared small target. Fig. 5(b) shows the visualization of feature map without spatial attention mechanism and Fig. 5(c) is the visualization of feature map with spatial attention mechanism. It is obvious that the response of the infrared small target has been enhanced by spatial attention mechanism.

2.3. Multi-task loss function

Multi-task loss function is used to train the proposed network in an end-to-end way. There are two parts in the loss function of ISTDet, namely, the loss for IFM and the loss for ISTDM.

For the former, we use Focal loss [27] to regress the confidence of the infrared small targets, aiming to adjust the imbalance between negative and positive samples and pay more attention to hard samples, which is represented as $Loss_{IFM}$:

$$Loss_{IFM} = \begin{cases} -\alpha(1-y_P)^{\gamma} \times \quad logy_P, & y_{GT} = 1 \\ -(1-\alpha)y_P^{\gamma} \times log(1-y_P), & y_{GT} = 0 \end{cases} \tag{1}$$

where y_{GT} denotes the ground truth confidence and y_P denotes the predicted confidence. α and γ are set by 0.25 and 2, respectively.

After that, we use the confidence map to activate the input infrared image by element-wise multiplication. Then, pass the activated infrared image to the ISTDM to speculate the category and the position of the infrared small targets. We use Generalized Intersection over Union loss (GIOU loss) [28] to regress the position of the infrared small targets. We can use the following formula to compute GIOU.

$$GIOU_{B_{GT},B_P} = \frac{|B_{GT} \cap B_P|}{|B_{GT} \cup B_P|} - \frac{|B \backslash (B_{GT} \cup B_P)|}{|B|} \tag{2}$$

where B_{GT} represents the ground truth box andB_P represents the predicted box. B denotes the smallest enclosing convex region between B_{GT} and B_P. Then we can calculate the GIOU loss by

$$Loss_{giou} = 1 - GIOU_{B_{GT},B_P} \tag{3}$$

The category of the infrared small targets is regressed by Binary Cross Entropy loss.

$$Loss_{cls} = C_{GT}logC_P - (1-C_{GT})log(1-C_P) \tag{4}$$

where C_P is the predicted category and C_{GT} is the ground truth category.

Then, the loss function for ISTDM can be obtained by summing $Loss_{giou}$and $Loss_{cls}$.

$$Loss_{ISTDM} = Loss_{giou} + Loss_{cls} \tag{5}$$

Finally, we can calculate the total loss of ISTDet by

$$Loss_{ISTDet} = Loss_{IFM} + Loss_{ISTDM} \tag{6}$$

Table 2
Detailed information of five infrared small target sequences.

	Image size	Frame	The details of the targets and background
Sequence1	256 × 256	3000	A long imaging distance, Ground background, Long time
Sequence2	256 × 256	763	Dim and small target, Low signal-to-noise ratio Ground background
Sequence3	256 × 256	751	Maneuvering target, Ground background
Sequence4	256 × 256	399	Target from near to far, Dim and small target, Ground background
Sequence5	256 × 256	500	Maneuvering target, Target from near to far, Ground background

3. Experiments

In this section, the experiment details will be introduced in the following five parts: (1) Dataset and Evaluation metrics; (2) Implementation details; (3) Ablation study; (4) Performance comparison. (5) Discussion.

3.1. Introduction of dataset and evaluation metrics

The experiments are conducted on five infrared small target sequences. These five infrared small target sequences are referred to [29]. This infrared dataset focuses on dim-small target detection and tracking of low altitude flying target, and provides fixed-wing UAV targets via outfield recording and post data processing. The scenario covers complex field background. Table 2 describes the detailed information of these infrared small target sequences. The examples of these five sequences are shown in Fig. 6. The red box denotes the position of the infrared small target.

The detection accuracy is evaluated by average precision (AP), recall rate and precision rate. The detection speed is evaluated by Frames per

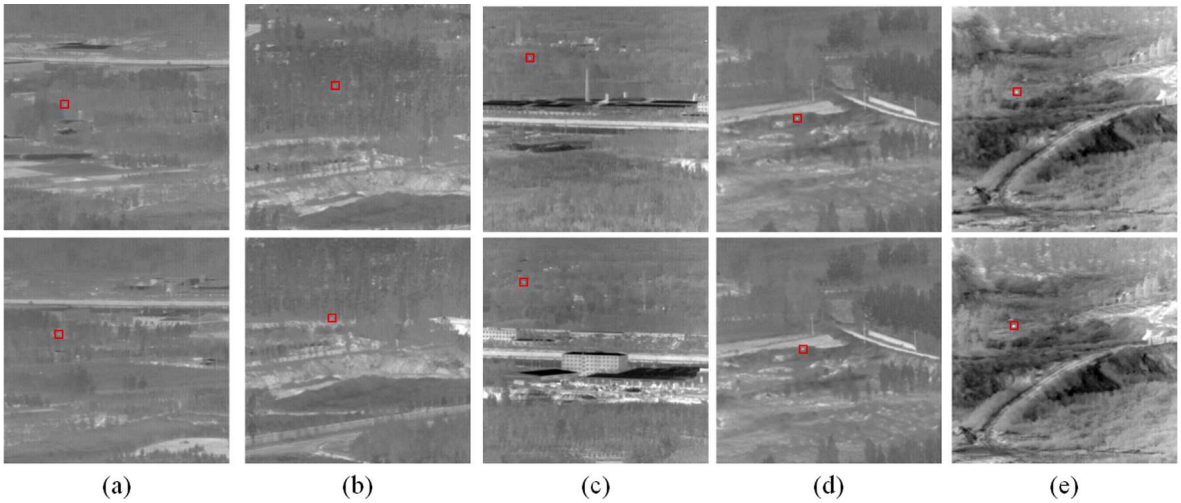

Fig. 6. Infrared small target dataset. (a) Sequence 1. (b) Sequence 2. (c) Sequence 3. (d) Sequence 4. (e) Sequence 5.

Table 3
Effectiveness of various design in ISTDet.

Baseline	With adaptive receptive field fusion module	With spatial attention mechanism	With Image filtering module	Sequence 1		Sequence 2		Sequence 3		Sequence 4		Sequence 5	
				AP (%)	FPS	AP (%)	FPS	AP (%)	FPS	AP (%)	FPS	AP (%)	FPS
√				89.05	184.97	66.54	185.84	78.55	182.12	76.13	184.67	80.53	183.11
√	√			90.40	177.10	67.02	175.27	80.33	176.17	77.42	176.39	81.07	176.46
√	√	√		91.62	167.29	71.54	162.75	81.77	164.16	79.00	165.41	83.65	164.39
√	√	√	√	**93.83**	**75.80**	**79.42**	**75.40**	**88.27**	**76.89**	**89.58**	**75.93**	**92.76**	**76.11**

second (FPS). We determine the predicted box is positive based on the common metric that the Intersection over Union (IOU) between the predicted box and the ground truth box is greater than 0.5.

$$Recall = \frac{X_{TP}}{X_{TP} + X_{FN}} \tag{7}$$

$$Precision = \frac{X_{TP}}{X_{TP} + X_{FP}} \tag{8}$$

where *Recall* and *Precision* represent recall rate and precision rate, respectively. X_{TP} denotes the number of the predictions that were correctly detected as true targets. X_{FP} represents the number of the predictions which were falsely detected as true targets. X_{FN} represents the number of the predictions which failed to be correctly detected.

From the perspective of recall rate and precision rate, AP is utilized to test the detection accuracy of the network.

3.2. Implementation details

The experiments are implemented on 1 Titan X GPU. The models are trained with Pytorch 0.4.1. The momentum is set to 0.9 and the weight decay is set to 0.0005. The initial learning rate is 0.0001. During training, it will decrease to 0.000001 based on the cosine learning rate schedule [30].

Table 4
The experimental results for sequence 1: AP (%).

Detector	YOLO V3	RFBnet	RefineDet	ISTDet*	ISTDet
Input	256	300	320	256	256
AP	72.51	76.34	88.06	91.62	93.83
FPS	63.86	71.87	45.24	167.29	75.80

3.3. Ablation study

We do the ablation study on five infrared small target sequences to verify the impact of various design in ISTDet, as shown in Table 3. The input resolution is 256 × 256. Baseline represents the version of ISTDet, which has not added adaptive receptive field module, spatial attention mechanism and image filtering module.

Baseline: As shown in Table 3, Baseline achieves 89.05% AP at 184.97 FPS in sequence 1, 66.54% AP at 185.84 FPS in sequence 2, 78.55% AP at 182.12 FPS in sequence 3, 76.13% at 184.67 FPS in sequence 4 and 80.53% at 183.11 FPS in sequence 5.

With adaptive receptive field fusion module: Compared with baseline, the AP has been improved from 89.05% to 90.40% in sequence 1, from 66.54% to 67.02% in sequence 2, from 78.55% to 80.33% in sequence 3, from 76.13% to 77.42% in sequence 4, and from 80.53% to 81.07% in sequence 5. As discussed in Section 2.2.3, adaptive receptive field fusion module is helpful to increase the context information around the targets.

With spatial attention mechanism: With spatial attention mechanism, the AP has been boosted by 1.22%, 4.52%, 1.44%, 1.58% and 2.58% in sequence 1, 2, 3, 4 and 5, respectively, which still maintains high detection speed at about 165 FPS. That is because spatial attention mechanism can model the relationships between widely separated spatial regions, which is helpful for ISTDet to have a global contextual view.

With image filtering module: It is obvious that the AP has been improved by 2.21%, 7.88%, 6.50%, 10.58% and 9.11% in sequence 1, 2, 3, 4 and 5, respectively with image filtering module. Although the detection speed has reduced, it still maintains high detection speed at about 75 FPS, which satisfies the real-time detection for infrared small targets. As discussed in Section 2.1, the response of background has been suppressed and the response of infrared small target has been enhanced by image filtering module, which is conducive to infrared small target detection.

Table 5
The experimental results for sequence 2: AP (%).

Detector	YOLO V3	RFBnet	RefineDet	ISTDet*	ISTDet
Input	256	300	320	256	256
AP	64.55	62.20	74.24	71.54	79.42
FPS	63.38	71.43	44.51	162.75	75.40

Table 6
The experimental results for sequence 3: AP (%).

Detector	YOLO V3	RFBnet	RefineDet	ISTDet*	ISTDet
Input	256	300	320	256	256
AP	76.66	78.37	82.03	81.77	88.27
FPS	62.91	71.16	45.69	164.16	76.89

Table 7
The experimental results for sequence 4: AP (%).

Detector	YOLO V3	RFBnet	RefineDet	ISTDet*	ISTDet
Input	256	300	320	256	256
AP	80.33	83.74	85.17	79.00	89.58
FPS	63.45	71.83	45.37	165.32	75.93

Table 8
The experimental results for sequence 5: AP (%).

Detector	YOLO V3	RFBnet	RefineDet	ISTDet*	ISTDet
Input	256	300	320	256	256
AP	80.12	84.16	88.35	83.65	92.76
FPS	62.97	71.59	45.74	163.51	76.11

3.4. Performance comparison

Tables 4–8 show the comparative results on five infrared small target sequences between ISTDet and the state-of-the-art target detection approaches. YOLO V3, RFBnet and RefineDet are the representative target detectors, which achieve good performance for generic target detection. It is noted that the abovementioned three target detection approaches all have a scale for small target detection. YOLO V3 adopts the idea of Feature Pyramid Network [31]. RFBnet is a fast and powerful target detector, which simulates the structure of receptive field in human visual systems [18]. RefineDet is a single-shot refinement CNN-based target detector, which owns the advantages of two-stage target detection approach and one-stage target detection approach [19]. ISTDet* represents the version without image filtering module. ISTDet represents the version with image filtering module.

3.4.1. Quantitative evaluation

As shown in Tables 4–8, compared with YOLO V3, RFBnet and RefineDet, the AP has been enhanced by 21.32%, 17.49% and 5.77% respectively in sequence 1. In sequence 2, the AP has been enhanced by 14.87%, 17.22% and 5.18% respectively. In sequence 3, the AP has been improved by 11.61%, 9.90% and 6.24%, respectively. In sequence 4, the AP has been improved by 9.25%, 5.84% and 4.41%, respectively. In sequence 5, the AP has been improved by 12.64%, 8.60% and 4.41%, respectively. ISTDet outperforms the other target detection approaches in sequence 1, 2, 3, 4 and 5. That is because adaptive receptive field fusion module has increased the context information around the small targets and spatial attention mechanism has made ISTDet have a global contextual view. What is more, image filtering module has effectively suppressed the response of background and enhanced the response of infrared small targets. The detection speed of ISTDet is about 75 FPS, which is the fastest of all. It is almost 1.2 times as fast as YOLO V3 and 1.7 times as fast as RefineDet. The reason can be concluded that ISTDet uses only 8 × down-sampled feature maps to detect targets because it focuses on infrared small target detection. However, YOLO V3, RFBnet and RefineDet use multiple scales to detect targets because they focus on generic target detection.

Table 9
Comparison of recall rate and precision rate between YOLO V3 and ISTDet for sequence 1.

Detector	X_{TP}	X_{FP}	X_{FN}	*Precision/%*	*Recall/%*
YOLO V3	404	71	76	85.05	84.17
ISTDet	462	15	18	96.85	96.25

Table 10
Comparison of recall rate and precision rate between YOLO V3 and ISTDet for sequence 2.

Detector	X_{TP}	X_{FP}	X_{FN}	*Precision/%*	*Recall/%*
YOLO V3	98	20	23	83.05	80.99
ISTDet	101	13	20	88.60	83.47

Table 11
Comparison of recall rate and precision rate between YOLO V3 and ISTDet for sequence 3.

Detector	X_{TP}	X_{FP}	X_{FN}	*Precision/%*	*Recall/%*
YOLO V3	103	16	17	86.55	85.83
ISTDet	112	8	8	93.33	93.33

Table 12
Comparison of recall rate and precision rate between YOLO V3 and ISTDet for sequence 4.

Detector	X_{TP}	X_{FP}	X_{FN}	*Precision/%*	*Recall/%*
YOLO V3	54	5	9	91.52	85.71
ISTDet	58	4	5	93.54	92.06

Table 13
Comparison of recall rate and precision rate between YOLO V3 and ISTDet for sequence 5.

Detector	X_{TP}	X_{FP}	X_{FN}	*Precision/%*	*Recall/%*
YOLO V3	69	9	11	88.46	86.25
ISTDet	77	6	3	92.77	96.25

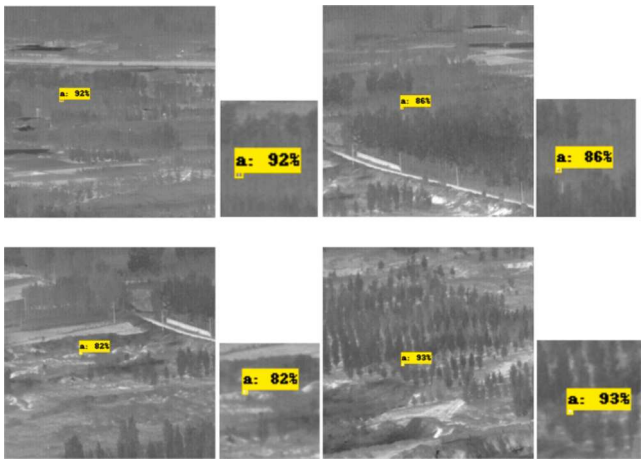

Fig. 7. Visual detection results of ISTDet in Sequence 1.

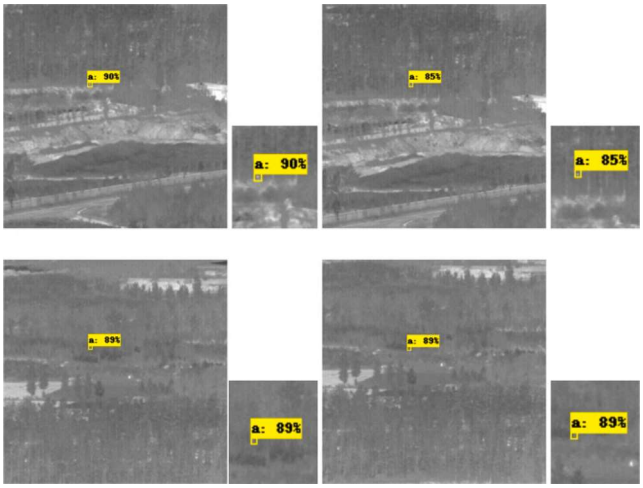

Fig. 8. Visual detection results of ISTDet in Sequence 2.

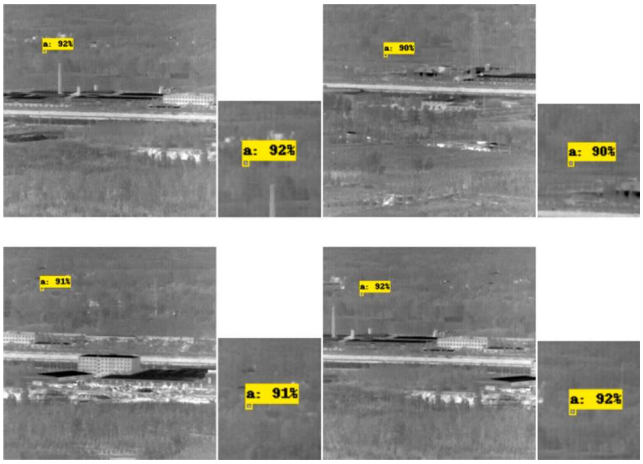

Fig. 9. Visual detection results of ISTDet in Sequence 3.

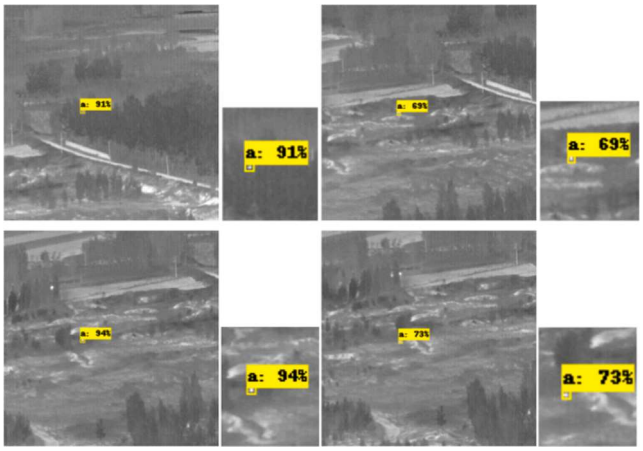

Fig. 10. Visual detection results of ISTDet in Sequence 4.

To further evaluate the performance of ISTDet, we also compare the recall rate and precision rate between ISTDet and YOLO V3, as shown in Tables 9–13. ISTDet outperforms YOLO V3 both on recall rate and precision rate in 5 real infrared small target sequences. The comparative results demonstrate that ISTDet is an efficient CNN-based target detector for infrared small target detection.

Fig. 11. Visual detection results of ISTDet in Sequence 5.

3.4.2. *Qualitative evaluation*

To qualitatively evaluate the detection performance of ISTDet, Figs. 7–11 show the visualization of the detection results of ISTDet in 5 real infrared small target sequences. For easy observation, we enlarged the area around the infrared small target. ISTDet is able to detect the infrared small targets accurately even if the small target is dim and the environment is complicated. The visual results indicate that ISTDet has good detection performance on infrared small targets.

To sum up, both the quantitative evaluation and the qualitative evaluation indicate that ISTDet is an efficient CNN-based infrared small target detector. In addition, the ablation study demonstrates that the various design of ISTDet is helpful to improve the detection performance for infrared small targets.

3.5. *Discussion*

ISTDet is an end-to-end CNN-based infrared small target detector, which combines the image filtering task and target detection task into one network. The experimental results on 5 real infrared small target sequences demonstrate ISTDet has good detection performance for infrared small targets. When the resolution of input image is 256×256, the detection speed can almost reach 77 FPS, which is suitable for real-time infrared small target detection. By both quantitative and qualitative evaluation, ISTDet outperforms other target detectors. That is because the IFM in ISTDet is helpful to enhance the response of infrared small targets and suppress the response of background. By introducing adaptive receptive field fusion module and spatial attention mechanism, the detection performance of ISTDet has been further improved. We hope our proposed method can inspire related work on infrared small target detection.

The CNN-based target detectors mainly depend on the datasets. However, there are relatively few open infrared small target datasets which limits the development of the CNN-based target detectors. Hence, it is essential to build more open dataset for infrared small target detection. In the future, we will do deep research on unsupervised CNN-based infrared small target detection methods.

4. Conclusion

In this paper, an efficient CNN-based target detector, ISTDet, is introduced to perform infrared small target detection. ISTDet is a novel single-shot based detector, composed of two inter-connected modules, namely the IFM and the ISTDM. The IFM is proposed to obtain the confidence map, aiming to enhance the response of infrared small targets and suppress the response of background. We use the confidence map to activate the input infrared image by element-wise multiplication. The ISTDM takes the activated infrared image as input and speculates

the category and the position of the infrared small targets. The ISTDet is trained in an end-to-end way by multi-task loss. Finally, we conduct comparative experiments on five infrared small target sequences to demonstrate the detection performance of ISTDet. The experimental results verify ISTDet has better speed-accuracy trade-off compared with other target detectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] X. Zhang, Q. Ding, H. Luo, et al., Infrared dim target detection algorithm based on improved LCM, Infrared Laser Eng. 46 (7) (2017) 0726002.
[2] V. Tom, T. Peli, M. Leung, J.B. Ondaryk, Morphology-based algorithm for point target detection in infrared backgrounds, Proc. SPIE 1954 (10) (1993) 2–11.
[3] X. Bai, F. Zhou, Analysis of new top-hat transformation and the application for infrared dim small target detection, Pattern Recogn. 43 (6) (2010) 2145–2156.
[4] S.D. Deshpande, M.H. Er, R. Venkateswarlu, P. Chan, Maxmean and max-median filters for detection of small targets, Proc. SPIE 3809 (1999) 74–83.
[5] C.L.P. Chen, H. Li, Y. Wei, T. Xia, Y.Y. Tang, A local contrast method for small infrared target detection, IEEE Trans. Geosci. Remote Sensing 52 (1) (2014) 574–581.
[6] Y. Chen, Y. Xin, An effcient infrared small target detection method based on visual contrast mechanism, IEEE Geosci. Remote Sens. Lett. 13 (7) (2016) 962–966.
[7] J. Han, Y. Ma, B. Zhou, F. Fan, K. Liang, Y. Fang, A robust infrared small target detection algorithm based on human visual system, IEEE Geosci. Remote Sens. Lett. 11 (12) (2014) 2168–2172.
[8] C. Gao, D. Meng, Y.i. Yang, Y. Wang, X. Zhou, A.G. Hauptmann, Infrared patch-image model for small target detection in a single image, IEEE Trans. Image Process. 22 (12) (2013) 4996–5009.
[9] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proc. NIPS, 2012, pp. 1097–1105.
[10] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchiesfor accurate object detection and semantic segmentation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 580–587.
[11] R. Girshick, Fast R-CNN, in: Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1440–1448.
[12] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149.
[13] Z. Cai, N. Vasconcelos, Cascade R-CNN: delving into high quality object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[14] W. Liu, et al., SSD: Single shot multibox detector, Computer Vision ECCV (2016) 21–37.
[15] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 779–788.
[16] J. Redmon, A. Farhadi, YOLO 9000: Better, faster, stronger, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6517–6525.
[17] J. Redmon, A. Farhadi. "YOLOv3: An incremental improvement." [Online]. Available: https://arxiv.org/abs/1804.02767.
[18] S. Liu, D. Huang, Receptive field block net for accurate and fast object detection, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 385–400.
[19] S. Zhang, L. Wen, X. Bian, Z. Lei, S.Z. Li, Single-shot refinement neural network for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4203–4212.
[20] B. Jiang, X. Ma, Y. Lu, Y. Li, L. Feng, Z. Shi, Ship detection in spaceborne infrared images based on Convolutional Neural Networks and synthetic targets, Infrared Phys. Technol. 97 (2019) 229–234.
[21] S. Chen, Z. Chen, X. Xu, N. Yang, X. He, Nv-Net: efficient infrared image segmentation with convolutional neural networks in the low illumination environment, Infrared Phys. Technol. 105 (2020), 103184.
[22] M. Shi, H. Wang, Infrared dim and small target detection based on denoising autoencoder network, Mobile Netw. Appl. 25 (4) (2020) 1469–1483.
[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[24] P. Hu, D. Ramanan, Finding tiny faces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 951–959.
[25] Y. Li, Y. Chen, N. Wang, Z. Zhang, Scale-aware trident networks for object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6054–6063.
[26] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in: International Conference on Machine Learning, 2019, pp. 7354–7363.
[27] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, IEEE Trans. Pattern Anal. Mach. Intelligence (2017) 2999–3007.
[28] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 658–666.
[29] B. Hui, Z. Song, H. Fan, et al., A dataset for dim-small target detection and tracking of aircraft in infrared image sequences, China Scientific Data (2019), https://doi.org/10.11922/csdata.2019.0074.zh.
[30] I. Loshchilov, F. Hutter, SGDR: stochastic gradient descent with warm restarts, [Online]. Available: https://arxiv.org/abs/1608.03983.
[31] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.