Check for updates

Received: 23 September 2022 | Revised: 13 December 2022 | Accepted: 28 December 2022

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12165

ORIGINAL RESEARCH

SDDNet: Infrared small and dim target detection network

Ma Long[1,2] | Shu Cong[1,2] | Huang Shanshan[1,2] | Wei Zoujian[1,2] | Wang Xuhao[1,2] | Wei Yanxi[3]

[1]School of Computer Science and Engineering, Xi'an Technological University, Xi'an, Shaanxi, China

[2]State and Local Joint Laboratory of Advanced Network and Monitoring, Xi'an, Shaanxi, China

[3]School of Computing University of Kent, Canterbury, UK

Correspondence

Ma Long, School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, Shaanxi, China.
Email: malong@xatu.edu.cn

Abstract

This study focuses on developing deep learning methods for small and dim target detection. We model infrared images as the union of the target region and background region. Based on this model, the target detection problem is considered a two-class segmentation problem that divides an image into the target and background. Therefore, a neural network called SDDNet for single-frame images is constructed. The network yields target extraction results according to the original images. For multiframe images, a network called IC-SDDNet, a combination of SDDNet and an interframe correlation network module is constructed. SDDNet and IC-SDDNet achieve target detection rates close to 1 on typical datasets with very low false positives, thereby performing significantly better than current methods. Both models can be executed end to end, so both are very convenient to use, and their implementation efficiency is very high. Average speeds of 540+/230+ FPS and 170+/60+ FPS are achieved with SDDNet and IC-SDDNet on a single Tesla V100 graphics processing unit and a single Jetson TX2 embedded module respectively. Additionally, neither network needs to use future information, so both networks can be directly used in real-time systems. The well-trained models and codes used in this study are available at https://github.com/LittlePieces/ObjectDetection.

KEYWORDS

deep learning, detection of moving objects

1 | INTRODUCTION

When sensors are used for long-distance detection, the targets of interest often have a small area and a low signal-to-noise ratio; this type of target is called a small and dim target. Small and dim target detection in complex backgrounds is a classic problem in automatic target recognition (ATR), which includes early warnings, space debris finding, and range measurements. In recent years, with the increasing number of consumer unmanned aerial vehicles, small and dim target detection has become a key area of interest for low altitude security and key area protection.

In this work, we focus on extremely small and dim infrared targets. The imaging area of this kind of target is smaller than 0.15% of the whole image (for an image sized 256 × 256, the target area is generally less than 9 × 9 pixels), and the signal-to-clutter ratio is less than 4 dB [1, 2]. Aircraft and other air flying targets are typical examples of these targets. This kind of target detection is very difficult mainly for the following reasons: Because of the long imaging distance, the shape, texture and other features of the target are basically lost, and the targets always appear as small and dim speckles in the visual field. Additionally, because infrared images are thermal images, the colour information of the target is lost. Infrared (IR) targets are usually considered to be unrelated to the surrounding backgrounds and are categorised as high-frequency images [3]. However, in heterogeneous scenes, such as complex clouds and ground backgrounds, there is always more than one strong IR radiation source. In such a highly cluttered environment, the high-frequency characteristics of the target are no longer

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

salient. The two aforementioned factors cause the apparent characteristics of the target to be very insignificant, and it is very difficult to detect these targets; regular detection approaches for optical images are no longer valid for IR images [1]. Currently, infrared small and dim target detection is still an open problem. In recent years, great success has been achieved in computer vision with deep learning technology, such as image classification, facial recognition, optical object detection, and tracking. A main reason for the effectiveness of this technology is that deep neural networks have very strong feature extraction abilities; these networks can learn effective multilevel features for visual tasks. Inspired by these results, this paper proposes a learning method for very small and dim target detection. We try to use the powerful feature extraction abilities of deep neural networks to capture effective features from the insignificant appearance of extremely small and dim infrared targets to improve the corresponding detection accuracy. We construct a neural network called SDDNet (small and dim target detection network) for single-frame images. The network yields target extraction results according to the original images. For multiframe images, we construct a network called IC-SDDNet, a combination of SDDNet and an interframe correlation (IC) network module. SDDNet and IC-SDDNet achieve target detection rates close to 1 on typical datasets with very low false positives, thereby performing significantly better than current methods. Both models can be executed end to end, so both are very convenient to use, and their implementation efficiency is very high. Additionally, neither network needs to use future information, so both networks can be directly used in real-time systems.

The rest of this paper is organised as follows. The previous works are presented in Section 2. The target detection problem is described in Section 3. Our methods are proposed in Section 4, where SDDNet is constructed in Subsection 4.1, the IC module is developed in Subsection 4.2, and the postprocessing methods are described in Subsection 4.3.

The network training details are described in Section 5. The experimental results are given in Section 6, and the conclusions of this work are given in Section 7.

2 | PREVIOUS WORKS

Small and dim target detection is a continuously popular research topic in ATR. The mainstream methods can be simply divided into the traditional and deep learning methods developed in recent years.

The traditional methods include the following categories. The first category includes background estimation and suppression methods [4–8], the basic flow of which was described in the next section. The second category includes contrast-based methods [9–16]. These methods assume that the target is significant in the feature space of intensity, scale or direction, and extract the most salient area in the image as the target. The third category includes infrared patch-image (IPI) model-based methods [1, 3]. These methods assume that the target patch image is a sparse matrix and that the background patch image is a low-rank matrix. The small target detection task is then transformed into the recovery of the low-rank and sparse matrices, which can be effectively achieved via an optimisation technique. The fourth category includes methods that use small target motion detectors. STMD-based models [17, 18], inspired by the insect motion-sensitivity mechanism, fuse information such as motion information and directional contrast to detect targets and eliminate fake targets. The target detection process of the above methods can be summarised as follows: first, the typical features of small and dim targets are defined, and a series of feature extraction operators are designed accordingly; second, the salient regions in the image are extracted as candidate targets by using these operators; finally, a series of prior decision criteria are used to screen the final target from the candidate targets.

Recently, with the development of deep learning, some neural networks with multiple layers have been used in small and dim target detection [19–24]. DRUNet [21] constructs neural networks to estimate the background and obtain the small target image by subtracting the background image from the original image. DNANet [23] designs a dense nested interactive module (DNIM) to achieve progressive interaction among high-level and low-level features. With the repeated interaction in DNIM, dim features of small targets can be maintained in deep layers. The local similarity pyramid model (LSPM) method [22] proposes a network that leverages the local similarity pyramid module and feature aggregate module to segment infrared small targets. The asymmetric contextual modulation (ACM) method [19] supplements a bottom-up modulation pathway based on point-wise channel attention for exchanging high-level semantics and subtle low-level details to better highlight small targets, in addition to top-down global contextual feedback. Similarly, CBP-Net [25] uses bottom-up and top-down information with a bidirectional pyramid structure. LPNet [26] jointly considers the global and local properties of infrared small target images for high-precision detection. IAANet [27] leverages a transformer encoder to obtain attention-aware features to determine whether a pixel belongs to the target or the background. MLCL-Net [28] builds a multiscale local contrast learning module to fully extract the target feature information. MDvsFA-cGAN [20] adopts a conditional generative adversarial network comprising two generators and one discriminator. Each generator strives for one subtask with each focussing on reducing either the number of miss detections or the number of false alarms (FAs), while the discriminator differentiates the three segmentation results from the two generators and the ground truth in the training stage.

The above methods are generally used to process a single-frame image. If multiple consecutive images are available, temporal information can be used to assist in target detection. One kind of multiframe method is 3D matched filtering [29]. This method performs moving target signature-matched filtering in the Fourier domain. The result is a set of matched-filter peaks indicating the detected target tracks with an enhanced signal-to-noise ratio (SNR), often far exceeding what would be normally obtained from separately spatially filtering each individual frame. A primary problem of 3-D

24682322, 2023, 4, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.12165 by Test, Wiley Online Library on [22/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

matched filtering is that the matched filter must be matched to a specific velocity profile, a known target, with a known intensity distribution moving at a known speed in a designated direction. Recently, a new multiframe method based on Markov random field (MRF) guided noise modelling was proposed [30]. This novel method treats small and dim targets as a special sparse noise component of complex background noise and adopts a mixture of Gaussian (MOG) with a MRF to model the detection problem. It is claimed that this method is robust against real infrared images with complex backgrounds.

This work models infrared images as the union of the target region and background region. Based on this model, the target detection problem is considered a two-class segmentation problem that divides an image into the target and the background. Therefore, we construct the segmentation neural network SDDNet for single-frame images and IC-SDDNet for multiframe images which can extract dim and small targets effectively and efficiently.

3 | DESCRIPTION

Generally, the infrared image model can be formulated as [1, 3].

$$f(x,y) = f_T(x,y) + f_B(x,y) + n(x,y) \tag{1}$$

where $f(x,y)$ is the intensity of pixel (x,y) and $f_T(x,y), f_B(x,y)$, and $n(x,y)$ are the intensity of the target, background, and noise, respectively, at (x,y). Without considering noise, the intensity of any pixel in the infrared image can be expressed by the sum of the infrared radiation intensity of the target and the background at that point. Many traditional detection methods, such as background estimation and suppression methods, are derived from this conclusion. These methods first adopt smoothness filters [31] to estimate the current background pixels by using their neighbouring areas; this step is referred to as background estimation. Second, according to (1), the potential targets can be detected by subtracting the estimated background from the original image [9]; this step is referred to as background suppression. The disadvantage of this method is that in background suppression, a background estimate (usually not 0) is subtracted from the target, thus weakening the strength of the target. When the background estimation method and suppression method are not carefully designed, the target saliency may decrease, thus reducing the detection accuracy.

In fact, when (x,y) is in the background area, the target radiation does not contribute to its intensity, so $f_T(x,y) = 0$, $f(x,y)$ is determined by the background radiation; when (x,y) is in the target area, the background radiation of this point is blocked by the target, so $f_B(x,y) = 0, f(x,y)$ is determined by the target radiation. However, Equation (1) easily leads to the misconception that both the target radiation and background radiation contribute to the imaging intensity of the target. To more clearly describe the relationship between the target and the background, we believe that the whole infrared image region r should be modelled as the union of the target regions r_T and the background regions r_B. That is,

$$r = r_T \cup r_B \tag{2}$$

According to the above equation, the problem of infrared dim small target detection can be regarded as a two-class segmentation problem that divides an image into target regions and background regions. Based on this analysis, a deep neural small and dim target detection network The SDDNet is constructed to detect small and dim targets end to end.

4 | METHODS

4.1 | SDDNet

SDDNet is essentially a two-class segmentation network. This network separates the target region from the background region in the image. The input is an image, and the output is the target-background segmentation map of the same size as the input image. The probability that a pixel belongs to the target is given in the segmentation map. Considering the operation efficiency, the structural design of SDDNet is very simple (Figure 1). The network uses as few network layers as possible and does not use the strategies of multiresolution processing or multilayer feature fusion, which are commonly used in image segmentation. This makes the SDDNet processing speed, which reaches more than 540 FPS on a typical single graphics processing unit (GPU), very fast. Additionally, the network is very suitable for small and dim target extraction; the network

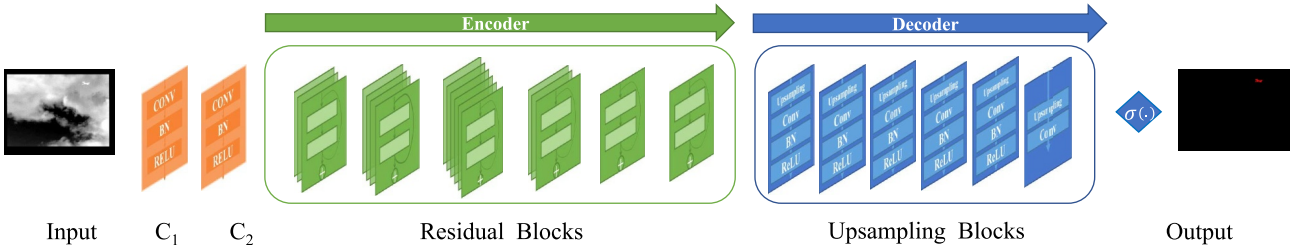

FIGURE 1 SDDNet. SDDNet adopts the encoder-decoder structure. The encoder, which extracts the features, adopts mainly the stack residual block structure. The decoder, which segments the target, adopts mainly the stack upsampling block structure. Starting from the end of the encoder, we adopt a layer-by-layer upsampling technique of nearest-neighbour interpolation following the convolution operation. By upsampling 6 times in succession, we obtain the final output of a single-channel segmentation map of the same size as the input IR image.

24682322, 2023, 4, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.12165 by Test, Wiley Online Library on [22/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

obtains a probability of detection (PD) close to 1 and maintains a very low false alarm (FA) rate.

SDDNet adopts the encoder-decoder structure commonly used in image segmentation networks [32–34]. The encoder, which extracts the features, adopts mainly the stack residual block structure, which is widely considered to have strong feature extraction ability [35]. We hope that this structure can extract salient features that are conducive to target segmentation from the indistinct appearance of small and dim targets. The decoder, which segments the target, adopts mainly the stack upsampling block structure. The decoder increases the size of the feature map layer by layer and ultimately yields a segmented image of the same size as that of the input image. Starting from the end of the encoder, we adopt a layer-by-layer upsampling technique of nearest-neighbour interpolation following the convolution operation. The relationship between the number of channels and the size of the feature map in these upsampling layers is as follows: whenever the size of the feature map doubles, the number of channels is reduced by half to maintain the scale of the information. This technique has also been used in classic networks, such as [35–37]. By upsampling six times in succession, we obtain the final output of a single-channel segmentation map of the same size as the input IR image.

Except the output layer, all the network layers use the rectified linear unit activation function to increase the networks' ability to perform non-linear fitting. The output layer uses the sigmoid activation function with a value range of (0, 1), which is consistent with the value range of the normalised labels used when training the networks. To prevent the loss of target information, we do not use the pooling layer in the whole network but use a convolution layer with a stride of 2 to reduce the dimension.

The small and dim target detection network will be trained under the supervised learning framework. We hope that the network can learn how to extract and enhance the characteristics of small and dim targets from the strong clutter background under the guidance of labels and finally accurately segment the targets. Therefore, the design of the network should be conducive to the end-to-end transmission of important information, avoid the forced abandonment of insignificant information that may contain the characteristics of small and dim targets, and be conducive to the integration of important information. Our SDDNet just fully meets the above requirements. First, in the feature extraction stage, the residual block adopts the shortcut structure [35], which is conducive to the smooth transmission of important information in the network. No pooling layer is used throughout the network, thus preventing important information from being discarded. The above two points ensure that which feature the network extracts completely depends on the guidance of supervision learning and avoid inaccurate feature extraction caused by information attenuation or incorrect abandonment. Second, in the target extraction stage, we use an encoding network composed of a series of upsampling layers. After supervised learning, the upsampling layers can reasonably fuse, filter and enhance the encoding features layer by layer; this process is conducive to the output of an accurate segmentation map.

4.2 | Interframe correlation module and IC-SDDNet

In practical applications, multiple consecutive frames of the target, that is, image sequences, can often be obtained. The intensity, shape and motion track of the target over a certain period are recorded in the image sequence. Compared with a single image, the image sequence provides more information. If this information can be effectively used, the target detection accuracy can be further improved. In this work, a neural network-based IC module is designed to capture and fuse the information. When consecutive multiframe images are available, this module can be connected to SDDNet to help effectively suppress the background clutter and extract the real target.

The IC module is used when consecutive multiframes are available. The function of this module is to use the segmentation results of N consecutive frames, provided by SDDNet, to modify the segmentation results of the current image. In the real-time system, when the current time is t, only the image at t and those images before t can be used. Therefore, our IC module uses only the N frames from $t-N+1$ to t. The network still uses an encoder-decoder architecture similar to that of SDDNet but has a more refined structure. The encoder uses only three residual blocks, which are used mainly for feature extraction and fusing the multiframe segmentation results; the decoder uses the same number of upsampling blocks, which are used mainly to increase the size of the feature map to the input image size layer by layer and to output the modified current segmentation results.

In the training stage, we input the original segmentation of N consecutive frames, including the current frame, into the module and require the module to output an accurate segmentation map of the current frame. We hope that the network can simultaneously learn the spatial and temporal characteristics of the real target and then realise the following functions through training (as shown in Figure 2a–c). 1. The module can eliminate the background clutter (usually called FAs) that can be incorrectly classified as the target due to the high similarity of the background clutter to the target in the single frame detection result. 2. For targets missed due to their weak

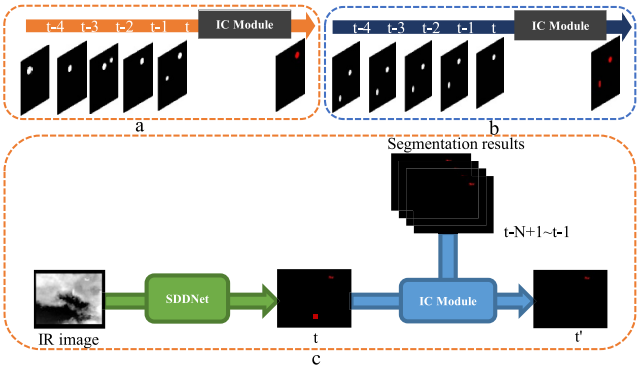

FIGURE 2 (a) Eliminating false alarm (FA) areas. (b) Filling in missed targets. (c) The information processing of IC-SDDNet.

24682322, 2023, 4, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.12165 by Test, Wiley Online Library on [22/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

appearance in the single frame detection results (usually called missed detections), the module can fill in these areas.

In use, the IC module can be directly connected to SDDNet to form IC-SDDNet. IC-SDDNet uses the image t and the previously saved segmentation map of $t-N+1$ to $t-1$ to output the corrected segmentation map of t in real time. The information processing of IC-SDDNet is shown in Figure 2c.

4.3 | Postprocessing

SDDNet and IC-SDDNet construct the segmentation map, and the probability that each pixel is part of the target is given on the segmentation map (in the ideal state, the target pixel probability is 1, and the background pixel probability is 0).

In practical applications, it is necessary to provide the location, size, and other information on the target. We extract this information from the segmentation map in two steps. First, we use the threshold Tr to binarise the segmentation map to obtain a black-white map. Then, we label the connected components and obtain the position and size of the target from the black-white map by using a block-based decision tree (BBDT) [38], the code for which is provided by OpenCV. Based on this information, we give the bounding box of the target. The BBDT exploits and extends the decision table formalism for connected component labelling by introducing OR-decision tables, in which multiple alternative actions are managed. An automatic procedure to select the most convenient alternative is proposed to obtain a single-entry decision table, and finally, a Boolean optimisation algorithm is adopted to automatically produce the optimal decision tree in terms of the number of evaluations. The BBDT requires very little memory access and computation, thereby making our postprocessing very efficient.

5 | TRAINING

We collect approximately 100 infrared image sequences, including approximately 80,000 images that contain small and dim targets. These sequences contain small and dim targets with various intensities in various scenes. We reserve 6 sequences of typical scenes as test sets and other sequences as training sets (Figure 3).

When training SDDNet, we extract 10% of the samples from the training set as the validation set. We label all the image sequences and train the network by using supervised learning. We input an infrared image and hope the network outputs the correct segmentation map. We calculate the loss between the output of the network and the truth segmentation and update the weight of the network by using the optimiser through the back-propagation process.

The IC module needs to take the current and the previous consecutive N-1 segmentation results as the input and output the modified current segmentation results. The training samples are obtained by SDDNet through cross testing. We divide the training set into 10 groups; one group is selected each time to be reserved data, and other data are used to train and validate SDDNet. Every time an SDDNet model is sufficiently trained, we use the model to process the reserved data to obtain the segmentation result of the reserved data. By collecting the segmentation results of each reserved dataset, we obtain enough samples for training IC modules. We use the initial segmentation results of N consecutive frames as input, truth segmentation as labels, and supervised learning to train the IC module.

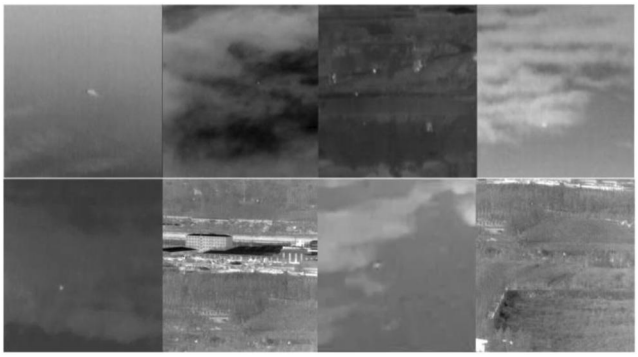

FIGURE 3 Samples of training data

Considering the problem of class imbalance between the target and background, we use the weighted binary cross entropy function when training the SDDNet and IC modules. For each pixel, its loss value

$$l = w \cdot y \cdot \log x + (1-w)(1-y)\log(1-x) \qquad (3)$$

where x is the output of the networks, y is its corresponding label, and w is the proportion of the background in the image, which is set to 0.9859 in this work according to the statistical results. Through the adjustment of w, the target loss and the background loss in an image are roughly equivalent. This approach can avoid the domination of the training process by the background, which is often much larger than small and dim targets, to better learn the characteristics of the target and improve the effectiveness of the network.

Training is performed on an NVIDIA DGX station with 4 T V100 GPUs in a distributed-data-parallel framework. We implement all of our models by using an optimised PyTorch deep learning framework in NVIDIA GPU-accelerated containers. The detailed training strategies we use are as follows: we flip the training image to the left-right for data enhancement to expand the training set. On each GPU, we process data with a minibatch size of 32, adopting batch normalisation immediately after each convolution and before activation and training both networks from scratch. For optimisation, we use adaptive moment estimation (Adam) [39] as a solver, with b1 and b2 set to 0.5 and 0.999 respectively. In the SDDNet and IC module trainings, the learning rates are set to 0.0001 and 0.001 respectively. Both of the networks are trained without dropout. We keep track of the validation error across iterations, and at the end of training, we use the weights that yield the lowest validation error.

24682322, 2023, 4, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.12165 by Test, Wiley Online Library on [22/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

6 | EXPERIMENTAL RESULTS

We use six image sequences of different scenes (see the first line of Figure 4 for example) to test the proposed method. In addition to the four datasets we collected, two public sets of representative small targets within natural scene datasets were used [40]. These datasets contain serious cloud or ground background clutter or both or contain considerable noise, which poses great challenges to the performance of the detection methods. Table 1 shows the descriptions of the six image sequences and their local signal-to-clutter ratio (LSCR) and global signal-to-clutter ratio (GSCR), which are defined as [3, 39].

$$LSCR = \frac{|\mu_T - \mu_B|}{\sigma_L} \tag{4}$$

$$GSCR = \frac{|\mu_T - \mu_B|}{\sigma_G} \tag{5}$$

where μ_T is the average intensity value of the target; μ_B and σ_L are the average intensity value and the standard deviation of the pixels in the neighbouring area (two times the target area) around the target, respectively; and σ_G is the background standard deviation of the whole image.

Both the PD and the FA rate are used to evaluate the performance of small target detection methods. The PD denotes the probability that targets are detected in images where targets truly exist, while the FA rate denotes the rate at which targets are detected in the images where targets do not exist. The PD and FA rate are described as follows [1, 9]:

$$PD = \frac{\#\text{ of true targets detected}}{\#\text{ of total true targets}} \tag{6}$$

$$FA = \frac{\#\text{ of false pixels detected}}{\#\text{ of total pixels in image}} \tag{7}$$

The detected result is considered correct if the pixel distance between the ground truth and the result is within a threshold T. We set T to 1.5e-5 times the image area. For an image of 256 × 256, $T \approx 1$ pixels.

6.1 | Results of SDDNet

Figure 4 shows the typical results obtained by SDDNet on six test datasets. The first line is the original infrared image fed into SDDNet, and the second line is the image segmentation

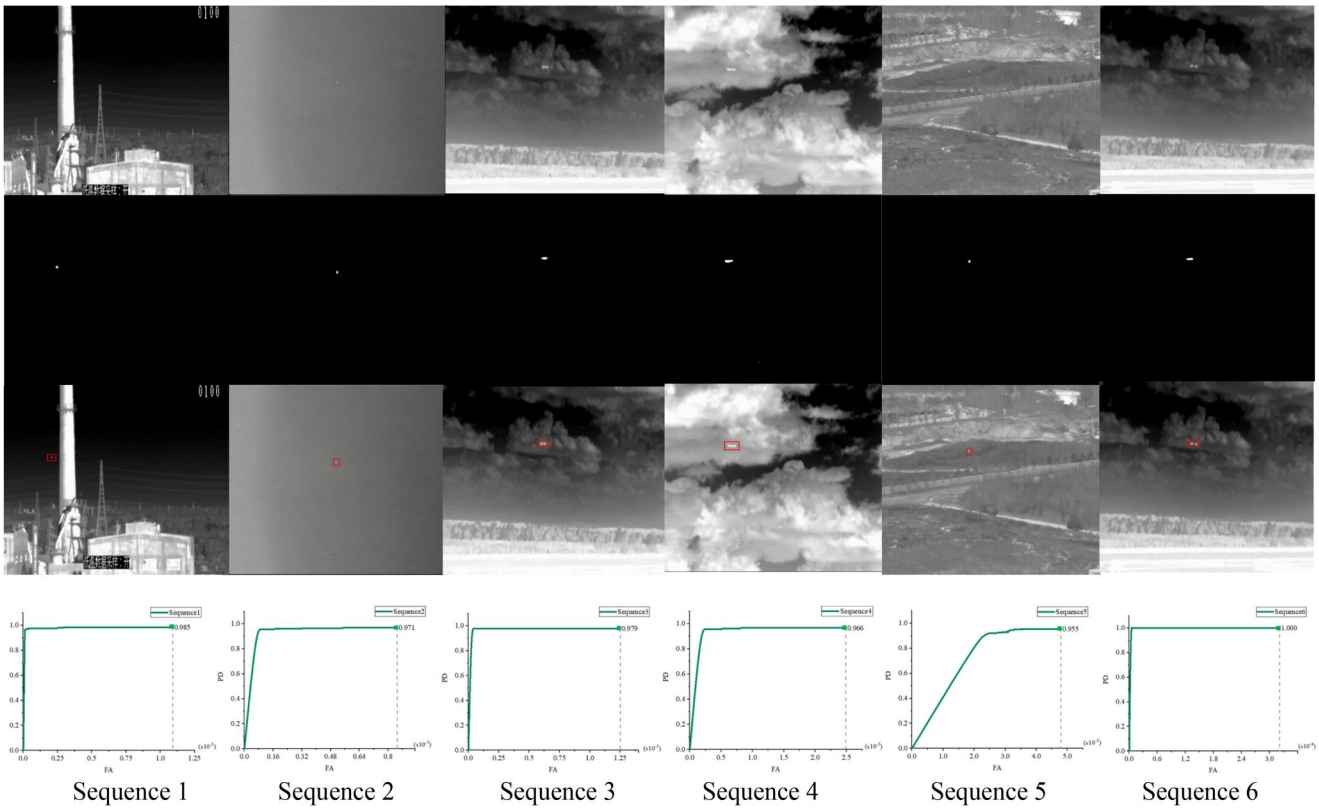

FIGURE 4 Typical results of SDDNet. The first line is the original infrared image, the second line is the image segmentation result, the third line is the detection result where target is indicated with bounding boxes, and the fourth line shows the receiver operating characteristic (ROC) curves with respect to six sets of data.

24682322, 2023, 4, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.12165 by Test, Wiley Online Library on [22/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

result output by SDDNet. The probability that each pixel belongs to the target is given in the image segmentation results. We need to binarise the image with a threshold to extract the target. The fourth line of Figure 4 shows the receiver operating characteristic (ROC) curves with respect to six sets of data. From the beginning of the lower FA rate, the PD reaches a relatively stable high level close to 1. The upper limit of the FA rates of existing detection methods is generally equal to 1, which corresponds to the lowest threshold. However, our FA rate has a very low upper limit (illustrated by the point at the end of the ROC curve in Figure 4) because the segmentation result output by SDDNet is very accurate. In this segmentation map, most of the background clutter is completely filtered and removed, and the value of these areas is determined to be 0. Therefore, even if $Tr = 0$, the final FA rate remains very low. However, the effect of the existing methods on background clutter suppression is often not very good, so the FA rate upper limit is high. In practical applications, the optimal threshold can be determined by the ROC. Here, to objectively evaluate our method, we use an empirical threshold [11, 12]:

$$Tr = m + 0.5 * (maxv - m) \tag{8}$$

where m is the average intensity of all pixels in the image to be segmented and $maxv$ is the maximum intensity of all pixels in the image to be segmented. Table 2 shows the results of our

TABLE 1 Descriptions of the test videos

Datasets	Resolution	Target size	LSCR|GSCR	Target details	Background details
Sequence 1	640 × 480	7 × 7	0.114|0.008	• Long imaging distance. • Small size and minimal changes. • Continuous movement in a straight line.	• Background with some buildings. • Obvious obstructions.
Sequence 2	320 × 256	7 × 7	3.158|0.243	• Long imaging distance. • Small size and many changes. • Fluctuation with much noise.	• Single background. • Heavy noise.
Sequence 3	640 × 480	11 × 11 11 × 11	0.158|0.013 0.146|0.013	• Long imaging distance. • Small size and many changes. • Continuous motion. • Similarity between two targets.	• Very cloudy sky as background clutter. • Background of open space boundary.
Sequence 4	640 × 480	28 × 10	0.068|0.011	• Long imaging distance. • Irregular shape. • Flying in a straight line.	• Very cloudy sky as background clutter.
Sequence 5	256 × 256	3 × 3	0.389|0.297	• Remote imaging distance. • Very small size.	• Background of complicated buildings. • Heavy noise.
Sequence 6	256 × 256	3 × 3	1.764|0.111	• Remote imaging distance. • Very small size.	• Complicated Earth background. • Close to target.

TABLE 2 Comparison of the results of the single frame methods (The best results are marked in bold and the second-best results are marked in italic)

PD, FA	Years	Sequence 1	Sequence 2	Sequence 3	Sequence 4	Sequence 5	Sequence 6
IPI [3]	2013	(0.633, 8e-4)	(0.398, *4.0e-5*)	(0.883, 4.4e-5)	(0.853, 5e-5)	(0.745, 1.2e-4)	**(1.0, 3.0e-6)**
RW [40]	2019	(0.015, ***1e-5***)	(0.010, 7.0e-5)	(0.936, 5.5e-5)	(0.784, 2.2e-4)	(0.325, 6.2e-3)	(0.730, 2.3e-4)
LCM [41]	2014	FAIL	(0.189, 2.6e-3)	FAIL	FAIL	(0.325, 6.5e-3)	(**1.0,** 1.7e-4)
MGDWE [14]	2016	FAIL	FAIL	FAIL	FAIL	(0.025, 9.0e-4)	(0.175, 2.6e-3)
DNANet [23]	2021	(0.784,5.58e-5)	(0.392,5.28e-5)	(0.883,1.98e-4)	(0.884, 6.55e-5)	(0.73,**1.91e-5**)	(1.0, 1.755e-5)
AGPCNet [24]	2021	(0.245,5.45e-4)	(0.158,7.01e-5)	(0.476,2.79e-4)	(0.548,2.27e-4)	FAIL	(0.92,2.4e-3)
LSPM [22]	2021	(0.834, 2.4e-4)	(0.869, 1.3e-4)	(0.992, 4.6e-4)	(0.925, 1.0e-3)	(0.28, 9.0e-4)	(1.0, 3.4e-4)
ACM [19]	2021	(0.339,9.3e-4)	(0.1,1.18e-4)	(0.533,4.27e-4)	(0.521,6.49e-4)	(0.251,4.64e-4)	(0.782,9.25e-5)
MDvsFA-cGAN [20]	2019	(0.518, 4.8e-3)	(0.377, 4.0e-4)	(0.990, 4.1e-4)	(0.88, 7.1e-4)	FAIL	(0.305, 7.1e-4)
SDDNet (Ours)	-	**(0.985, 2.0e-6)**	**(0.966, 3.0e-6)**	(*0.979, 1.9e-5*)	**(0.966, 1.0e-5)**	(*0.950, 4.0e-5*)	(**1.0,** *5.0e-6*)

Abbreviations: ACM, asymmetric contextual modulation; IPI, infrared patch-image; LCM, local contrast method; LSPM, local similarity pyramid model; MGDWE, multiscale gray difference weighted image entropy; RW, random walker.

24682322, 2023, 4, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.12165 by Test, Wiley Online Library on [22/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

method with this threshold. Table 2 also shows the target detection results of the representative state-of-the-art methods, including four traditional methods, such as IPI [3], random walker [42], the local contrast method [43], and multiscale grey difference weighted image entropy [14], and five deep learning methods, such as DNANet [23], AGPCNet [24], ACM [19], LSPM [22], and MDvsFA-cGAN [20], using the same threshold calculation method. "FAIL" in the table indicates that most parts of the image are marked as target areas, so the FA rate is extremely high. As shown in the table, our method achieves the best results on most of the test data. For a few datasets, our PD or FA rate may not be the best, but the overall performance is still the best; that is, a high PD is obtained while keeping the FA rate low.

Additionally, we record the execution efficiency of SDDNet (Table 3). After acceleration with TensorRT, the average frame rate can reach more than 540 fps on a single NVIDIA Tesla V100 GPU; even on a single NVIDIA Jetson TX2 embedded artificial intelligence (AI) computing device, the average frame rate can reach more than 230 fps. Therefore, SDDNet is very efficient and can be implemented in real time. Additionally, we can see that the implementation speed of SDDNet is hardly affected by the complexity of the image scene; this is very beneficial to the stability of the detection system.

6.2 | Results of IC-SDDNet

To correct the results of the current segmentation, IC-SDDNet needs to use N consecutive segmentations from previous times. The input frame number N importantly affects

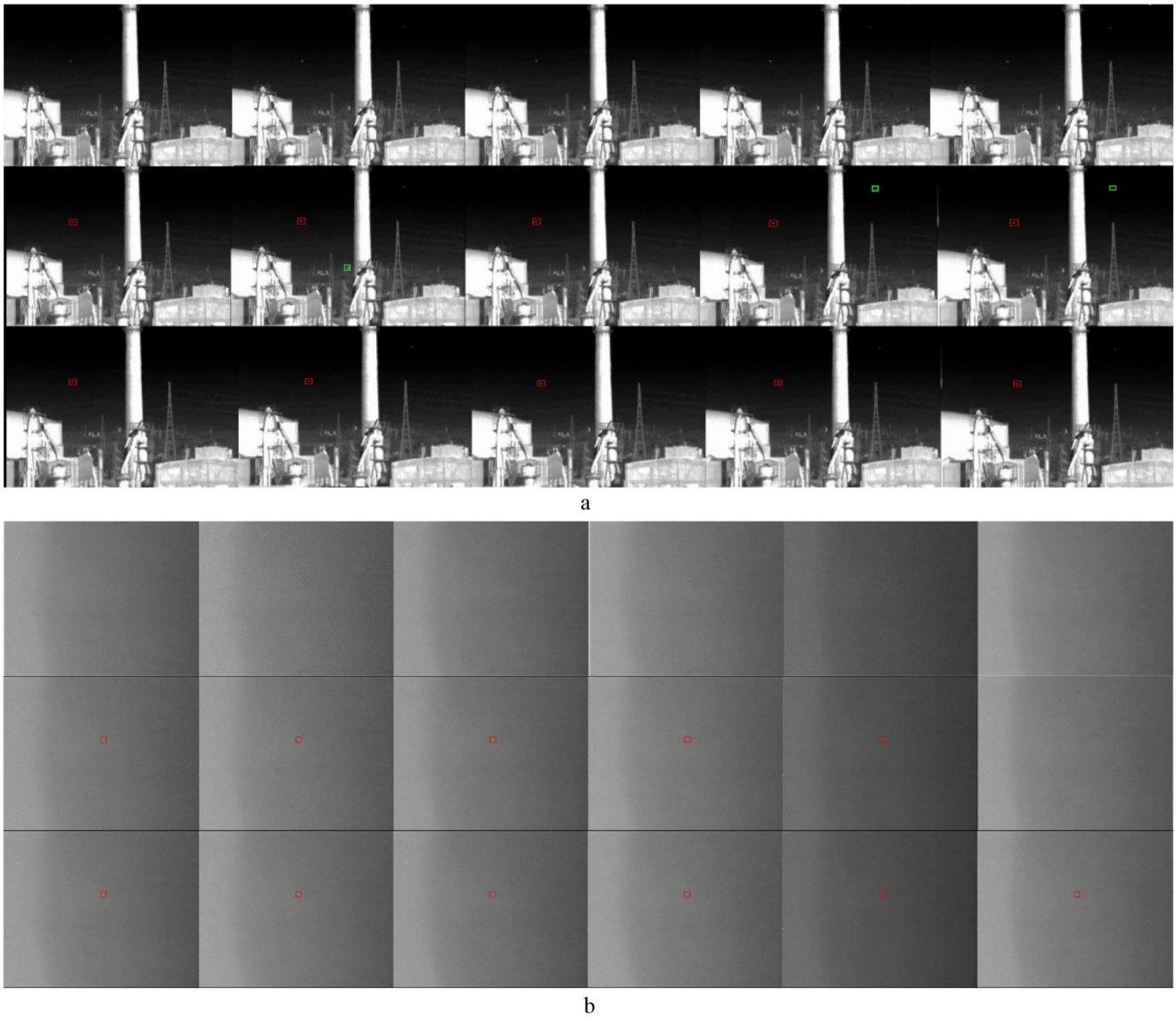

FIGURE 5 Results of IC-SDDNet and SDDNet. (a) Reduction in false alarm (FA) rate. (b) Filled in missing detection. The first line is the original image, the second line is the results of SDDNet, and the third line is the results of IC-SDDNet.

24682322, 2023, 4, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.12165 by Test, Wiley Online Library on [22/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the performance of the IC module. We recommend using $N = 5$ in the application according to our test. Figure 5 shows the target detection results obtained by postprocessing the output results of IC-SDDNet with $N = 5$. In Figure 5a, SDDNet mistakenly detects some background clutter that is extremely similar to the target radiation characteristics as a target, thus resulting in an FA; in Figure 5b, because the target radiation in some image frames is too weak, SDDNet fails to detect these targets, thus resulting in missed detection. In these test images, the detection results of IC-SDDNet do not show FAs or missed detections mainly because the IC module can eliminate false detection targets according to the space-time characteristics of the target and fill in some missed targets.

Table 4 shows the PDs and FA rates of the target detection results of different methods. Compared with the FA rates and PDs of SDDNet, the FA rates and PDs of IC-SDDNet are significantly lower and higher respectively. This indicates that the IC module can significantly reduce the FA rate and improve PDs. The results of another very recent multiframe detection method, MOG [30], are also given in Table 4. To our surprise, this method yields very poor results when our test data are used and even fails when other data are used, even though we have tried our best to tune the parameters. This approach may not be applicable to our scenario. Another typical multitarget detection method is the 3-D matched filtering method, but as described in the second section of this paper, a primary problem of 3-D matched filtering is that the matched filter must be matched to a specific velocity profile, a known target with a known intensity distribution moving at a known speed in a designated direction. For our test data, the target velocity is unknown, so we do not give the results of these methods.

Table 3 shows the execution time of IC-SDDNet. Compared with that of SDDNet, the execution speed of IC-SDDNet is slower mainly because of the execution efficiency of the IC module. However, after acceleration with TensorRT, the average frame rate still reaches more than 170 fps on a single NVIDIA Tesla V100 GPU and more than 60 fps on a single NVIDIA Jetson TX2 embedded AI computing device. This execution speed can still meet the requirements of most applications.

7 | CONCLUSION

In this work, we propose a deep learning-based method for small and dim target detection. We construct SDDNet, which can effectively detect targets in single-frame images; additionally, we construct an IC module that can improve the detection results of the current frame according to the historical target information. IC-SDDNet, which combines the IC module and SDDNet, can achieve end-to-end target detection, and the results are generally better than those of SDDNet. The target detection results of the two networks are obviously better than those of the state-of-the-art methods. Moreover, the two networks have very high efficiencies. On an embedded GPU platform, the execution speed of SDDNet and IC-SDDNet can meet the real-time requirements. Additionally, neither model uses any future information. Therefore, these models can be directly used in real-time systems; consequently, these models are very practical.

Although our networks have achieved good results, they still make some detection errors. Typical error results of SDDNet are shown in Figure 5, and typical error results of IC-SDDNet are given in Figure 6. Among these errors, the FA

TABLE 3 Time consumption comparison

FPS		Sequence 1	Sequence 2	Sequence 3	Sequence 4	Sequence 5	Sequence 6
SDDNet	Jetson TX2	276	248	236	249	244	270
	Tesla V100	617	704	546	544	693	689
IC module	Jetson TX2	85	83	82	87	86	85
	Tesla V100	269	263	265	265	235	280
IC-SDDNet	Jetson TX2	65	62	61	64	64	64
	Tesla V100	188	192	178	178	175	199

TABLE 4 Comparison of the results of the multiframe methods (The best results are marked in bold)

PD, FA	MOG [30]	SDDNet	IC-SDDNet
Sequence 1	(0.011, 1.2e-2)	(0.985, 2.0e-6)	**(0.990, 1.2e-6)**
Sequence 2	(0.652, 2.0e-5)	(0.966, 3.0e-6)	**(0.975, 1.1e-6)**
Sequence 3	(0.351, 4.4e-3)	(0.979, 1.9e-5)	**(0.979, 1.7e-6)**
Sequence 4	(0.098, 5.5e-3)	(0.966, 1.0e-5)	**(0.966, 4.7e-6)**
Sequence 5	FAIL	(0.950, 4.0e-5)	**(0.950, 1.6e-5)**
Sequence 6	FAIL	(1.0, 5.0e-6)	**(1.0, 2.0e-6)**

24682322, 2023, 4, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.12165 by Test, Wiley Online Library on [22/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

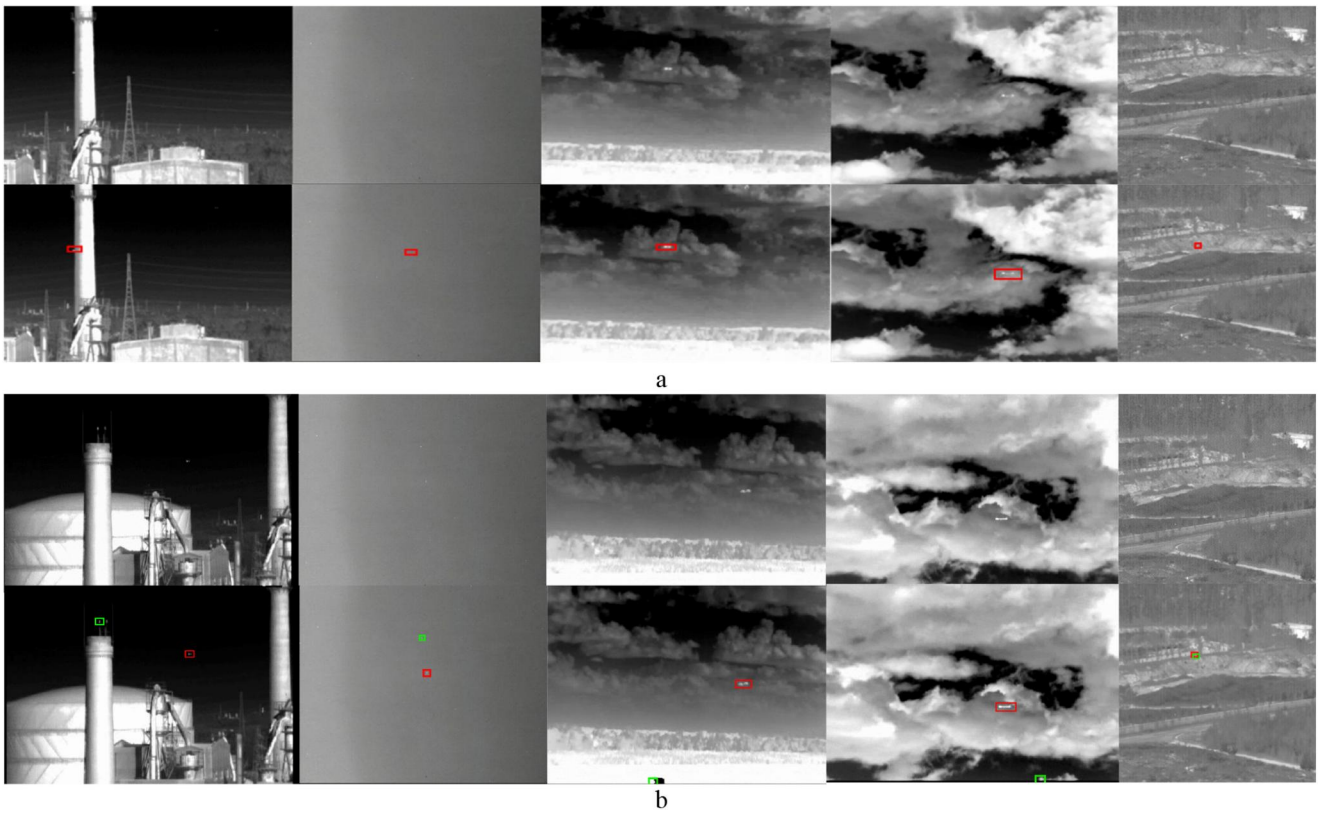

FIGURE 6 Poor results of IC-SDDNet. (a) The first line shows missed detection results, and the second line shows the actual target positions with red bounding boxes. (b) The first line shows the original images. In the second line, the red bounding boxes indicate correct detections, and the green boxes indicate falsely detected background clutter.

rate is due to the high similarity between the radiation characteristics of the background clutter and those of small and dim targets; the missed detection is due to the overly low contrast between the dim target and the background. In the future, we will try to improve the network structure so that the network can capture stronger target features. Additionally, expanding the training set to enhance the generalisation ability of the network is necessary.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interests regarding the publication of this paper.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study.

ORCID

Wang Xuhao https://orcid.org/0000-0001-5340-1084

REFERENCES

1. Wang, X., et al.: Infrared dim and small target detection based on stable multisubspace learning in heterogeneous scene. IEEE Trans. Geosci. Rem. Sens. 55(10), 5481–5493 (2017). https://doi.org/10.1109/TGRS.2017.2709250
2. Xi, J.: Rapid Detection and Processing of Dim and Small Moving Objects. Doctor dissertation. Univ. Chin. Acad. Sci., Huairou, China (2017)
3. Gao, C., et al.: Infrared patch-image model for small target detection in a single image. IEEE Trans. Image Process. 22(12), 4996–5009 (2013). https://doi.org/10.1109/TIP.2013.2281420
4. Wang, P., Tian, J.W., Gao, C.Q.: Infrared small target detection using directional highpass filters based on LS-SVM. Electron. Lett. 45(3), 156–158 (2009). https://doi.org/10.1049/el:20092206
5. Bai, X., Zhou, F.: Analysis of new top-hat transformation and the application for infrared dim small target detection. Pattern Recogn. 43(6), 2145–2156 (2010). https://doi.org/10.1016/j.patcog.2009.12.023
6. Shao, Z., Zhu, X., Liu, J.: Morphology infrared image target detection algorithm optimized by genetic theory. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 37, 1299–1304 (2020)
7. Bai, X., Zhou, F.: Infrared small target enhancement and detection based on modified top-hat transformations. Comput. Electr. Eng. 36(6), 1193–1201 (2010). https://doi.org/10.1016/j.compeleceng.2010.05.008
8. Bai, X., Zhou, F.: Hit-or-miss transform based infrared dim small target enhancement. Opt Laser. Technol. 43(7), 1084–1090 (2011). https://doi.org/10.1016/j.optlastec.2011.02.003
9. Bai, X., Bi, Y.: Derivative entropy-based contrast measure for infrared small-target detection. IEEE Trans. Geosci. Rem. Sens. 56(4), 2452–2466 (2018). https://doi.org/10.1109/TGRS.2017.2781143
10. Han, J., et al.: An infrared small target detecting algorithm based on human visual system. Geosci. Rem. Sens. Lett. IEEE 13(3), 452–456 (2016). https://doi.org/10.1109/LGRS.2016.2519144
11. Liu, J., et al.: Tiny and dim infrared target detection based on weighted local contrast. Geosci. Rem. Sens. Lett. IEEE 15(11), 1780–1784 (2018). https://doi.org/10.1109/LGRS.2018.2856762
12. Han, J., et al.: A robust infrared small target detection algorithm based on human visual system. Geosci. Rem. Sens. Lett. IEEE 11(12), 2168–2172 (2014). https://doi.org/10.1109/LGRS.2014.2323236

24682322, 2023, 4, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.12165 by Test, Wiley Online Library on [22/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

13. Wei, Y., You, X., Li, H.: Multiscale patch-based contrast measure for small infrared target detection. Pattern Recogn. 58, 216–226 (2016). https://doi.org/10.1016/j.patcog.2016.04.002
14. Deng, H., et al.: Infrared small-target detection using multiscale gray difference weighted image entropy. IEEE Trans. Aero. Electron. Syst. 52(1), 60–72 (2016). https://doi.org/10.1109/TAES.2015.140878
15. Deng, H., et al.: Small infrared target detection based on weighted local difference measure. IEEE Trans. Geosci. Rem. Sens. 54(7), 4204–4214 (2016). https://doi.org/10.1109/TGRS.2016.2538295
16. Cui, Z., et al.: Target detection algorithm based on two layers human visual system. Algorithms 8(3), 541–551 (2015). https://doi.org/10.3390/a8030541
17. Wang, H., Peng, J., Yue, S.: A directionally selective small target motion detecting visual neural network in cluttered backgrounds. IEEE Trans. Cybern. 50(4), 1541–1555 (2018). https://doi.org/10.1109/TCYB.2018.2869384
18. Wang, H., et al.: A robust visual system for small target motion detection against cluttered moving backgrounds. IEEE Transact. Neural Networks Learn. Syst. 31(3), 839–853 (2019). https://doi.org/10.1109/TNNLS.2019.2910418
19. Dai, Y., et al.: Asymmetric contextual modulation for infrared small target detection. In: 2021 IEEE Winter Conf. Appl. Comput. Vision (WACV), Waikoloa, HI, pp. 949–958 (2021)
20. Wang, H., Zhou, L., Wang, L.: Miss detection vs. false alarm: adversarial learning for small object segmentation in infrared images. In: 2019 IEEE/CVF Int. Conf. Comput. Vision (ICCV), Seoul, Korea, pp. 8508–8517 (2019)
21. Fang, H., et al.: Infrared small UAV target detection based on residual image prediction via global and local dilated residual networks. Geosci. Rem. Sens. Lett. IEEE 19, 1–5 (2022). https://doi.org/10.1109/LGRS.2021.3085495
22. Huang, L., et al.: Infrared small target segmentation with multiscale feature representation. Infrared Phys. Technol. 116, 103755 (2021). https://doi.org/10.1016/j.infrared.2021.103755
23. Li, B., et al.: Dense nested attention network for infrared small target detection. arXiv:2106.00487 (2021)
24. Zhang, T., et al.: AGPCNet: attention-guided pyramid context networks for infrared small target detection. arXiv:2111.03580 (2021)
25. Bai, Y., et al.: Cross-connected bidirectional pyramid network for infrared small-dim target detection. Geosci. Rem. Sens. Lett. IEEE 19, 1–5 (2022). https://doi.org/10.1109/LGRS.2022.3145577
26. Chen, F., et al.: Local patch network with global attention for infrared small target detection. IEEE Trans. Aero. Electron. Syst. 58(5), 3979–3991 (2022). https://doi.org/10.1109/TAES.2022.3159308
27. Wang, K., et al.: Interior attention-aware network for infrared small target detection. IEEE Trans. Geosci. Rem. Sens. 60, 1–13 (2022). https://doi.org/10.1109/TGRS.2022.3163410
28. Yu, C., et al.: Infrared small target detection based on multiscale local contrast learning networks. Infrared Phys. Technol. 123, 104107 (2022). https://doi.org/10.1016/j.infrared.2022.104107
29. Li, M., et al.: Moving weak point target detection and estimation with three-dimensional double directional filter in IR cluttered background. Opt. Eng. 44(10), 107007 (2005). https://doi.org/10.1117/1.2056586
30. Gao, C., et al.: Infrared small-dim target detection based on Markov random field guided noise modeling. Pattern Recogn. 76, 463–475 (2018). https://doi.org/10.1016/j.patcog.2017.11.016
31. Deshpande, S.D., et al.: Max-mean and max-median filters for detection of small targets. Signal and Data Processing of Small Targets 1999 3809 (1999). SPIE
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., et al. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241. Springer International Publishing, Cham (2015)
33. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39(12), 2481–2495 (2017). https://doi.org/10.1109/TPAMI.2016.2644615
34. Zheng, X., et al.: Unsupervised change detection by cross-resolution difference learning. IEEE Trans. Geosci. Rem. Sens. 60, 1–16 (2022). Art no. 5606616. https://doi.org/10.1109/TGRS.2021.3079907
35. He, K., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conf. Comput. Vision Pattern Recognit. (CVPR), Las Vegas, NV, pp. 770–778 (2016)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR (2014). abs/1409.1556
37. Zheng, X., et al.: Generalized scene classification from small-scale datasets with multitask learning. IEEE Trans. Geosci. Rem. Sens. 60, 1–11 (2022). Art no. 5609311. https://doi.org/10.1109/TGRS.2021.3116147
38. Grana, C., Borghesani, D., Cucchiara, R.: Optimized block-based connected components labeling with decision trees. IEEE Trans. Image Process. 19(6), 1596–1609 (2010). https://doi.org/10.1109/TIP.2010.2044963
39. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: Int. Conf. Learn. Representations, San Diego, CA, pp. 1–15 (2014)
40. Bagheri, Z., et al.: Performance of an insect-inspired target tracker in natural conditions. Bioinspiration Biomimetics 12(2), 025006 (2017). https://doi.org/10.1088/1748-3190/aa5b48
41. Chenqiang, G., Tianqi, Z., Qiang, L.: Small infrared target detection using sparse ring representation. IEEE Aero. Electron. Syst. Mag. 27(3), 21–30 (2012). https://doi.org/10.1109/MAES.2012.6196254
42. Qin, Y., et al.: Infrared small target detection based on facet kernel and random walker. IEEE Trans. Geosci. Rem. Sens. 57(9), 7104–7118 (2019). https://doi.org/10.1109/TGRS.2019.2911513
43. Chen, C.L.P., et al.: A local contrast method for small infrared target detection. IEEE Trans. Geosci. Rem. Sens. 52(1), 574–581 (2014). https://doi.org/10.1109/TGRS.2013.2242477

How to cite this article: Long, M., et al.: SDDNet: Infrared small and dim target detection network. CAAI Trans. Intell. Technol. 8(4), 1226–1236 (2023). https://doi.org/10.1049/cit2.12165

24682322, 2023, 4, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.12165 by Test, Wiley Online Library on [22/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License