STDMANet: Spatio-Temporal Differential Multiscale Attention Network for Small Moving Infrared Target Detection

Puti Yan, Runze Hou, Xuguang Duan, Chengfei Yue, *Member, IEEE*, Xin Wang, *Member, IEEE*, and Xibin Cao

***Abstract*— Infrared target detection has important applications in rescue and Earth observation. However, the disadvantages of low signal-to-clutter ratios and severe background noise interference for infrared imaging pose great challenges to the detection technology for infrared dim targets. Most algorithms only extract features from the spatial domain, while the lack of temporal information results in unsatisfactory detection performance when the difference between the target and the background is not significant enough. Although some methods utilize temporal information in the detection process, these nonlearning-based methods fail to incorporate the complex and changeable background, and need to adjust parameters according to the input. To tackle this problem, we proposed a Spatio-Temporal Differential Multiscale Attention Network (STDMANet), a learning-based method for multiframe infrared small target detection in this article. Our STDMANet first used the temporal multiscale feature extractor to obtain spatiotemporal (ST) features from multiple time scales and then resorted them to the spatial multiscale feature refiner to enhance the semantics of ST features on the premise of maintaining the position information of small targets. Finally, unlike other learning-based networks that require binary masks for training, we designed a mask-weighted heatmap loss to train the network with only center point annotations. At the same time, the proposed loss can balance missing detection and false alarm, so as to achieve a good balance between finding the targets and suppressing the background. Extensive quantitative experiments on public datasets validated that the proposed STDMANet could improve the metric F_1 score up to 0.9744, surpassing the state-of-the-art baseline by 0.1682. Qualitative experiments show the proposed method could stably extract foreground moving targets from video sequences with various backgrounds while reducing false alarm rate better than other recent baseline methods.**

***Index Terms*— Attention mechanism, infrared target, target detection.**

Manuscript received 13 September 2022; revised 29 November 2022 and 13 January 2023; accepted 18 January 2023. Date of publication 1 February 2023; date of current version 9 February 2023. This work was supported in part by the Science Center Program of National Natural Science Foundation of China under Grant 62188101; in part by the National Natural Science Foundation of China under Grant 61833009, Grant 11972130, Grant 61690212, Grant 62222209, and Grant 62102222; and in part by the Heilongjiang Touyan Team. *(Corresponding authors: Chengfei Yue; Xin Wang.)*

Puti Yan and Xibin Cao are with the Department of Aerospace Engineering, Harbin Institute of Technology, Harbin 150001, China.

Runze Hou is with the Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Beijing 518055, China.

Xuguang Duan and Xin Wang are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: xin_wang@tsinghua.edu.cn).

Chengfei Yue is with the Institute of Space Science and Applied Technology (ISSAT), Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China (e-mail: yuechengfei@hit.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3241311

NOMENCLATURE

STDMANet	Spatio-Temporal Differential Multiscale Attention Network.
HVS	Human visual system.
LoG	Laplacian of Gaussian.
LIG	Local intensity and gradient.
AAGD	Absolute gray-scale difference.
LoPSF	Laplacian of point spread function.
ACM	Asymmetric contextual modulation.
ALCNet	Attentional local contrast network.
DNANet	Dense nested attention network.
DNIM	Dense nested interactive module.
NPSTT	Nonoverlapping patch Spatio-Temporal tensor.
TCNN	Tensor capped nuclear norm.
CSAM	Channel and spatial attention module.
CTSAM	Channel-temporal and spatial attention module.
CTA	Channel-temporal attention.
MLP	Multilayer perceptron.
TP	True positive.
FN	False negative.
FP	False positive.
BSF	Background suppression factor.
CG	Contrast gain.

I. INTRODUCTION

WITH the development of science and technology, infrared imaging technology is widely used in ground observation, night rescue, forest firefighting, and other fields due to its advantages of strong environmental adaptability and high resolution [1], [2], [3], [4]. Among its various applications, infrared target detection technology, as an extremely important part of the infrared imaging system, has been studied by scientists and scholars [5] for decades. However, due to the lack of effective information, such as the shape and texture of infrared dim targets, the fast motion of the target, and the influence of background noise and interference radiation [6], the infrared image usually contains a severe undulating background, and the targets are submerged. Thus, the targets often appear to be several pixels in the complex background, which makes infrared dim target detection very difficult.

Today, the mainstream methods of air-to-ground infrared target detection are mainly divided into two categories: single-frame-based and multiframe-based methods [7]. The single-frame infrared target detection method mainly distinguishes the target from the background by means of background features and local contrast [8]. However, due to the characteristics of the infrared dim targets, there are limitations.

1) The signal-to-clutter ratio (SCR) is low.
2) The size is small, and the relative area is less than 0.01% of the entire image.
3) The background is extremely complex.
4) The relative target moves fast.

Therefore, a single-frame infrared image cannot provide enough information for stable weak and small target detection.

However, the multiframe method detects more temporal context information than the single-frame method [9] by correlating multiple consecutive frames to improve the performance of small target detection, and the method based on multiple frames usually performs better than that based on a single frame. These temporal clues are important in robust small infrared object detection, especially for complex cases [10].

There are two main methods of multiframe data target detection through temporal context: one is to reduce noise and interference through temporal context, thereby decreasing false detection due to flicker and other factors [11]; the other is to enhance the extraction of target characteristics through temporal context to better solve the difficulties caused by motion blur and small object area [12].

There is a certain difference between a large number of small aerial target image sequences captured by an airborne imaging platform and ordinary infrared images. In air-to-ground data, the target moves relatively faster, the target SCR is low, and ground radiation and weather interference are very serious considerations. Dim infrared aerial targets often appear as point targets in infrared images. They only occupy a few pixels in the image. Due to infrared imaging characteristics, the edges of objects are blurred. The scarcity of intrinsic properties for the target means that the shape of the target is of little significance for practical applications. We can simply use the center position of the spot to represent the small infrared aerial target in the image.

The targets in a cluttered background may not show a high correlation in an image sequence, especially if the background changes due to the motion of the airborne infrared detection system. In this case, traditional methods such as background alignment alone cannot achieve satisfactory results for robust small infrared aerial target detection and will introduce other redundant information and noise. Thus, even if it can provide some implicit information about motion, some additional feature processing is required.

In this article, the temporal attention mechanism was introduced to infrared dim target detection for the first time. The proposed method contained a temporal multiscale feature extractor module, a spatial multiscale feature refiner module, and a weighted loss compensation module. The temporal multiscale feature extractor module extracted both dynamic motion and static contrast features through several feature paths in different time scales. In the spatial multiscale feature refiner module, the semantics of these extracted Spatio-Temporal (ST) features were further refined to better distinguish targets from a variety of backgrounds while maintaining position information through dense connection. For the temporal multiscale module and spatial multiscale module, we customized different attention mechanisms to complete their respective tasks. Different attention mechanisms were designed for the selection of different steps for the temporal section in particular, the continuous motion accumulation in the temporal domain, and the selection of local contrast enhancement in the spatial domain, so as to improve the quality of feature extraction. In addition, our mask-weighted heatmap loss could perform end-to-end training with only point annotations, eliminating the need for segmentation masks. We evaluated the proposed method on public datasets (DSAT [13] and SIATD [14]) and compared it with existing methods. Experimental results demonstrated that our method could robustly detect small infrared airborne targets against complex ground backgrounds and outperform existing methods.

II. Related Work

A. Single-Frame Infrared Small Target Detection

The mainstream single-frame detection method of infrared dim target highlights small targets, suppresses background noise through image preprocessing, then uses threshold segmentation to extract suspected targets, and, finally, confirms targets based on feature information, without considering time series operations. This has been extensively studied.

In 2012, according to the contrast mechanism of HVS, Shao et al. [15] suppressed noise and enhanced target intensity through the LoG filter, thereby enhancing monitoring performance. In 2012, Qi et al. [16] proposed a salient region detection method combined with an attention mechanism to detect small infrared targets in complex backgrounds. In 2013, Gao et al. [17] used an adaptive infrared patch image model constructed by local patches to segment targets and suppress a different clutter interference. Also, in 2013, Chen et al. [18] proposed an algorithm based on the contrast mechanism of the HVS and a derived kernel model to segment targets through local contrast and adaptive thresholds. In 2014, Han et al. [19] proposed a threshold operation and fast traversal mechanism method based on an HVS attention transfer mechanism for rapid target acquisition. Although these methods can effectively improve detection performance, they are still not applicable to complex backgrounds. In 2018, Zhang et al. [20] computed LIG maps from raw infrared images to enhance targets and suppress clutter. In 2018, Moradia et al. [11] modeled point targets through multiscale mean AAGD and LoPSF to reduce the FP rate (FPR).

Although the research on single-frame infrared small target detection has been going on for decades, traditional methods still cannot cope with changeable scenes in complex backgrounds. These methods usually adjust the parameters of each scene, which challenges their application for the actual scene. Recently, deep learning methods have been widely used in various visual tasks. With the release of infrared target detection datasets [21], [22], deep neural networks have

naturally migrated to the field of single-frame infrared small target detection. The annotations provided by these datasets are usually bounding boxes and segmentation masks, among which segmentation masks are the most widely used.

For those methods utilizing bounding box annotations, they detect small infrared targets by modifying the general detection network [23]. In contrast, most methods obtained the contour of the target by predicting the segmentation mask of the input image and then determining the detection boxes. Zhao et al. [24] proposed TBC-Net, which utilized a simple spatial multiscale network to extract the location of the target and a simple classification network to obtain the label of the target. Wang et al. [22] regarded infrared small target detection as a balance between miss detection and false alarm. They achieved a balance between the two indicators through a deep adversarial learning framework [22]. Dai et al. [25] proposed ACM, which combined high-level semantics and low-level location details by designing a comprehensive top-down and bottom-up attention modulation path. They further introduced the idea of local contrast in the subsequent ALCNets to refine the characteristics of small targets [26]. In addition, the interior attention-aware network (IAANet) combined the relationship between pixels to enhance the correlation of the target through coarse-to-fine attention [27]. DNANet proposed a DNIM module, which improved the effect of feature extraction and preservation of targets through dense connection and spatial pyramid fusion [21]. Liu et al. [28] incorporated the popular structure Transformers used in computer vision and natural language processing into infrared small target detection and achieved good results.

Deep-learning-based single-frame infrared small target detection methods are inspired by both deep neural network design (such as dense connection, attention mechanism, and feature pyramid networks) and traditional infrared small target detection methods (such as local contrast) so that the performance of single-frame detection methods has significantly improved.

B. Multiframe Infrared Small Target Detection

In contrast, multiframe methods need to deal with several consecutive frames. Therefore, multiframe methods must be able to conduct ST processing in order to obtain dynamic information, such as motion and brightness changes, in addition to static information, such as local contrast.

In 2011, Qi and An [29] proposed an improved optical flow infrared target detection algorithm to improve adaptive ability and prevent the suppression of reliable optical flow. In 2019, Lv et al. [30] proposed a method of building a background dictionary based on an online learning double sparse model to improve the traditional algorithm to suppress background noise. Zhao et al. [31] proposed an ST consistency detection method for motion trajectories based on optical flow to distinguish objects from backgrounds in 2019. In 2020, Li et al. [32] proposed an adaptive infrared dark target detection method utilizing a multiframe screening strategy based on optical flow combined with dynamic pipeline filtering to identify targets and reduce the false alarm rate. In 2021, Wang et al. [33] obtained nonoverlapping patches in adjacent images by using sliding windows, established an NPSTT model, and introduced TCNN to improve the robustness of detection.

However, the problem of applying deep learning technology to small infrared target detection in videos has not been fully explored. The release of relevant datasets [13], [14] recently provides a good premise for this problem. The simplest method is to detect small targets in every video frame [34], which cannot consider temporal context information. Some methods carried out the super-resolution operation on the current frame through the dynamic information of the front and back frames to enhance the details and significance of the targets in the current frame [35], [36]. This operation also enabled the application of single-frame small target detection methods. However, such methods required multiplying the image resolution, significantly increasing computation consumption. Recently, Yao et al. [37] made a multiframe image input into a single-frame detection method through maximum filter preprocessing.

The idea of these methods is to use a single-frame detector to detect multiframe input. The core problem in this process is to convert multiple-frame images to a single frame and retain dynamic information as much as possible. However, these methods are not specially designed for multiframe scenes, and there is still much space for exploration in the use of temporal information. Recently, some methods [38], [39] have explored the specific designs for multiframe cases, but there is still a problem on the edge of small targets in their requirements for detection boxes, which increases the difficulty of annotating. Their exploration of spatial multiscale information is not sufficient, which has been validated in recent single-frame detection methods that spatial multiscale information is effective for small targets. Based on this, this article made a detailed analysis of the ST characteristics in multiscale manners of multiframe input and designs in STDMANet, which achieved the efficient utilization of ST information under a similar network complexity to that of the single-frame detection method.

III. Proposed Method

A. Overview

The overall structure of our proposed STDMANet is shown in Fig. 1. STDMANet takes k infrared frames as input. With the sequential processing of the temporal multiscale feature extractor (see Section III-B), spatial multiscale feature refiner (see Section III-C), and prediction head module (see Section III-D), the output of our model is the prediction mask p_t. The generated segmentation mask is processed differently in the training phase and the testing phase. In the training phase, we used mask-weighted heatmap loss (see Section III-E) to minimize the difference between the segmentation mask and the 2-D Gaussian heatmap. In the test phase, we used instance extraction and center computation (see Section III-F) to obtain the position of small targets in the segmentation mask.

Specifically, the given input $X_t \in \mathbb{R}^{k\times H\times W}$ at time step t consists of k infrared frames, i.e., $X_t = [x_{t-k}, \ldots, x_{t-1}, x_t]$,

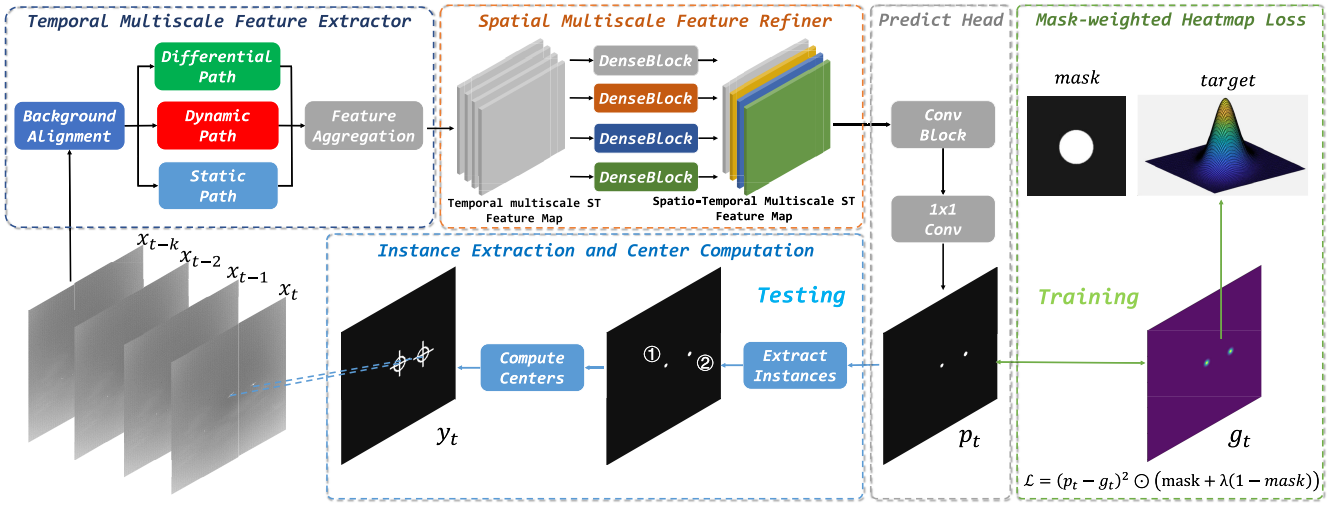

Fig. 1. Structure diagram of our proposed STDMANet. The network consists of three parts: the temporal multiscale feature extractor, the spatial multiscale feature refiner, and the prediction head. The temporal multiscale feature extractor is proposed to extract ST features in different time scales. Then, the spatial multiscale feature refiner can refine and preserve the ST features on the multispatial scale. Finally, the prediction head is applied to produce the prediction map. Then, different operations are applied in the training and testing phase to train the whole network and generate the position output, respectively.

where x_t is corresponded to the infrared frame at time step t. Here, we treated infrared images as single-channel gray-scale images rather than RGB images so that the multiple frames could be treated as several input channels of the network. After passing through the temporal multiscale feature extractor module and the spatial multiscale feature refiner module, the output feature maps are $F_t^T \in \mathbb{R}^{C_T \times H \times W}$ and $F_t^{\text{ST}} \in \mathbb{R}^{C_{\text{ST}} \times H \times W}$, respectively. Then, the final segmentation mask $p_t \in \mathbb{R}^{H \times W}$ generated by the prediction head module represents the probability that each pixel becomes the center of a small target. In the following, we would explain in detail the motivation and design of each module and describe the training and testing process for the model to clarify the reason that our model could solve the problem of point target position prediction in multiframe infrared images.

B. Temporal Multiscale Feature Extractor

The multiple infrared frames can be regarded as a multichannel single image input into the network because the infrared image can be regarded as a single-channel gray-scale image. Therefore, the single-frame detection network can be applied to multiframe scenarios by changing only the first layer of the network and will not introduce too much computational burden.

However, there seems to be some impropriety in inputting continuous infrared sequences into the network equally. On the one hand, the purpose of multiframe infrared target detection is to locate the target position of the current frame according to the continuous k images in the past, so the current frame should clearly have a larger weight than other frames. On the other hand, when the motion state of the target is not fixed, the frames that produce more apparent displacement should also have a larger weight.

Based on the above analysis and the traditional methods of infrared small target detection, we designed a temporal multiscale feature extractor module, which was composed of the motion path, the dynamic path, and the static path. These paths could extract features on different time scales by customizing the input information of different paths. Specifically, the input of the static path is the current frame, corresponding to the static time scale. The input of the dynamic path is all k frames, corresponding to the dynamic time scale. The input of the differential path is the subtraction between the current frame and each past frame, corresponding to the time scale from 1 to $k-1$. The specific design of these paths can be seen in Fig. 2.

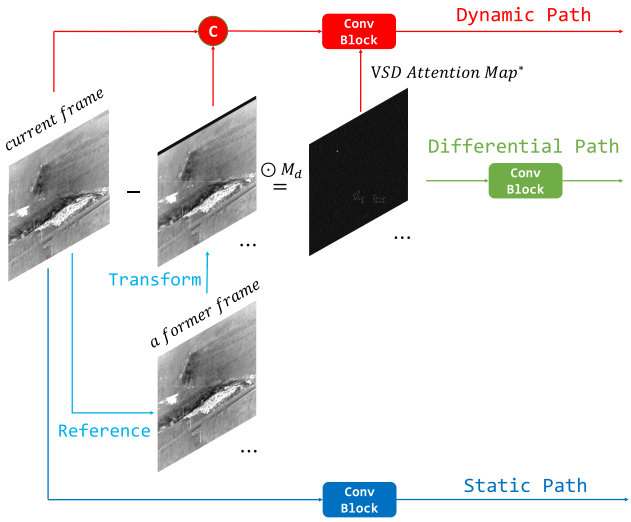

Fig. 2. Illustration of proposed temporal multiscale feature extractor module. The basic idea is to achieve the feature extraction on different time scales by customizing the input image sequences of different paths and then through aggregation to get the ST features containing multiscale time information.

Before the infrared sequence fed into the network, we perform background alignment to separate the motion of the background and the motion of the target, which made the differential path have more physical meaning, which was

the change in pixel brightness in the same position. Specifically, we used the common background alignment pipeline. For a previous frame $x_{t-i}, i = 1, \ldots, k$ and the current frame x_t, we extracted their scale-invariant feature transform (SIFT) [40] features, respectively. Then, we identified the matching pairs between the two. After this, the transformation matrix is calculated through the RANdom SAmple Consensus (RANSAC) [41] algorithm, and finally, x_{t-i} is transformed to the perspective of x_t, which we denote it as $\tilde{x}_{t-i}$.

After the above alignment operation, we could formalize the input of different paths. The input of differential path $\mathrm{DI}_t \in \mathbb{R}^{(k-1)\times H\times W}$ can be formulated as

$$\mathrm{DI}_t = M_t^d \odot [x_t - \tilde{x}_{t-k}, \ldots, x_t - \tilde{x}_{t-1}] \tag{1}$$

where $M_t^d = [m_{t-k}^d, \ldots, m_{t-1}^d]$ is a group of masks that excludes the blank area after transformation. m_{t-i}^d can be obtained from

$$m_{t-i}^d(m,n) = \begin{cases} 0, & \tilde{x}_{t-i}(m,n) = 0 \\ 1, & \text{else} \end{cases} \tag{2}$$

where $m \in [0, W-1]$ and $n \in [0, H-1]$ is the position index of the image. Because there is a background shift between the previous frame x_{t-i} and the current frame x_t, there will be a certain blank area in $\tilde{x}_{t-i}$ when the previous frame is transformed to the perspective of the current frame. If these blank spaces are ignored, there will be a large response here when taking the differential operation because it is zero here in $\tilde{x}_{t-i}$ and nonzero in x_t. Therefore, we introduced additional m_{t-i}^d to exclude this area to prevent the model from focusing on the new background introduced by the camera motion rather than the moving targets. The design of the differential path was inspired by the frame differential methods commonly used in infrared small target detection. The differential results of different time scales were obtained by using the current frame to apply differential operations to each previous frame. Thus, the model was provided with brightness-aware information across time scales.

In addition, the input of a dynamic path $\mathrm{DY}_t \in \mathbb{R}^{k\times H\times W}$ can be formalized as

$$\mathrm{DY}_t = \mathrm{abs}([\mathrm{DI}_t, I]) \odot [\tilde{x}_{t-k}, \ldots, x_t] \tag{3}$$

where I is the identity matrix.

We padded DI_t here to k channels to serve as the variable step differential attention (VSDA) map for clearer highlighting of the areas in the image sequence that had changed significantly. The dynamic path was designed to capture continuous motion and changes in image sequences.

Finally, the input of the static path is x_t. The static path is designed to obtain spatial information, such as local contrast and background information.

By passing the customized inputs through their respective convolution blocks, the output features F_t^{DI} of the differential path, F_t^{DY} of the dynamic path, and F_t^{SP} of the static path could be obtained. Then, these features were concatenated, and the output of the temporal multiscale feature extractor module is obtained through a feature aggregator, which was also a convolution block that reduced the dimension of the collected feature map. Formally,

$$F_t^{\mathrm{DI}} = \mathrm{ConvBlock}(\mathrm{DI}_t) \tag{4}$$

$$F_t^{\mathrm{DY}} = \mathrm{ConvBlock}(\mathrm{DY}_t) \tag{5}$$

$$F_t^{\mathrm{SP}} = \mathrm{ConvBlock}(x_t) \tag{6}$$

$$F_t^{T} = \mathrm{ConvBlock}\left(\left[F_t^{\mathrm{DI}}, F_t^{\mathrm{DY}}, F_t^{\mathrm{SP}}\right]\right). \tag{7}$$

For the design of the convolution module, we followed DNANet [21] to perform feature enhancement adaptively through sequentially applied channelwise and spatialwise attention. In the original paper, this module was called CSAM. However, here, we took an infrared frame as a gray-scale image to serve as an input channel for the network. Thus, the channel dimension of the feature map in our model was not only the semantic channels but contained temporal information and clues. As such, we clarified that, in our model, the basic convolution module should be called the CTSAM. The core of the module was the acquisition of CTA map $M_{\mathrm{CT}} \in \mathbb{R}^{C_{\mathrm{in}}\times 1\times 1}$ and spatial attention (SA) map $M_S \in \mathbb{R}^{1\times H\times W}$, where C_{in} is the input feature channels. For input feature map H_{in}, the CTA map M_{CT} can be calculated by

$$H = \mathrm{Conv2}(\mathrm{Conv1}(H_{\mathrm{in}})) \tag{8}$$

$$M_{\mathrm{CT}} = \sigma([W(\mathrm{Maxpool}_{\mathrm{HW}}(H)), W(\mathrm{Avgpool}_{\mathrm{HW}}(H))]) \tag{9}$$

where Conv1 and Conv2 are the combination of Conv-BN-ReLU. The subscript HW in Maxpool and Avgpool means taking pooling operations in width and height dimensions. W is a two-layer MLP that performs nonlinear feature projection. σ is sigmoid nonlinearity. After this, the feature map becomes

$$H_{\mathrm{CT}} = M_{\mathrm{CT}} \otimes H. \tag{10}$$

Then, the SA map M_S and the output feature map H_{out} can be obtained by

$$M_S = \sigma(\mathrm{Conv}[\mathrm{Maxpool}_{\mathrm{CT}}(H_{\mathrm{CT}}), \mathrm{Avgpool}_{\mathrm{CT}}(H_{\mathrm{CT}})]) \tag{11}$$

$$H_{\mathrm{out}} = M_S \otimes H_{\mathrm{CT}}. \tag{12}$$

In the following spatial multiscale feature refiner module, we continued using the standard structure of CTSAM. However, it should be noted that, for the static path, temporal information does not exist, so the CTSAM module here is simplified to a CSAM module.

C. Spatial Multiscale Feature Refiner

Although the temporal multiscale feature contains rich information, such as motion, brightness change, local contrast, and background in the input infrared sequence, its semantic information is not rich enough to clearly distinguish the target from the background. Therefore, it is necessary to extract high-level semantic information, which can be achieved by stacking subnetworks and spatial multiscale fusion in computer vision tasks. However, the cascaded subnetworks will still experience suboptimal small target detection. Because the target size is so small, targets only account for a small part of the feature map. In the sequential processing of subsequent subnetworks, the features of a small target are easy to obscure and weaken. In order to solve this problem, the dense connection is

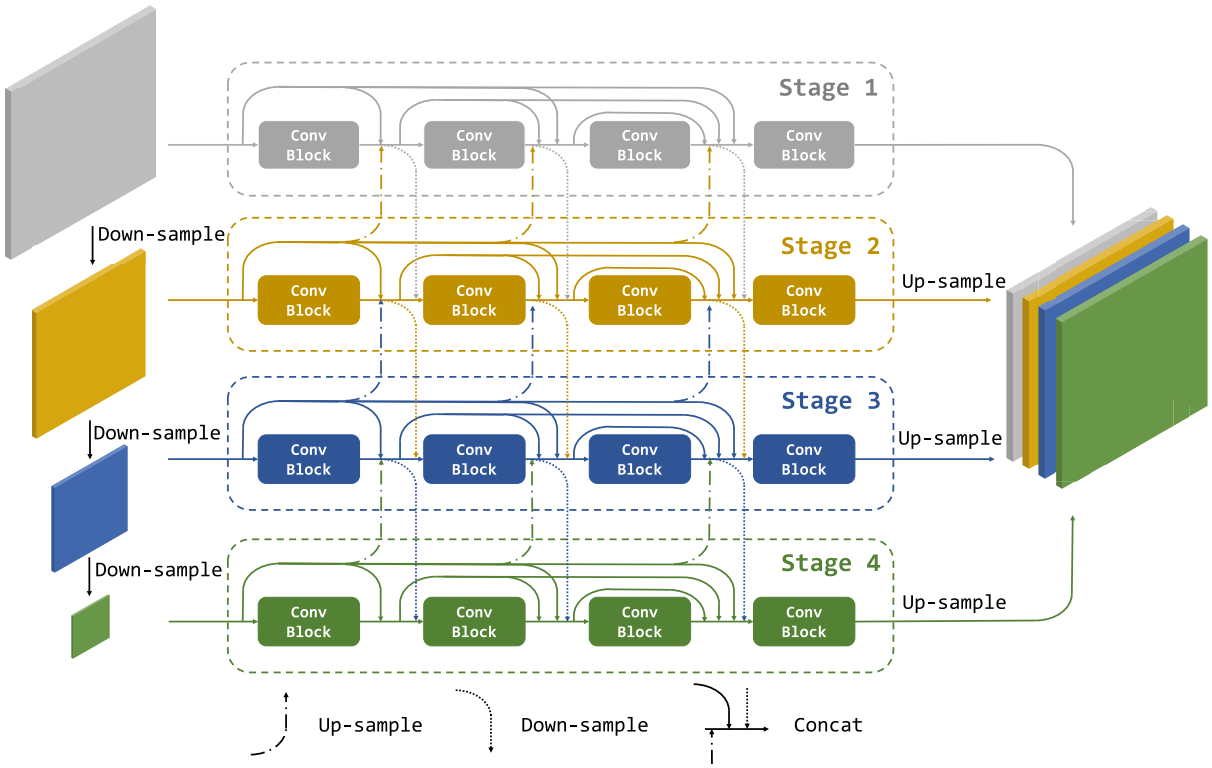

Fig. 3. Illustration of the proposed spatial multiscale feature refiner module. This module divides four different spatial scales and refines the semantic features of the target at each scale. We design sufficient spatial scale interaction paths between different spatial scales to integrate different scale features and enrich semantic information. Finally, the module outputs ST multiscale features based on the input temporal multiscale information.

an essential component of neural network design in small target detection [21], [36]. Based on the above principles, we designed the spatial multiscale feature refiner module, as shown in Fig. 3.

In this article, we set four spatial scales: the original scale, 1/2 scale, 1/4 scale, and 1/8 scale, corresponding to the four stages of the module. The reason for setting multiple scales was first to enrich semantic information through the interaction between multiscales and second to provide different feature paths for small targets of different sizes because, in the detection task, feature maps of different sizes could be used to detect targets of different sizes. We did not further expand the fifth spatial scale, and it was almost impossible for extremely small feature maps to retain information about small targets. For each spatial scale, we stacked L CTSAM convolution modules. In addition, with the combination of dense connection and spatial fusion, each convolution module could reuse all previous features of the current scale and receive the features from the upper scale and the bottom scale.

Specifically, we first downsampled the temporal multiscale features to obtain the original feature inputs $G_t^{s,0}$ at different spatial scales and stages, i.e.,

$$G_t^{s,0} = \text{DownSample}_{\text{HW}}^{2^s \times 2^s}\left(F_t^T\right) \tag{13}$$

where s is index of stage and the superscript of DownSample represents the kernel size of downsampling.

After that, the output of the lth module of the sth stage was denoted as $G_t^{s,l}$, which can be calculated by

$$G_t^{s,l} = \text{ConvBlock}\left(\left[G_t^{s,0}, \ldots, G_t^{s,l-1}, \check{G}_t^{s-1,l-1}.\hat{G}_t^{s+1,l-1}\right]\right) \tag{14}$$

where $\check{G}_t^{s-1,l-1}$ is 2×2 downsampling from $G_t^{s-1,l-1}$ and $\hat{G}_t^{s+1,l-1}$ is 2×2 upsampling from $G_t^{s+1,l-1}$. For the modules of the first and last stages, their inputs lacked the output of the upper module and the bottom module, respectively.

Finally, the features generated by the modules in different stages were aggregated, and finally, the spatial multiscale features F_t^{ST} were obtained

$$F_t^{\text{ST}} = \left[G_t^{0,L}, u_2\left(G_t^{1,L}\right), u_4\left(G_t^{2,L}\right), u_8\left(G_t^{3,L}\right)\right] \tag{15}$$

where u_2, u_4, and u_8 represent 2×2, 4×4, and 8×8 upsampling, respectively.

D. Prediction Head

The prediction head consists of a CTSAM module and a 1×1 convolution. The function of the convolution module is to further process the previously obtained ST multiscale features and reduce the dimension to produce reasonable prediction results. The 1×1 convolution generates heatmap prediction result $p_t \in \mathbb{R}^{H \times W}$ based on the final ST features, and each pixel in the heatmap represents the probability that the pixel is the center of a small target.

E. Mask-Weighted Heatmap Loss

Most detection methods of infrared small target use datasets annotated with segmentation binary masks as ground truth. However, when the target further shrinks and does not exceed the size of 3×3, it is not easy to discern a clear edge, and even we can only use the central point to annotate the targets. Hence, the segmentation mask is difficult to apply in such a scenario. Therefore, how to use point annotations to train the network in the detection of tiny targets has become a problem that needs to be solved. If we use the binary mask directly, only the center point is 1, and the rest of the background is 0, in which case the model will easily recognize all the pixels as the background and fail to generate the correct prediction.

Using a 2-D Gaussian distribution to model infrared small targets can alleviate the problem that point annotations are challenging for model training [42], [43]. The 2-D Gaussian distribution softens the binary mask into a continuous spatial probability distribution, that is, the probability of the annotated point as the target center point is 1, and the probability decreases with the increasing distance from the annotated center points. Specifically, we can build ground truth g_t in the following ways:

$$g_t(m,n) = \exp\left(-\frac{1}{2\epsilon^2}\left[\left(m-\mu_x^o\right)^2 + \left(n-\mu_y^o\right)^2\right]\right) \quad (16)$$

where ϵ is the standard deviation and o is the index of independent targets ($o = 1, 2, \ldots, O$). (μ_x^o, μ_y^o) is the point annotation of the oth small target. The coefficient and covariance matrix of the distribution were ignored because we did not need to guarantee that the integral of the probability density function was 1, and we assumed that the trend of probability reduction was isotropic.

The ground truth using a 2-D Gaussian distribution expands the area of the nonzero part of the heatmap, preventing the model from cheating by predicting the whole image as the background. However, it should be noted that the area of the small target is very small, so the area that a 2-D Gaussian can significantly cover will not be very large. In order to further alleviate the imbalance between the target region loss and the background region loss, we introduced M_t

$$M_t(m,n) = \begin{cases} 1, & p_t(m,n) > \delta \\ 0, & \text{else} \end{cases} \quad (17)$$

where δ is the threshold of significant probability. To summarize, we used mask-weighted heatmap loss to train our model

$$L = (p_t - g_t)^2 \odot (M_t + \lambda(1 - M_t)) \quad (18)$$

where λ is the weighting factor, which was used to balance the loss of the background region and the target region. Here, we set $\lambda = 0.25$ to reduce the proportion of the loss of the background region in the backward gradient.

F. Instances Extraction and Center Computation

Instance extraction and center computation were operations in the test phase. The function of instance extraction is to determine each independent target from the predictive heatmap [44]. Then, through the center computation, the center of the closed graphic composed of all the target points was obtained as the center point of the prediction. In this way, we received the desired result: the center point of each target in the image.

IV. Experiments

To qualitatively and quantitatively verify the performance of the proposed method, we conducted experiments on the public datasets DSAT (a dataset for infrared image dim-small aircraft target detection and tracking underground/air background) and SIATD (a dataset for small infrared moving target detection under clutter background). We also compared our method with existing representative methods, including single-frame-based methods and multiframe-based methods. We set the parameters of these baseline methods according to the original paper for a fair comparison.

We would investigate the following questions in our experimental evaluation.

1) *Q1:* Faced with the thermal radiation effects of different ground disturbances, such as forests, rivers, and buildings in the background, we would study the impact of different paths on capturing dim targets.
2) *Q2:* Our proposed method can be viewed as a temporal attention mechanism. We would examine the effect of different time windows on detection performance separately.
3) *Q3:* We would study loss weight settings for background and target regions to achieve a balance between suppressing noise (false detection and false alarm rate) and enhancing the target (miss detection).
4) *Q4:* We would analyze the comparison of the proposed method with other state-of-the-art or data-driven methods.

A. Dataset

For the experimental evaluation, we chose DSAT and SIATD datasets, focusing on solving the following challenges: 1) target detection for multiscene images; 2) how to balance detection probability and false alarm rates; and 3) discovery and capture of dim targets in complex environments.

1) DSAT: This dataset mainly solved the problem of infrared sequence image detection and tracking of fixed-wing unmanned aerial vehicle (UAV) targets. For the DSAT dataset, the typical wavelengths range from 3 to 5 μm, and it belongs to mid-wave infrared. 22 typical scenarios were designed, which mainly included a single target underground complex background, target from far to near, target from near to far, target leaving the field of view, target returning to the field of view, and so on. Most scenes in the dataset were ground backgrounds, including forests, rivers, buildings, and so on. Each typical scene corresponded to one data segment, for a total of 22 data segments, 30 tracks, 16 177 frames of images, and 16 944 targets and labels. Each infrared image has a resolution of 256×256 pixels and a bit depth of 8 bits. An example is shown in Fig. 4.

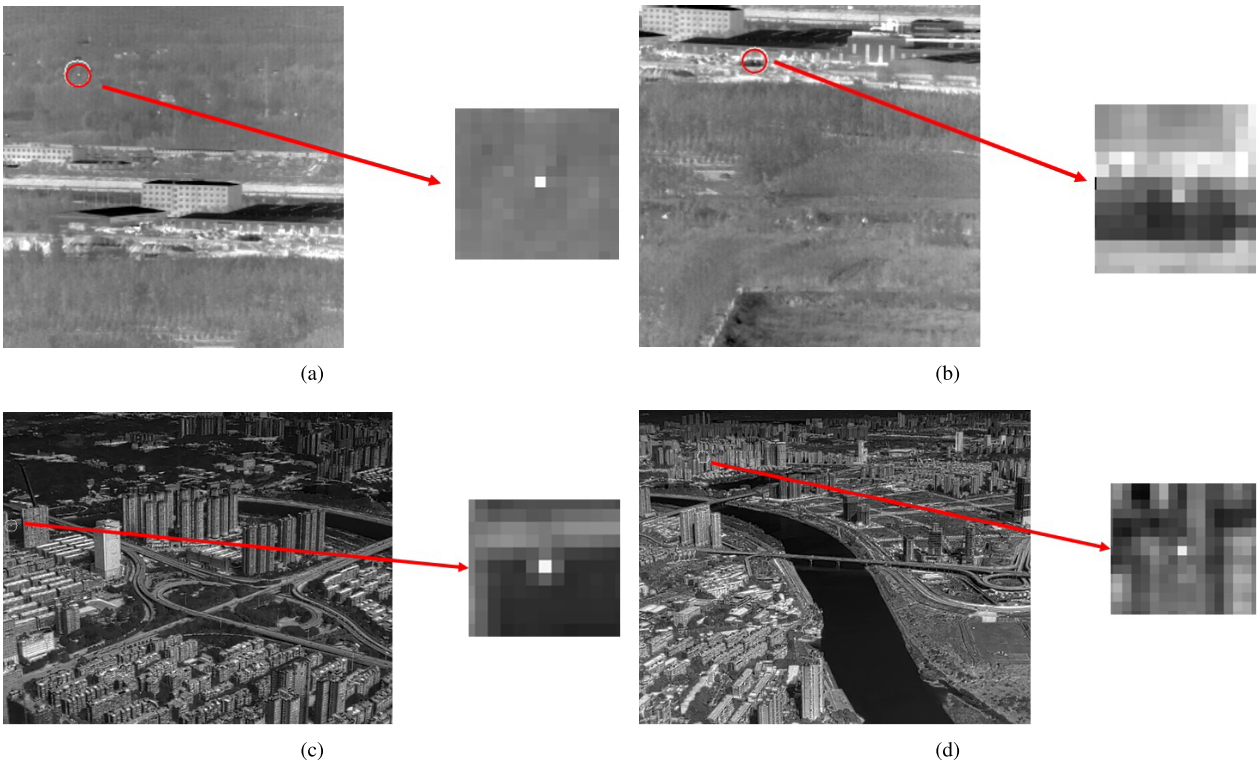

Fig. 4. Two instances in the DSAT and SIATD datasets. We show the easy and difficult parts of the same sequence, respectively. For the easy part, even if the target is small, it still can be clearly distinguished from the background. However, in the difficult part, the target moves to a complex part of the background, which makes it difficult to distinguish the target from the background in appearance, local contrast, and brightness, so it is necessary to use temporal information to determine the target position. (a) Easy part of DAST. (b) Difficult part of DAST. (c) Easy part of SIATD. (d) Difficult part of SIATD.

2) SIATD: This dataset contained a total of 350 image sequences, and the targets were objects such as airplanes flying in the air. Each target category also included five types of backgrounds: looking up, looking down, looking down on vegetation, looking down on water surfaces, and looking down on buildings. For the SIATD dataset, the wavelength ranges from 8 to 14 μm, which belongs to long-wave infrared. Although the dataset used a semisimulation method to generate image sequences, the background images were all captured on-site by a small UAV platform equipped with infrared detection equipment. As a result, the radiation characteristics corresponding to the scene environment in the image were relatively true and reliable. The target size contained in it did not exceed 7×7 at most, which met the standard of dim targets. An example of the dataset is also shown in Fig. 4.

B. Evaluation Metrics

First, to measure the weakness of the target in a complex background and, thus, indirectly evaluate the difficulty of detection, we calculated the SCR of each target in each image

$$\text{SCR} = \frac{|m_t - m_b|}{\sigma_b} \tag{19}$$

where m_t and m_b are the mean of the target and the background area, and σ_b is the variance of the background area. The target area refers to the square with the size of $a \times b$ centered on the target center, while the background area is the square with the same center as $(a + 2d) \times (b + 2d)$. In our experiments, $a = b = 3$ and $d = 10$ by default. The definition comes from [45]. For some baseline methods with different settings, we followed their settings for a fair comparison. The SCR of the DSAT dataset under different splits is shown in Table I, and the SCR of the SIATD dataset in test split is 3.33.

Second, we employed a direct and publicly available metric, the F_1 score that combined the Precision and Recall, to quantitatively evaluate the performance of point target detection. It is defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{20}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{21}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{22}$$

which is the same as the definition in [37].

We stipulated that the TP (i.e., the correct prediction) referred to the distance between the predicted center point and the actual center point being within d pixels, the FN (i.e., the missing prediction) referred to the fact there was no prediction center around an actual center point within d pixels, and the FP (the false detection) assumed that there was no actual center around a predicted center point with d pixels. d was set to 3 in SIATD and 10 in DSAT because there were relatively large targets whose point annotation was not centroid in some sequences of the DSAT dataset. Among the above metrics, Precision is actually the ratio of the number of correctly predicted targets to the total number of predicted targets, while Recall represents the ratio of the number of

correctly predicted targets to the number of correct targets. In this case, these two metrics are directly related to detection probability P_d and false alarm rate F_a, that is,

$$P_d = \text{Recall} \tag{23}$$

$$F_a = 1 - \text{Precision}. \tag{24}$$

The definition comes from [46].

To further evaluate our model in a more stable way, we calculate the receiver operation characteristic (ROC) metric, as described in [47], [48], and [49]. Here, we follow the definition of TP rate (TPR) and FPR in [21], [32], [36], [42], [43], and [45], where

$$\text{TPR} = \frac{\text{number of true detections}}{\text{number of total targets}} \tag{25}$$

$$\text{FPR} = \frac{\text{number of false detections}}{\text{number of image pixels}}. \tag{26}$$

In addition, some DSAT baseline methods did not use the F1 score but other metrics to evaluate the method performance. In these cases, we also calculated the corresponding metrics of our method. These metrics are BSF and CG. They are defined as

$$\text{BSF} = \frac{\sigma_{\text{in}}}{\sigma_{\text{out}}} \tag{27}$$

$$\text{CG}_1 = \frac{|\text{Max}_t - m_b|_{\text{out}}}{|\text{Max}_t - m_b|_{\text{in}}} \tag{28}$$

$$\text{CG}_2 = \frac{|m_t - m_b|_{\text{out}}}{|m_t - m_b|_{\text{in}}} \tag{29}$$

where σ_{in} and σ_{out} are standard deviations of the input image and the output probability map. Max_t represents the max value in the target area, and m_t and m_b are the mean values in the target and the background area. Here, we used two definition of CG according different methods, where CG_1 and CG_2 come from [45] and [50]. The definition of BSF comes from [50].

Finally, we add inference time and frames per second (fps) for evaluating and comparing the speed of our methods and other baselines. In our method, we set the batch size to 1 while calculating inference time.

C. Implementation Details

In our implementation, the sample frames k were set to 20 for the DSAT dataset and 5 for the SIATD dataset. For the DSAT dataset, we fed the infrared frames directly into the network using the original frame size (256×256). For the SIATD dataset, we cropped the original frame (640×512) at 256×256 to save video memory. In addition, the video frames were randomly mirrored and normalized before being sent to the network. The feature dimensions of the differential path, the dynamic path, and the static path were 96, 32, and 32, respectively. The output channel dimensions of feature aggregation and spatial multiscale feature refiner were 32. We initialized the model weights with the Xavier [51] method. The standard deviation ϵ of the 2-D Gaussian distribution was set to 3, and the weight factor λ in the loss was set to 0.25. Our method was developed based on the OpenCV [52] and PyTorch [53] libraries.

During the training process, we followed the official train/test split of the SIATD dataset. For the DSAT dataset, we used the low SCR and challenging sequences as the test set for performance evaluation and ablation experiments. For some of the baseline methods with different splits, we adopted their splits and metrics to facilitate fair comparison. We used the Adam [54] optimizer to train our model with a batch size of 8 and an initial learning rate of 0.005. We used cosine annealing to adjust our learning rate. For the DAST dataset and the SIATD dataset, we trained 200 and 20 epochs, respectively. We applied deep supervision to promote faster and more stable convergence of the network. All models were trained and tested on Nvidia GeForce RTX 3090 GPU.

D. Quantitative Results

1) Performance on the DSAT Dataset: We showed the comparison performance over different splits and metrics in Table I. Our proposed STDMANet was better than the baseline methods in most cases. In experiment B, the CG performance of our method was better, and the BSF performance was worse. The reason was that we used 2-D Gaussian distribution to represent point targets, resulting in a larger nonzero area for each target in the final probability map, which led to an increase in the standard deviation of the output probability map and made the BSF metric slightly worse. We further show the inference time and fps in Table II. Our method achieves better performance while maintaining comparable speed as the fastest method.

2) Performance on the SIATD Dataset: As shown in Table III, we provided the F_1 score of our model on the SIATD dataset versus the baselines. For those methods that used the SIATD dataset and reported performance by F_1 score, we used the result in the original paper, and we also selected some of the baseline methods that they implemented. For other methods, we tested them with checkpoints provided by their official open-source code. We could not retrain them because they were difficult to manage with point annotations.

In comparison with recent baselines, our proposed STDMANet improves the overall F_1 score by 0.1682, achieving the best performance among all methods. In addition, our model performed well in terms of precision and recall. The precision is only 0.0115 lower than the highest IAANet, while the recall is the highest and surpasses the second highest ALCNet at 0.1336. In contrast, the baseline methods could only perform well in one metric between precision and recall, while our model achieved similar performance on two metrics.

Since the SIATD dataset contains rich scenarios, we select several typical scenes and evaluate the performance of those sequences. The results are shown in Table IV. The experimental results show that our method maintains high detection accuracy and consistency in all kinds of scenarios.

3) Performance on Targets With Different Speeds: This article focuses on the effectiveness of temporal information for small target detection. However, different targets have different speeds. In this part, we design experiments to explore the performance differences between our methods in detecting targets with different speeds. We believe that, for those targets with a strong contrast with the background, regardless of the

TABLE I

RESULTS OF SEVERAL METRICS ON SEVERAL DSAT DATASET SPLITS. WE SHOW MULTIPLE RESULTS OF DIFFERENT METHODS WITH RESPECT TO THEIR METRICS. FOR THE RESULTS OF EACH SEQUENCE, THE FIRST LINE IS THEIR NAME IN THE CORRESPONDING PAPER, AND THE SECOND LINE IS THEIR SEQUENCE INDEX IN DSAT

Exp	Test Sequences	Average SCR	Metric	Method	Results					
					Overall	Data2 Seq 8				
A	8	2.91	$P_d \uparrow$	Yan et al. [46]	**0.9858**	**0.9858**				
				STDMANet (ours)	0.9812	0.9812				
			$F_a \downarrow$	Yan et al. [46]	0.1435	0.1435				
				STDMANet (ours)	**0.0025**	**0.0025**				
					Overall	Sequence1 Seq 8	Sequence2 Seq 2			
B	2,8	5.47	Precision ↑	Yao et al. [37]	0.987	**0.984**	0.989			
				STDMANet (ours)	**0.988**	0.981	**0.993**			
			Recall ↑	Yao et al. [37]	0.994	0.992	0.996			
				STDMANet (ours)	**0.997**	**0.998**	**0.997**			
					Overall	Sequence 4 Seq 10	Sequence 5 Seq 6	Sequence 6 Seq 18		
C	6,10,18 [1]	3.13	$CG_2 \uparrow$	ASTTV-NTLA [45]	13.89	**27.49**	5.88	**8.30**		
				STDMANet (ours)	**15.94**	16.42	**30.83**	3.73		
					Overall	scene(f) Seq 6	scene(g) Seq 20	scene(h) Seq 22	scene(i) Seq 8	
D	6,8,20,22	3.21	BSF ↑	Wu et al. [50]	**23.098**	**16.391**	**31.017**	25.615	**18.686**	
				STDMANet (ours)	19.714	14.899	18.297	**29.446**	13.725	
			$CG_1 \uparrow$	Wu et al. [50]	1.395	1.498	1.465	1.260	1.391	
				STDMANet (ours)	**1.596**	**1.958**	**1.669**	**1.378**	**1.434**	
					Overall	Seq 3	Seq 4	Seq 16	Seq 18	
E	3,4,16,18	5.96	Precision ↑	Zhu et al. [55]	0.9759	**1.0000**	**1.0000**	0.9898	0.9380	
				STDMANet (ours)	**0.9866**	0.9141	0.9786	**0.9960**	**1.0000**	
			Recall ↑	Zhu et al. [55]	0.9727	**0.9865**	**0.9988**	0.9838	0.9380	
				STDMANet (ours)	**0.9920**	0.9192	0.9975	**0.9960**	**1.0000**	
			$F_1 \uparrow$	Zhu et al. [55]	0.9717	**0.9932**	**0.9994**	0.9789	0.9380	
				STDMANet (ours)	**0.9893**	0.9166	0.9880	**0.9960**	**1.0000**	
					Overall	Seq 7	Seq 10	Seq 14	Seq 15	Seq 17
F	Our Split 7,10,14,15,17	1.83	Precision ↑	STDMANet (ours)	0.8378	0.9598	0.6696	0.7431	0.9397	0.9950
			Recall ↑		0.9031	0.9623	0.6900	0.8623	0.9987	1.0000
			$F_1 \uparrow$		0.8692	0.9610	0.6796	0.7987	0.9683	0.9975

TABLE II

INFERENCE TIME AND fps BETWEEN OUR MODEL AND OTHER METHODS

Method	Input Size	Inference Time ↓	FPS ↑
Yan et al. [46]	(256, 256)	0.0596	16.78
Yao et al. [37]	(256, 256)	**0.0282**	**35.50**
Wu et al. [50]	(256, 256)	0.0693	14.43
ASTTV-NTLA [45]	(250, 250)	282.76	0.0035
Zhu et al. [55]	(256, 256)	1.4354	0.6967
STDMANet(ours)	(256, 256)	0.0367	27.26

TABLE III

RESULTS ON SIATD DATASET. WE SHOW THE RESULTS OF DIFFERENT METHODS WITH RESPECT TO THE PRECISION, RECALL, AND F_1 SCORE. THE HIGHEST F_1 SCORE IS IN BOLDFACE, AND THE SECOND HIGHEST IS UNDERLINE

Method	Precision	Recall	F_1 score
RIPT [56]	0.30	0.75	0.43
TIPI [57]	0.01	0.09	0.02
FKRW [58]	0.0033	0.0338	0.0061
ASTTV-NTLA [45]	0.60	0.52	0.56
Sun et al. [14]	0.92	0.63	0.75
MDvsFA [22]	0.9801	0.5623	0.7146
ACM [25]	0.0197	0.7986	0.0384
ALCNet [26]	0.4285	0.8397	0.5674
IAANet [27]	**0.9870**	0.6814	0.8062
DNANet [21]	0.0672	0.0682	0.0677
STDMANet(ours)	0.9755	**0.9733**	**0.9744**

speed, the detection performance can be guaranteed. However, for those targets with little difference from the background, if the speed is too slow, we cannot get enough information to distinguish the target; on the contrary, if the speed is too fast, the motion blur of the target will also make the detection more difficult. In order to verify this, we first count the speed of the target, where the speed is defined as the pixel displacement under the fixed time window k ($k = 20$ in DSAT and $k = 5$ in SIATD). After that, we calculate the detection

[1]To mention that the original paper only use part of sequences and there is no description of the start and the end frames, so the results are for reference only.

TABLE IV

SEVERAL TYPICAL SCENARIOS ON SIATD DATASET. WE SHOW THE RESULTS OF DIFFERENT SCENARIOS WITH F_1 SCORE TO EVALUATE THE PERFORMANCE OF OUR MODEL ON DIFFERENT SCENES. A: BUILDING; B: RIVER; C: BRIDGE; D: ROAD; E: FIELD; F: WOODS; G: HILLY; H: CLOUD; AND I: CITY

Perspective	Down	Down	Down	Down	Down	Down	Down	Down	Down	Down	Down	Down	Down	Down	Down
Scene	A	B	E	F	G	AB	AD	AF	BC	BE	DE	EF	EG	FG	ABD
F_1 score	0.9729	0.9520	0.9693	0.9910	0.9807	0.9871	0.9704	0.9655	0.9676	0.9854	0.9889	0.9551	0.9694	0.9045	0.9786
Perspective	Down	Down	Down	Down	Down	Down	Down	Down	Down	Down	Down	Down	Down	Middle	Up
Scene	ABE	ABF	BCD	BCE	BDF	BEF	EFG	ABCD	ABCF	ABEF	ABFG	ABEFG	ABCDEF	HI	H
F_1 score	0.9800	0.9791	0.9929	0.9885	0.9893	0.9708	0.9423	0.9933	0.9339	0.9923	0.9867	0.9627	0.9973	0.9775	0.9695

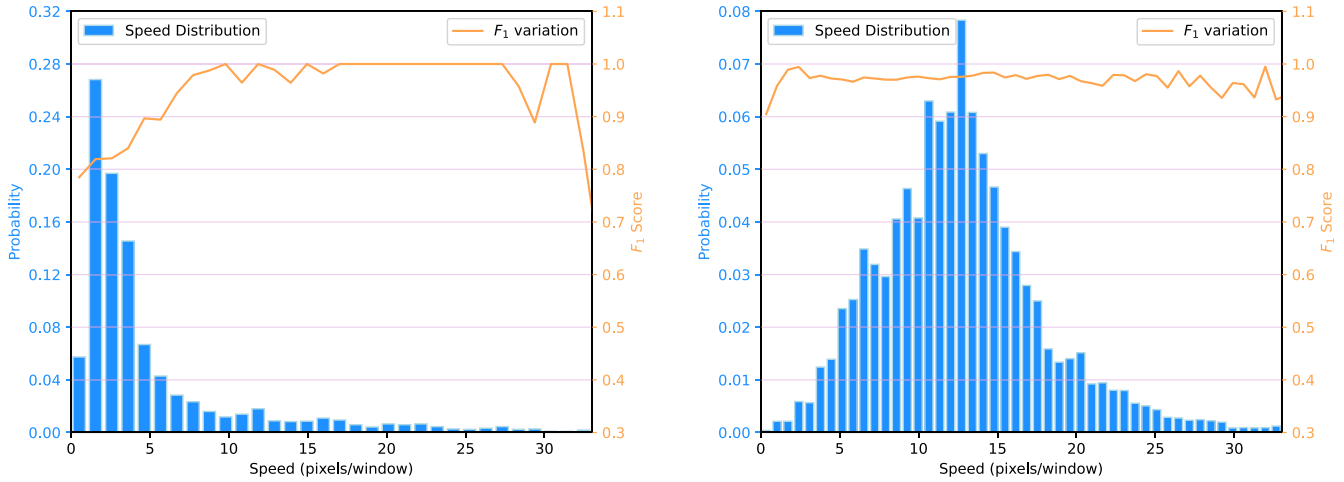

Fig. 5. Chart of target speed distribution and F_1 metric variation in (a) DSAT and (b) SIATD datasets. The abscissa is the speed range of the targets, the blue bar chart is the probability distribution of the target speed in the dataset, and the orange line chart is the average F_1 variation of target detection according to different speed intervals.

TABLE V

ABLATION RESULTS ON DSAT DATASET. WE SHOW THE PERFORMANCE GAP UNDER DIFFERENT FEATURE EXTRACTION STRATEGIES TO DEMONSTRATE THE EFFECTIVENESS OF EACH COMPONENT

Index	Background Alignment	Differential Path	Dynamic Path	Static Path	Precision	Recall	F_1 score	F_1 Diff
1	✓	✓	✓	✓	0.8378	0.9031	0.8692	0.0000
2		✓	✓	✓	0.7437	0.7685	0.7559	-0.1133
3	✓		✓	✓	0.8332	0.8654	0.8490	-0.0202
4	✓	✓		✓	0.8412	0.8668	0.8538	-0.0154
5	✓	✓	✓		0.8453	0.8795	0.8621	-0.0071
6	✓	✓			0.8303	0.8637	0.8467	-0.0225
7	✓		✓		0.8169	0.8341	0.8254	-0.0438
8	✓			✓	0.6053	0.6189	0.6120	-0.2572

performance of our model for targets with different speeds, as shown in Fig. 5. The experimental results show that our method shows performance degradation when the speed is too slow and too fast. At the same time, the decline of the SIATD dataset with high SCR is smaller, while the performance decline of the DSAT dataset with lower SCR is greater. These results are consistent with our previous analysis.

4) Ablation Study on the DSAT Dataset: In order to verify the effectiveness of our model components, we conducted several ablation experiments on the DSAT dataset. We evaluated the F_1 score after removing specific components to assess their contribution to the model. The experimental results are shown in Table V. The first row of the table shows the performance of our entire model. The second row shows the

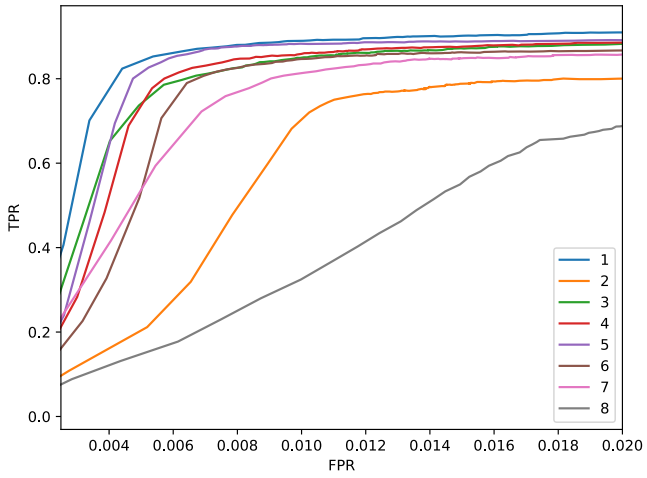

Fig. 6. ROC results of ablation study. The legends in the figure correspond to the index of the ablation experiment in Table V.

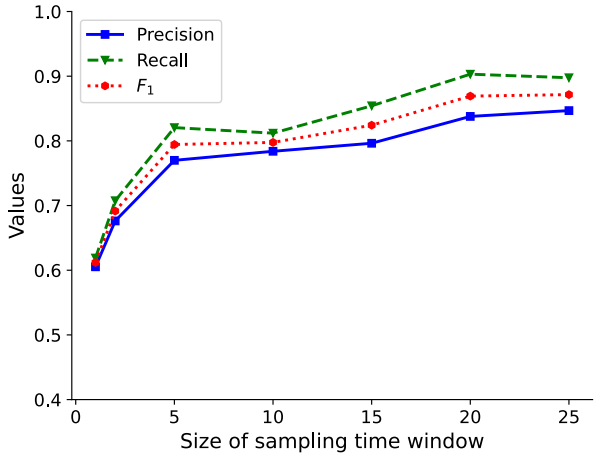

Fig. 7. Performance under different sampling frames, where the y-axis represents the value of different metrics and the x-axis represents the size of the time window.

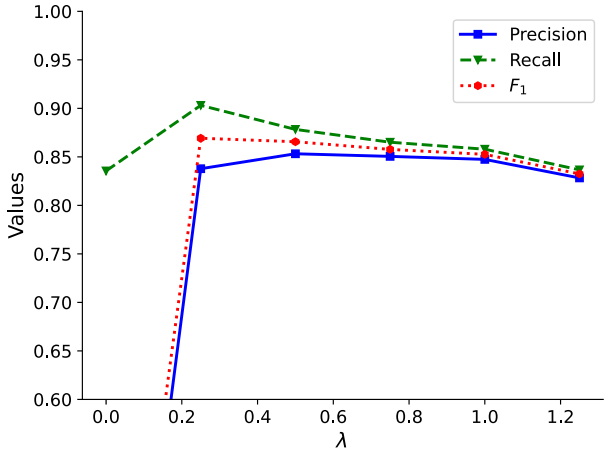

Fig. 8. Performance under different λ's. To better visualize the changing trend of various metrics with the increase in λ, our chart does not include precision when λ equals 0, which is 0.0813.

result of removing the background alignment. The other lines show the ablation of feature extraction paths at different time scales. Fig. 6 further proves the credibility of the ablation by Monte Carlo simulation.

From the second row of the table, we can see that, after removing the background alignment, the result of the model drops by 0.1133, indicating that feature extraction based on background alignment was vital for small target detection. From the results of the third row to the fifth row, we can see that combining any two feature paths could better complete the task of small target detection, but they were slightly lower than the combination of three feature paths. The results of the last three rows show that only one feature path could be used to complete the detection task, but its performance was limited. Among them, the performance of the differential path alone was the best, that of the dynamic path was the second, and that of the static path was the worst. This result was consistent with our analysis that it was far from enough to use only single-frame information when the target was very small (even less than 3×3). We must use temporal, motion, and brightness change information in multiple frames.

5) Variation of Time Window k on the DSAT Dataset: To verify the necessity of using multiframe images to obtain temporal information in dim and small infrared target detection, we transformed the sampling frame number k from 1 to 25 to explore the influence of the time window on final performance. The results are shown in Fig. 7. As can be seen from the figure, the result of the model with multiframe input was better than the performance of models using only the current frame to extract spatial information. When the sampling frame is greater than 20 frames, the final F_1 score is even higher than that of a single frame by more than 0.25, which verified that dim targets needed to be detected through temporal information. On the other hand, we observed that the final performance of the model no longer increases significantly when the time window exceeds 20, indicating that most of the motions in the DSAT dataset were completed within 20 frames. Therefore, we used 20 frames as the default time window size for the DSAT dataset. Similarly, for the SIATD dataset, the size of the time window was set to 5. In our further experiments, the effect of increasing the time window of the SIATD dataset was completely within the error range.

6) Variation of Weighting Factor λ on the DSAT Dataset: We explored the changes in several metrics with λ changes, as shown in Fig. 8. When λ is 0, only the loss of the target area is calculated, and the precision is reduced to 0.0813, because we do not add constraints to the background, resulting in a large number of false detection and a significant reduction in precision. When we gradually increase λ from 0.25, it means that we increase the constraint on the background area and reduce the constraint on the target area, and the task of suppressing the background will occupy more weight than finding the target; thus, the importance of reducing false detection increases, and the importance of reducing missed detection decreases. The curve in the figure proves this analysis. The decrease in precision is less than that of recall with the increase in lambda, which means that the network pays more attention to reduce false detection rather than missed detection. The experiment shows that missed detection and false detection can be balanced by λ. We set lambda to 0.25 to maximize the F_1 score.

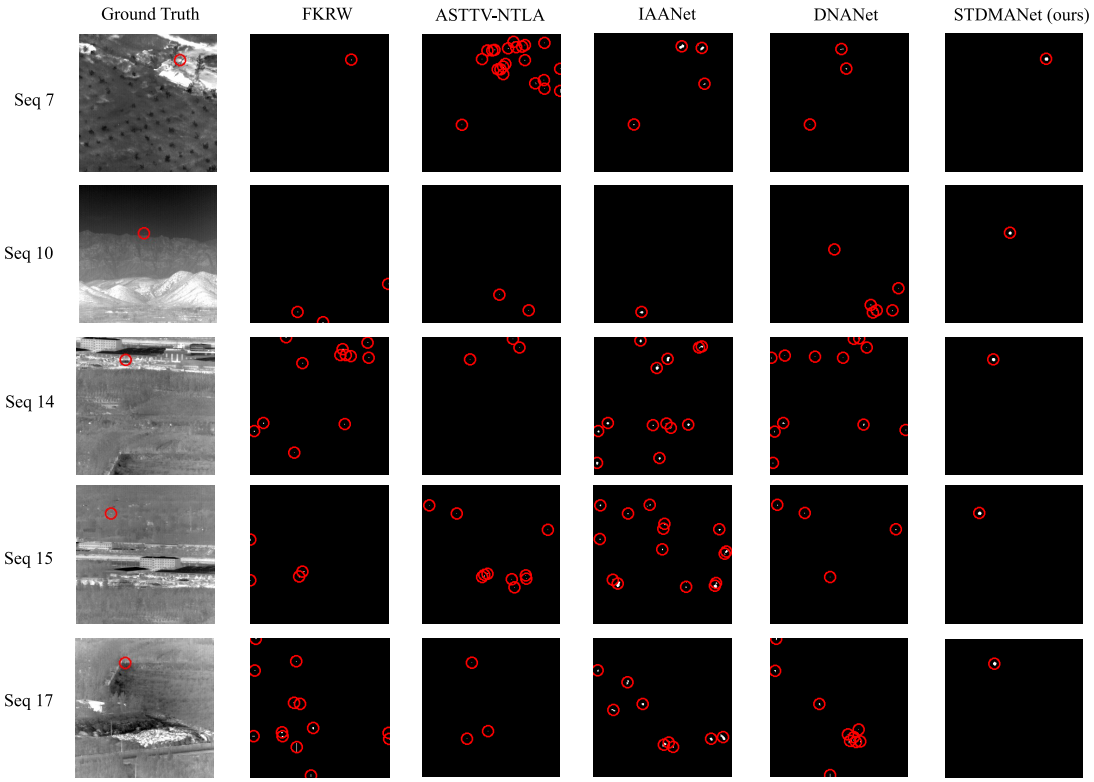

Fig. 9. Visualization results on the DSAT dataset. The time indices of given image sequences Data7, Data10, Data14, Data15, and Data17 are 40, 29, 295, 30, and 32, respectively.

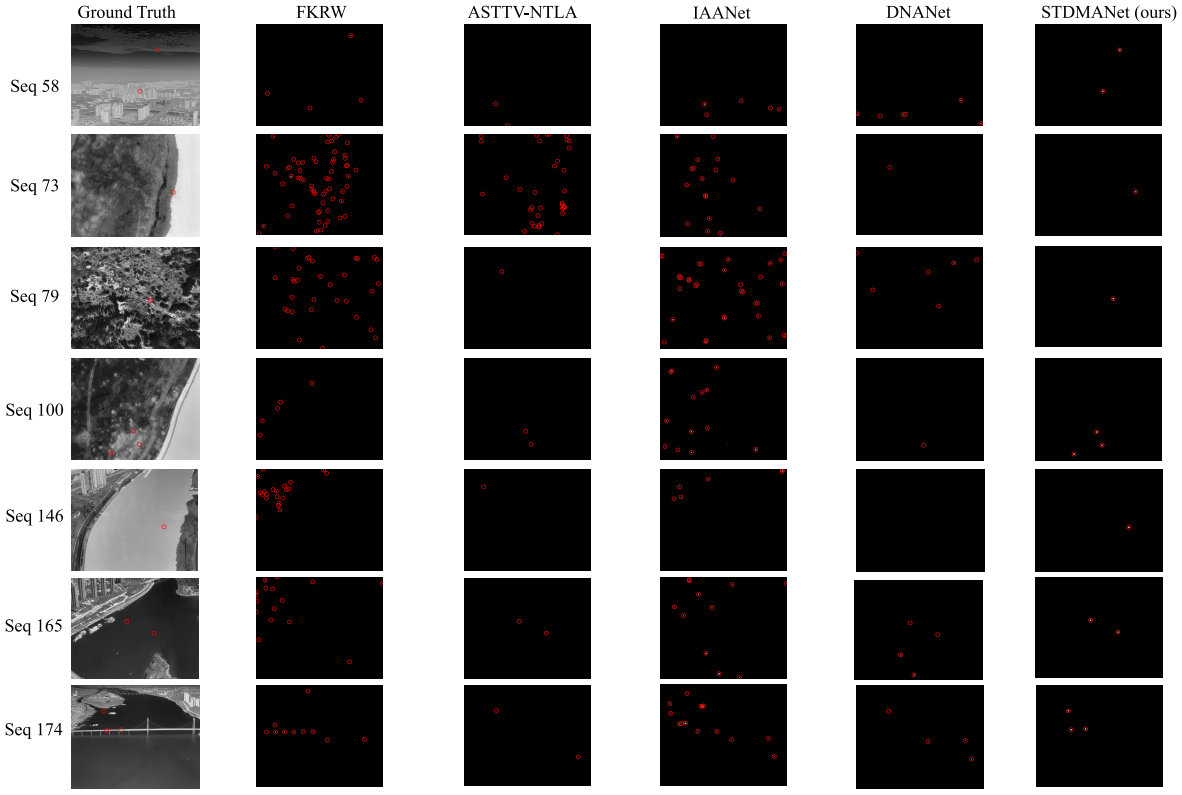

Fig. 10. Visualization results on the SIATD dataset. The time indices of given image sequences 58, 73, 79, 100, 146, 165, and 174 are 116, 360, 130, 127, 189, 100, and 106, respectively.

E. Qualitative Results

The visualization of the DSAT and SIATD datasets over our STDMANet and other methods is shown in Figs. 9 and 10, respectively. For those methods that need to adjust the parameters, we adjust the parameters of each sequence so that their result could include the correct target as much as possible.

In the visualization of the DSAT dataset, we showed the results of five scenes. Among them, three sequences (Seq7, Seq14, and Seq17) showed that the target entered an area where the background was similar to the appearance of the target, and two scenes (Seq10 and Seq15) showed that the target was extremely dim and difficult to discern. From the results of five scenes, the observation was that our proposed STDMANet could deal with complex backgrounds—our model not only could detect dim targets through spatial information but also could detect targets that could not be discerned by spatial information through temporal clues.

Similarly, in the SIATD dataset with higher resolution, more targets, and richer scenes, we also selected different sequences for five scenes. In sequence 58, the two targets entered clouds in the sky and buildings on the ground, respectively. In sequence 79, the target entered a strong reflection area caused by the superposition of clouds and roads. In sequence 100, all three targets entered the clouds from an overlooking perspective. In sequences 146 and 165, the background of the target was a strong reflection on the water surface and a weak reflection on the water surface, respectively. From the visualization results, we could see that our proposed STDMANet performed well in all scenarios, which further verified the ability of our model to deal with complex backgrounds.

The video frames of the above sequence were specially selected, in which it was difficult to distinguish the target from the background in the current frame. In this case, our model could make good use of the historical motion of the target and find clues to the location of the target in the current frame. Other baseline methods either added a large number of false detection to include the correct target or could only locate the high-brightness small patches in the video frame and could not correctly identify the target, while our proposed STDMANet could maintain both false detection and missed detection at a low level at the same time.

To summarize, through several modules inspired by the characteristics of infrared small targets, our model could extract high-quality ST features for small targets under a reasonable time window, generate prediction outputs of low false detection rate and low missed detection rate, and clearly exceed the latest baseline methods. The extensive experimental results showed that we had successfully solved the problems that we raised in the experimental design process.

V. Conclusion

In this article, we presented the STDMANet for infrared small target detection. A network with temporal and SA mechanisms was established through the temporal multiscale feature extractor and the spatial multiscale feature refiner, which was suitable for multiframe infrared dim target detection. The temporal multiscale feature extractor extracted the ST features accumulated by the continuous motion of the target under different time steps and improved the quality of feature extraction through the attention mechanism. At the same time, in order to prevent small objects from being annihilated by noise, we maintained the feature map of the original image size throughout the process and maintained the features related to the shallow position (position-aware) and deep semantic (semantic-aware) through feature reuse. We conducted extensive ablation studies and conducted a comparison with today's mainstream methods. The performance of the proposed STDMANet on the open SIATD and DSAT datasets showed that the proposed method significantly outperformed the compared pure model-driven methods and pure data-driven networks. This indicated that one should pay more attention to the detection of infrared dim targets through the combination of a deep network, temporal information, spatial information, and differential information. In the future, we can find more efficient ways to simplify computation and reduce the runtime of processing.

Acknowledgment

The authors would like to thank the editor and the anonymous reviewers who gave constructive comments and helped to improve the quality of this article.

References

[1] Y. Gu, C. Wang, B. Liu, and Y. Zhang, "A kernel-based nonparametric regression method for clutter removal in infrared small-target detection applications," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 3, pp. 469–473, Jul. 2010.

[2] X. Wang, Z. Peng, D. Kong, and Y. He, "Infrared dim and small target detection based on stable multisubspace learning in heterogeneous scene," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5481–5493, Oct. 2017.

[3] X. Sun, X. Liu, Z. Tang, G. Long, and Q. Yu, "Real-time visual enhancement for infrared small dim targets in video," *Infr. Phys. Technol.*, vol. 83, pp. 217–226, Jun. 2017.

[4] T. Zhang, Z. Peng, H. Wu, Y. He, C. Li, and C. Yang, "Infrared small target detection via self-regularized weighted sparse model," *Neurocomputing*, vol. 420, pp. 124–148, Jan. 2021.

[5] Q. Song, Y. Wang, K. Dai, and K. Bai, "Single frame infrared image small target detection via patch similarity propagation based background estimation," *Infr. Phys. Technol.*, vol. 106, May 2020, Art. no. 103197.

[6] K. Luo, "Space-based infrared sensor scheduling with high uncertainty: Issues and challenges," *Syst. Eng.*, vol. 18, no. 1, pp. 102–113, Jan. 2015.

[7] J. Guo and G. Chen, "Analysis of selection of structural element in mathematical morphology with application to infrared point target detection," *Proc. SPIE*, vol. 6835, pp. 178–185, Jan. 2008.

[8] Y. He, M. Li, J. Zhang, and Q. An, "Small infrared target detection based on low-rank and sparse representation," *Infr. Phys. Technol.*, vol. 68, pp. 98–109, Jan. 2015.

[9] R. Kerekes and B. V. K. V. Kumar, "Enhanced video-based target detection using multi-frame correlation filtering," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 45, no. 1, pp. 289–307, Jan. 2009.

[10] P.-Y. Lv, S.-L. Sun, C.-Q. Lin, and G.-R. Liu, "Space moving target detection and tracking method in complex background," *Infr. Phys. Technol.*, vol. 91, pp. 107–118, Jun. 2018.

[11] S. Moradi, P. Moallem, and M. F. Sabahi, "A false-alarm aware methodology to develop robust and efficient multi-scale infrared small target detection algorithm," *Infr. Phys. Technol.*, vol. 89, pp. 387–397, Mar. 2018.

[12] X. Bai and Y. Bi, "Derivative entropy-based contrast measure for infrared small-target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2452–2466, Apr. 2018.

[13] Z. Song and B. Hui, "A dataset for infrared image dim-small aircraft target detection and tracking under ground/air background," *Sci. Data Bank*, vol. 5, p. 12, Jan. 2020. [Online]. Available: https://www.scidb.cn/en/detail?dataSetId=720626420933459968&dataSetType=journal

[14] X. Sun et al. (Feb. 2022). *A Dataset for Small Infrared Moving Target Detection Under Clutter Background*. [Online]. Available: https://www.scidb.cn/en/detail?dataSetId=80802594687025152O& dataSetType=journal

[15] X. Shao, H. Fan, G. Lu, and J. Xu, "An improved infrared dim and small target detection algorithm based on the contrast mechanism of human visual system," *Infr. Phys. Technol.*, vol. 55, no. 5, pp. 403–408, Sep. 2012.

[16] S. Qi, J. Ma, C. Tao, C. Yang, and J. Tian, "A robust directional saliency-based method for infrared small-target detection under various complex backgrounds," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 495–499, May 2013.

[17] C. Q. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4996–5009, Dec. 2013.

[18] C. L. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581, Jan. 2014.

[19] J. Han, Y. Ma, B. Zhou, F. Fan, K. Liang, and Y. Fang, "A robust infrared small target detection algorithm based on human visual system," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2168–2172, Dec. 2014.

[20] H. Zhang, L. Zhang, D. Yuan, and H. Chen, "Infrared small target detection based on local intensity and gradient properties," *Infr. Phys. Technol.*, vol. 89, pp. 88–96, Mar. 2018.

[21] B. Li et al., "Dense nested attention network for infrared small target detection," 2021, *arXiv:2106.00487*.

[22] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8509–8518.

[23] B. McIntosh, S. Venkataramanan, and A. Mahalanobis, "Infrared target detection in cluttered environments by maximization of a target to clutter ratio (TCR) metric using a convolutional neural network," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 1, pp. 485–496, Feb. 2021.

[24] M. Zhao, L. Cheng, X. Yang, P. Feng, L. Liu, and N. Wu, "TBC-Net: A real-time detector for infrared small target detection using semantic constraint," 2019, *arXiv:2001.05852*.

[25] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 950–959.

[26] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Attentional local contrast networks for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9813–9824, Nov. 2021.

[27] K. Wang, S. Du, C. Liu, and Z. Cao, "Interior attention-aware network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5002013.

[28] F. Liu, C. Gao, F. Chen, D. Meng, W. Zuo, and X. Gao, "Infrared small-dim target detection with transformer under complex backgrounds," 2021, *arXiv:2109.14379*.

[29] Y. Qi and G. An, "Infrared moving targets detection based on optical flow estimation," in *Proc. Int. Conf. Comput. Sci. Netw. Technol.*, Dec. 2011, pp. 2452–2455.

[30] Y. Lu, S. Huang, and W. Zhao, "Sparse representation based infrared small target detection via an online-learned double sparse background dictionary," *Infr. Phys. Technol.*, vol. 99, pp. 14–27, Jun. 2019.

[31] F. Zhao, T. Wang, S. Shao, E. Zhang, and G. Lin, "Infrared moving small-target detection via Spatio-Temporal consistency of trajectory points," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 1, pp. 122–126, Jan. 2020.

[32] S. Li, C. Li, X. Yang, K. Zhang, and J. Yin, "Infrared dim target detection method inspired by human vision system," *Optik*, vol. 206, Mar. 2020, Art. no. 164167.

[33] G. Wang, B. Tao, X. Kong, and Z. Peng, "Infrared small target detection using nonoverlapping patch spatial–temporal tensor factorization with capped nuclear norm regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5001417.

[34] C. Kwan and D. Gribben, "Practical approaches to target detection in long range and low quality infrared videos," *Signal Image Process., Int. J.*, vol. 12, no. 3, pp. 1–16, Jun. 2021.

[35] C. Kwan, D. Gribben, and B. Budavari, "Target detection and classification performance enhancement using super-resolution infrared videos," *Signal Image Process.: Int. J.*, vol. 12, no. 2, pp. 33–45, Apr. 2021.

[36] X. Ying et al., "Local motion and contrast priors driven deep network for infrared small target superresolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5480–5495, 2022.

[37] S. Yao, Q. Zhu, T. Zhang, W. Cui, and P. Yan, "Infrared image small-target detection based on improved FCOS and spatio-temporal features," *Electronics*, vol. 11, no. 6, p. 933, Mar. 2022.

[38] J. Du et al., "Multiple frames based infrared small target detection method using CNN," in *Proc. 4th Int. Conf. Algorithms, Comput. Artif. Intell.*, Dec. 2021, pp. 1–6.

[39] J. Du et al., "A spatial–temporal feature-based detection framework for infrared dim small target," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3000412.

[40] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, Jul. 2003.

[41] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[42] T. Ma, Z. Yang, J. Wang, S. Sun, X. Ren, and U. Ahmad, "Infrared small target detection network with generate label and feature mapping," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6505405.

[43] S. Zhou, Z. Gao, and C. Xie, "Dim and small target detection based on their living environment," *Digit. Signal Process.*, vol. 120, Jan. 2022, Art. no. 103271.

[44] S. Suzuki and K. Be, "Topological structural analysis of digitized binary images by border following," *Comput. Vis., Graph., Image Process.*, vol. 30, no. 1, pp. 32–46, 1985.

[45] T. Liu et al., "Nonconvex tensor low-rank approximation for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5614718.

[46] P. Yan, S. Yao, Q. Zhu, T. Zhang, and W. Cui, "Real-time detection and tracking of infrared small targets based on grid fast density peaks searching and improved KCF," *Infr. Phys. Technol.*, vol. 123, Jun. 2022, Art. no. 104181.

[47] J. Ai, R. Tian, Q. Luo, J. Jin, and B. Tang, "Multi-scale rotation-invariant haar-like feature integrated CNN-based ship detection algorithm of multiple-target environment in SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10070–10087, Dec. 2019.

[48] J. Ai et al., "Robust CFAR ship detector based on bilateral-trimmed-statistics of complex ocean scenes in SAR imagery: A closed-form solution," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 3, pp. 1872–1890, Jan. 2021.

[49] J. Ai, Y. Mao, Q. Luo, L. Jia, and M. Xing, "SAR target classification using the multikernel-size feature fusion-based convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5214313.

[50] L. Wu, S. Fang, Y. Ma, F. Fan, and J. Huang, "Infrared small target detection based on gray intensity descent and local gradient watershed," *Infr. Phys. Technol.*, vol. 123, Jun. 2022, Art. no. 104171.

[51] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[52] G. Bradski, "The OpenCV library," *Dr. Dobb's J. Softw. Tools Prof. Program.*, vol. 25, no. 11, pp. 120–123, 2000.

[53] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[55] R. Zhu and L. Zhuang, "Unsupervised infrared small-object-detection approach of spatial–temporal patch tensor and object selection," *Remote Sens.*, vol. 14, no. 7, p. 1612, Mar. 2022.

[56] Y. Dai and Y. Wu, "Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3752–3767, Aug. 2017.

[57] C. Gao, L. Wang, Y. Xiao, Q. Zhao, and D. Meng, "Infrared small-dim target detection based on Markov random field guided noise modeling," *Pattern Recognit.*, vol. 76, pp. 463–475, Apr. 2018.

[58] Y. Qin, L. Bruzzone, C. Gao, and B. Li, "Infrared small target detection based on facet kernel and random Walker," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7104–7118, Sep. 2019.

Puti Yan received the B.S. degree in electronical information science and technology and the M.S. degree in aerospace science and technology from the Harbin Institute of Technology, Harbin, China, in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree in aerospace science and technology.

He has published research papers in IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Runze Hou received the B.E. degree from the School of Automation, Southeast University, Nanjing, China, in 2021. He is currently pursuing the M.S. degree with the Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Beijing, China.

His research interests include deep learning, multimodal learning, and infrared target detection.

Xuguang Duan received the B.E. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2018, where he is currently pursuing the master's degree with the Department of Computer Science and Technology.

He has published some research papers in top conferences and journals, including International Conference on Advances in Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (ICML), IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), and ACM Multimedia. His research interests include machine learning, neural-symbolic systems, and video understanding.

Chengfei Yue (Member, IEEE) received the B.E. degree from the Harbin Institute of Technology, Harbin, China, in 2013, and the Ph.D. degree from the National University of Singapore, Singapore, in 2019.

He is currently an Associate Professor with the Harbin Institute of Technology (Shenzhen) (HITSZ), Shenzhen, China, where he is also the Lab Director of the On-Orbit Manipulation Dynamics and Control.

Dr. Yue is also a Committee Member of the Committee of Space Intelligence, Chinese Society of Space Research, and a member of American Institute of Aeronautics and Astronautics (AIAA). He was also awarded as an Outstanding Reviewer of *Mechatronics* (IFAC) journal. He is a reviewer of more than ten reputational journals and an Associate Editor of *Frontiers in Control Engineering*.

Xin Wang (Member, IEEE) received the B.E. degree from Zhejiang University, Hangzhou, China, in 2011, the Ph.D. degree in computing science from Simon Fraser University, Burnaby, BC, Canada, in 2016, and the Ph.D. degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2017.

He is currently an Assistant Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. He has published over 100 high-quality research papers in top conferences and journals, including International Conference on Machine Learning (ICML), International Conference on Advances in Neural Information Processing Systems (NeurIPS), IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), ACM International Conference on Knowledge Discovery and Data Mining (KDD), International World Wide Web Conference (WWW), ACM International Conference on Special Interest Group on Information Retrieval (SIGIR), and ACM Multimedia. His research interests include multimedia intelligence and recommendation in social media.

Dr. Wang was a recipient of the 2020 ACM China Rising Star Award and the 2022 IEEE TCMC Rising Star Award.

Xibin Cao received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1985, 1988, and 1991, respectively.

Since 1991, he has been with the School of Astronautics, Harbin Institute of Technology, where he is currently a Full Professor. Since 2009, he has been the Dean of the Astronautics School, Harbin Institute of Technology. Since 2015, he has been an Assistant President of the Harbin Institute of Technology.

Prof. Cao is a fellow of the Chinese Academy of Engineering. He won the Distinguished Professor of Yangtze River Scholar, Ministry of Education of China, in 2005.