文章编号：1001-9014(2021)01-0122-11

DOI:10.11972/j.issn.1001-9014.2021.01.017

# Point target detection based on deep spatial-temporal convolution neural network

LI Miao[1*], LIN Zai-Ping[1], FAN Jian-Peng[1], SHENG Wei-Dong[1], LI Jun[1], AN Wei[1], LI Xin-Lei[2]

(1. College of electronic science and technology, National University of Defense Technology, Changsha 410073, China;
2. The Xian Chinese Space Tracking Control Center, Xian, Shanxi 710000, China)

**Abstract**: Point target detection in Infrared Search and Track (IRST) is a challenging task. due to less information. Traditional methods based on hand-crafted features are hard to finish detection intelligently. A novel deep spatial-temporal convolution neural network is proposed to suppress background and detect point targets. The proposed method is realized based on fully convolution network. So input of arbitrary size can be put into the network and correspondingly-sized output can be obtained. In order to meet the requirement of real time for practical application, the factorized technique is adopted. 3D convolution is decomposed into 2D convolution and 1D convolution, and it leads to significantly less computation. Multi-weighted loss function is designed according to the relation between prediction error and detection performance for point target. Number-balance weight and intensity-balance weight are introduced to deal with the imbalanced sample distribution and imbalanced error distribution. The experimental results show that the proposed method can effectively suppress background clutters, and detect point targets with less runtime.

**Key words**: point target detection, infrared search and track (IRST), background suppression, convolution neural network (CNN), spatial-temporal detection

**PACS**: : 84.40.Xb

# 基于深度时空卷积神经网络的点目标检测

李 淼[1*], 林再平[1], 樊建鹏[1], 盛卫东[1], 李 骏[1], 安 玮[1], 李昕磊[2]

(1. 国防科技大学 电子科学学院,湖南 长沙 410073;
2. 西安卫星测控中心,陕西 西安 710000)

**摘要**:由于点目标可用信息少,点目标检测技术是红外搜索与跟踪系统(IRST)中的挑战性难点。基于人工提取特征的传统目标检测,智能化水平低,对点目标检测的难度大。针对此问题,提出一种新的基于深度时空卷积神经网络的点目标检测方法。该方法采用全卷积架构,输入输出尺度相同,可用于处理任意尺度图像。为了提高实时性,卷积分解技术被引入3D时空卷积处理中,将复杂3D时空卷积分解为低复杂度的2D空域卷积和1D时域卷积。根据点目标特点,多权值损失函数被提出,分别采用样本均衡因子和能量均衡因子降低样本不均衡和误差分布不均衡对点目标检测性能的影响。测试结果表明,该方法能够有效抑制复杂背景杂波,并以较低计算量实现点目标检测。

**关 键 词**:点目标检测;红外搜索与跟踪(IRST);背景抑制;卷积神经网络(CNN);时空检测

**中图分类号**:TP753 **文献标识码**:A

## Introduction

Infrared Search and Track (IRST) systems were developed to automatically search, capture and track small incoming targets from infrared sequences. They have been wildly applied to many important fields, including unmanned aerial vehicle (UAV) defense, territory surveillance, space situation awareness (SSA), precise

**Received date**: 2020-01-09, **revised date**: 2020-07-15 **收稿日期**:2020-01-09,**修回日期**:2020-07-15

**Foundation items**: Supported by the National Natural Science Foundation of China (61921001)

**Biography**: LI Miao (1988-), male, Weifang, Lecturer, Doctor. Research area involves Spatial information processing. E-mail: lm8866@nudt.edu.cn.

* **Corresponding author**: E-mail: lm8866@nudt.edu.cn

guidance, and so on[1-5]. Obviously, small target detection is the key process in IRST, and the detection performance decides whether the IRST is erected success or not. Unfortunately, point target detection is still a great challenge for many negative effects. (1) Although the targets having a total spatial extent of less than 80 pixels (9×9) are defined as small target by SPIE, the targets in IRST may occupy fewer pixels (e. g. , 3×3 pixels)[6]. (2) Very few or even no obvious textural features can be extracted by hand due to limited spatial resolution[7]. (3) The intensity of target signal is very weak because of distant observation. (4) The sensor noise and background clutter may be salient. In fact, the scattered clouds appearing at the edge of the big clouds are very similar as point targets. For the above reasons, the weak point targets are usually submerged by noises and clutters, and they are hard to be detected simultaneously with low false alarm ratio, high probability of detection and strong robustness.

Many methods have been presented in the past decades, such as mean subtraction filter [8], median subtraction filter [9], TopHat filter [10], Max-Mean/Max-Median filter [11], spatial-temporal accumulative difference [12], matched filter [13], and local contrast measure (LCM)[7]. However, these methods cannot detect point targets in complex background. Because they were designed based on hand-crafted features, the performances heavily depend on the completeness and accuracy of the features. Unmatched hand-crafted features will result in the performance degradation, and consequently the missed detection ratio and false-detecting ratio will increase rapidly.

Generally, those assumptions often adopted to support hand-crafted filters include: (1) point target can be modeled as 2D Gauss model[3, 14]. (2) The target has a signature of discontinuity with its neighboring regions and concentrates in a relatively small region. (3) The background is consistent with its neighboring regions. However, in practical application, single hand-crafted filter cannot well deal with different scenes, especially when there are heavy clutters in the background. In order to overcome the defects of single hand-crafted filter, the combination of multiple filters based on different features is often used in practical system. But the combination and order of different filters are still an intractable difficulty. Besides, large number parameters of multiple filters must be tuned very carefully. Unfortunately, this work is very hard for human, and the algorithm optimization heavily depend on the experience of designer.

In this paper, a deep spatial-temporal convolution neural network based on deep learning theory is proposed to detect point targets intelligently. To detect targets from any image size, the fully convolution is adopted. Thus, the proposed network and be trained by small-size images, and tested by large-size images without any modification. In order to meet the requirement of real time, the 3D convolution in the proposed method is factorized into 2D spatial convolution and 1D temporal convolution, fewer parameters are needed and the computing burden is greatly decreased. Additional, the specified loss function is introduced to take the number imbalance and error imbalance into account simultaneously. Simulation results demonstrate that the proposed approach can robustly and effectively suppress background clutters and detect weak point targets in infrared sequences.

The rest of the paper is organized as follows: Section 1 reviews the related work about point target detection and deep learning based methods. Section 2 shows the deep spatial-temporal convolution neural network. The network architecture, factorized 3D convolution, fully convolution and novel loss function are introduced in detail. The overall performance of the proposed method and comparison results with other methods are presented in Section 3. Finally, conclusions are drawn in Section 4.

## 1 Related Work

Benefiting from the enhancement of computer, deep learning techniques have been recently used in object detection, visual recognition, and time series classi fi cation. Especially, the deep convolutional neural network (CNN) achieved impressive results at 2015 ImageNet contest. After that, many deep learning based methods are proposed in the field of target location and identification, such as R-CNN [15], Fast R-CNN[16], Faster R-CNN [17], Mask-RCNN [18], and YOLO [19-20] and. It has been proved by many researchers that deep learning based methods can automatically excavate more deep and obscure features from the raw images directly from a mass of training images, which are more beneficial to discriminate different objects than hand-crafted features.

Although many deep learning based methods are studied in recent years, the targets involved in the above methods are large targets, called as area targets, such as human face, vehicle, and animal. The features of area targets are distinctly different from point targets in IRST. In fact, the area target may extend to hundreds of pixels with abundant texture information, geometry information, and color information, which can provide plenty of details for processing. However, the point target is extremely not obvious in shape, size and color characteristics, because they are generated from point source at long distance, and only gray information is obtained by infrared sensor. As a result, the existing detection methods based on deep learning for area targets are not suitable for point target detection in IRST.

Some methods inspired by deep neural networks to detect point targets have been proposed recently. In brief, these methods can be divided into three categories.

(1) Some methods simply convert the detection problem into pattern recognition problem. Ming LIU used traditional 2D CNN networks to judge whether there are infrared small targets in infrared patches [21]. The input is fixed-sized in these methods. The network cannot be self-adaptive for arbitrary size, because of the fully connected layers. Essentially, these approaches can be regarded as fixed patch-wised recognition methods.

(2) Many researchers try to combine traditional predetection and deep learning based recognition. At May

2018, DDP-CNN was proposed based on data-driven proposal and deep learning-based classification by Ryu [1], including region proposal step and classification step. The region proposal step is realized based hand-crafted features (mean subtract filter) to find suspected target regions, and the classification step is finished by deep learning (consisting of two convolution layers, one max pooling and two fully connected layers). So, the DDP-CNN can be regarded as partial intelligent recognition method. Besides, the pooling layers lead to low positioning accuracy.

(3) 2D CNN networks based on spatial features are used to detect small targets by some researchers. Lin's method is designed to detect infrared targets in oversampling images [22]. The oversampling sensors are very different from that of usual sensors. As a result, the target size in oversampling system may be increased several times as much, and they are considered as extended target. Larger size leads to more texture features. Besides, only spatial features are not enough for the point target detection. Thus, Lin's method does not well deal with point target detection.

This work will focus on detecting point targets with high performance and less runtime by deep spatial-temporal convolution network.

## 2 Proposed method

### 2.1 Network architecture

For some reasons, exiting CNN networks cannot be directly used to detect point targets. Firstly, the traditional CNN networks are used to detect large area targets, and accurate spatial coordinates are thrown away due to pooling layers and fully connected layers. However, the targets in IRST are point targets, which must be located precisely by the pixel or sub-pixel. Secondly, for point target detection, each pixel may present a small target, and infrared image should be processed pixel by pixel. Thus, batch detection method is wanted. Thirdly, the input of traditional networks usually is fixed-size because of fully connected layers, which limits the flexibility in practical application. Thus, the special network must be designed based on the characters of point target.

Fortunately, we have found some characters of point target, and they can make it possible to overcome the above shortcomings. For example, the main features of point target can be obtained by statistical analysis based on its small neighboring region. Thus, smaller receptive filed (RF) of network is enough for point targets in comparison with area targets. As a result, one hand, fewer stacked layers in network are needed, which can reduce the complexity and computation. On the other hand, pooling layers and fully connected layers, used to enlarge receptive filed by image compression and feature integration, can be given up.

For the above reasons, the proposed network is hierarchically constructed by stacking different convolutional layers. The network architecture is shown in Fig. 1.

The proposed point target detection network takes video clip as input, and produces a residual image with the same size of the input. The residual image represents the estimated point target intensity after background suppression. The final target index can be obtained by threshold segmentation easily.

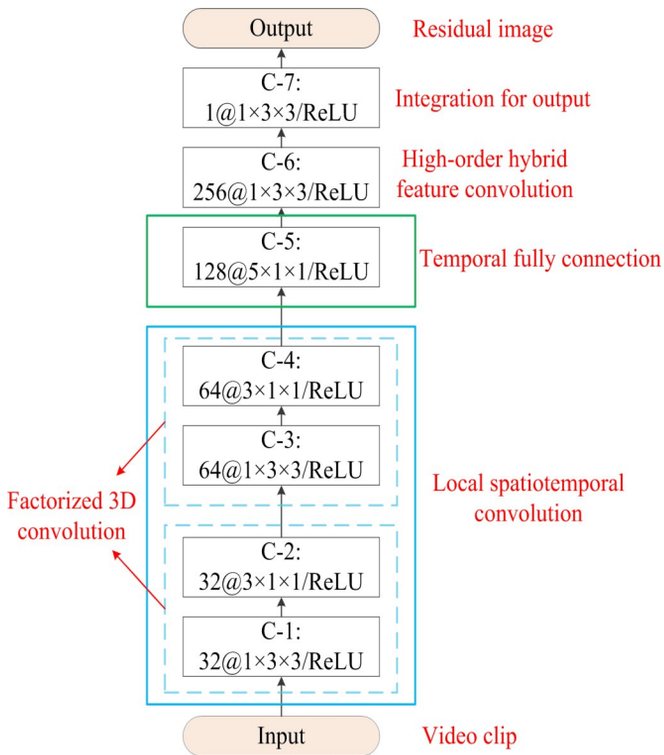


Fig. 1 The proposed network architecture.
图1 本文所提出网络架构

In Fig. 1, "C-n" indicates the index of convolution layer. The number of feature maps is denoted by the number before @, and the size of kernel is represented by the number after @. For example, the kernel with 1 × 3 × 3 represents the depth, height and width are 1 pixel, 3 pixels and 3 pixels, respectively.

This method consists of three parts. Firstly, the bottom of this network (C-1 to C-4) is a stack of 3D convolutional layers, which are focus on low-order spatiotemporal features. To improve their efficiency, the factorized 3D convolution is adopted, as introduced in Section 2.2. In second part, the 3D spatiotemporal feature maps generated from video clip are compressed into one 2D hybrid feature map. This operation is carried out by the convolution over the whole video clip in time dimension. In third part, the high-order hybrid features are intensively learned by more convolution kernels. Finally, feature fusion across different channels is achieved by 1 × 1 convolution in channel dimension, which makes sure that the output size is same as that of input. The 1 × 1 convolution is equivalent to cross-channel parametric pooling layer, and allows complex interactions of cross channel information.

The process is modeled with a fully convolutional network. Convolutional layer is architecture with shared parameters, so all pixels can be processed by the same operation. The feature maps, input and output of each

convolutional layer, can be modeled as feature results with size $d \times c \times h \times w$, where $d$, $c$, $h$ and $w$ are depth, the number of channels, height, width, respectively. For the first convolutional layer, the input is the infrared video clip, and the size is $h \times w$, the length of video clip is $d$ ($d$ is set to 5 in this paper), the number of channel is 1 (because the output of infrared sensor is gray image). The output feature map indicates a particular feature representation extracted based on all locations on the input, which can be obtained via convolving the input feature map with a trainable kernel and adding a trainable bias parameter. In this work, the input feature map is denoted as $X$. The weights and bias of convolution kernel are represented by $W$ and $b$. Thus, the output feature map can be computed by [23]

$$f_s(X; W, b) = W *_s X + b \qquad ,(1)$$

where $*_s$ denotes the convolution operation with stride $s$ ($s$ is 1 in this work). Feature representation ability can be enhanced by point-wise nonlinearity operation following with each convolutional layer, and ReLU is adopted in this method. Unlike traditional methods, non-linear down-sampling operation (e. g. , max pooling) is thrown away, because pixel-wise prediction is very important for point target detection.

## 2. 2　Factorized 3D convolution

Point target can be detected by 2D convolutional neural networks under smooth background. However, these methods do not provide robust detection in complex background, especially when the clutters are strong as shown in Section 3. 2. The reason is that only spatial information is not enough to discriminate true or false targets. Thus, both spatial and temporal information must be fully utilized.

2D convolution is performed only spatially, and temporal information of the input is lost. 3D convolution is done spatiotemporally, and both spatial information and temporal information of the input are preserved. Thus, 3D convolution is well-suited for spatiotemporal feature learning, and it is adopted in the proposed method.

Traditional 3D convolution can be regarded as that 2D spatial convolution and 1D depth projections are performed simultaneously. Thus, the cost of computational complexity is exceptionally high, even higher than the peak of common computers. Although many studies have proven that deep 3D convolutional neural networks can obtain spatiotemporal features even better than human level accuracy, it is beyond the applicable level. To solve the problem of real time and limit memory space in applicable application, the factorized 3D is adopted, which unravels spatial and temporal convolutions apart[24, 25]. It means that 2D spatial convolution layer and 1D temporal convolution layer are sequentially carried out as shown in Fig. 2. In this figure, $(k, i, j)$ represents the pixel at $k$ frame with the coordination$(i, j)$. Acceptable accuracy with significantly less computation can be achieved by stacking them together.

As shown in Fig. 2, the factorized 3D convolution is equivalent to 3D convolution. The factorization can be

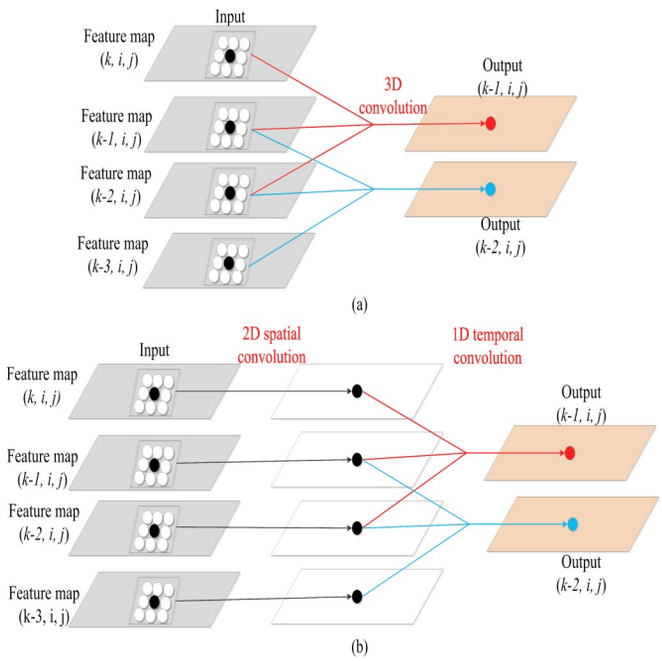


Fig. 2　The sketches of 3D covolution and factorized 3D convolution: (a) 3D covolution; (b) factorized 3D convolution.
图2　3D卷积和分解3D卷积示意图:(a) 3D卷积;(b) 3D卷积分解

modeled by

$$Conv_{3D} \approx Conv_{2D} \otimes Conv_{1D} \qquad ,(2)$$

where $\otimes$represents the Keonecker product. $Conv_{3D}$ is the 3D convolution. $Conv_{2D}$ and $Conv_{1D}$ denote the 2D convolution (spatial convolution) and 1D convolution (temporal convolution), respectively.

The computation comparison of 3D convolution and factorized 3D convolution is listed as Table 1. The number of parameters of traditional 3D convolution with 1024 3D kernels is about 28. 67K, and the requirement of computation is about 0. 06GFlops. It can be approximated by 32 2D kernels and 32 1D kernels. As a result, only 9. 38K parameters and 0. 03GFlops are required by factorized 3D convolution. So, the factorization can reduce the computation burden by 50%.

**Table 1　Computation comparison of different convolutions.**
**表1　不同卷积的计算量比较**

| Item | Flops (G) | Parameters (K) |
|---|---|---|
| 3D Conv. | 0. 06 | 28. 67 |
| Factorized 3D Conv. | 0. 03 | 9. 38 |

## 2. 3　Fully convolution

Pooling layer and fully connected layer take important role in traditional deep-learning methods for area target detection.

Typical pooling operations include average pooling and max pooling. They can be considered as non-linear down-sampling. For example, the size of feature map is reduced to a quarter of original after 2 × 2 pooling layer.

Obviously, the compressed feature maps are coarse and reduced-resolution, and lots of detailed information has been lost [27]. Even so, such coarse features (including color, shape, texture, and so on) are enough to locate and recognize area targets. For point target, single pixel may represent one important target, so any information lost can cause unexpected consequence. Thus, the whole feature maps should be fully and intensively analyzed, and pooling layer should not appear in this network.

Furthermore, the fully connected layer brings multiple local features from different regions together, but it limits the input size. It means that the size (height and width) of train image must be same as the size of test image.

To overcome those shortages, the proposed network only consists of many convolution layers, called fully convolution method. The 2D convolution layers with $1 \times H \times W$ kernel are used to extract spatial features, while the 1D convolution layers with $D \times 1 \times 1$ kernel are adopted for feature extraction in time domain. Different convolution layers are connected in series, and it makes the features from both spatial convolution and temporal convolution to be deeply integrated. By choosing slip step=1, the intensity generated from point target is estimated pixel by pixel. So, pixel-wise processing is carried out by the proposed fully convolution network.

Additionally, it has been demonstrated that small receptive fields of $3 \times 3$ convolution kernels can better learn complex features with deeper architectures [26]. Consequently, the spatial receptive field is set to $3 \times 3$ in the proposed network, while the temporal depth of the 3D convolution kernels are adjusted as needed.

### 2.4 Multi-weighted loss function

The multi-weighted L1 norm loss function is proposed in this work, and can be expressed as

$$\ell = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} w_{(i,j)}^{NumBal} \cdot w_{(i,j)}^{IntBal} \cdot \left| y_{(i,j)} - y'_{(i,j)} \right| \quad ,(3)$$

where $N$ is the number of training samples, $M$ is the number of pixels in training sample, $y_{(i,j)}$ is the ground truth for the $j$th pixel in $i$th training sample. $y'_{(i,j)}$ is the output of the proposed network. $w_{(i,j)}^{NumBal}$ is the number-balance weighting parameter, and $w_{(i,j)}^{IntBal}$ is the intensity-balance weighting parameter. The different weighting parameters are jointly used to trade off the false alarms and missing alarms. The loss function is minimized during the training, and it indicates that the predicted target intensity gradually reaches the truth.

The sample imbalance encountered in training of point target detection may bring extreme error. The imbalance between target samples and background samples can overwhelm training and lead to degenerate model, especially when background samples are far more than target samples. In fact, the background samples generally belong to majority class, while the target samples are in the minority. In practical application, the background images can be easy and often obtained, however the true targets are relatively rare. If the sample imbalance cannot be solved, the training is inefficient. For example, the extreme sample imbalance may lead to true target can be completely ignored. In order to alleviate the bias in performance caused by imbalanced sample distribution, the number-balance weight $w_{(i,j)}^{NumBal}$ is assigned to each sample to weaken the relative impact of background samples, while strengthen the impact of target samples. $w_{(i,j)}^{NumBal}$ can be calculated by

$$w_{(i,j)}^{NumBal} \propto \frac{N_i}{N^{Background}} \quad (4)$$

$$N_i = \begin{cases} N^{Background} & \text{if } i^{th} \text{ sample is background} \\ N^{True} & \text{if } i^{th} \text{ sample is true target} \end{cases} \quad (5)$$

where $N^{Background}$ and $N^{True}$ are the number of background samples and the number of target samples, respectively. The training sample includes the background samples and the target samples, i. e. $N = N^{True} + N^{Background}$.

For point target, the detection result can be obtained by threshold segmentation based on the output after background suppression. Although the error sum may be same for different predicted results, the detection results are obviously different as shown in Fig. 3. In this figure, each rectangle represents one pixel, and indicated by $(i, v)$, where $i$ is the index of pixel and $v$ is the predicted value. The ground truth is shown in Fig. 3 (a), and all pixels belong to background. Two possible error distributions are shown in Fig. 3 (b) and (c). Though the error sums of both (b) and (c) are 0.9, the result of (c) may bring more false alarms, because the 5th pixel can extremely likely over the threshold during segmentation. Thus, different weighting parameters should be adaptively given to different error distributions, called intensity-balance weight.

In this work, the intensity-balance weight is represented by $w_{(i,j)}^{IntBal}$, and can be calculated by

$$w_{(i,j)}^{IntBal} = \begin{cases} 1, & if \left| y_{(i,j)} - y'_{(i,j)} \right| \leqslant d_{th} \\ \dfrac{\left| y_{(i,j)} - y'_{(i,j)} \right|}{d_{th}}, & if \left| y_{(i,j)} - y'_{(i,j)} \right| > d_{th} \end{cases} \quad (6)$$

The calculation of $w_{(i,j)}^{IntBal}$ can be shown as Fig. 4.

As described in Fig. 4, when the original error is greater than $d_{th}$, the weighting parameter of the pixel is assigned with larger value such that the network is trained with less false alarms. $d_{th}$ is set based on the prior segmentation threshold.

## 3 Simulation results

### 3.1 Experiment scheme and evaluation metrics

In this section, three experiments are performed to evaluate the performance of the proposed method. A large amount of infrared samples are generated based on point target model and real background images shown as Fig. 5, and the sequences are 5 frames long.

In this paper, the weak target is regarded as point target, because of long-range observation. The point tar-

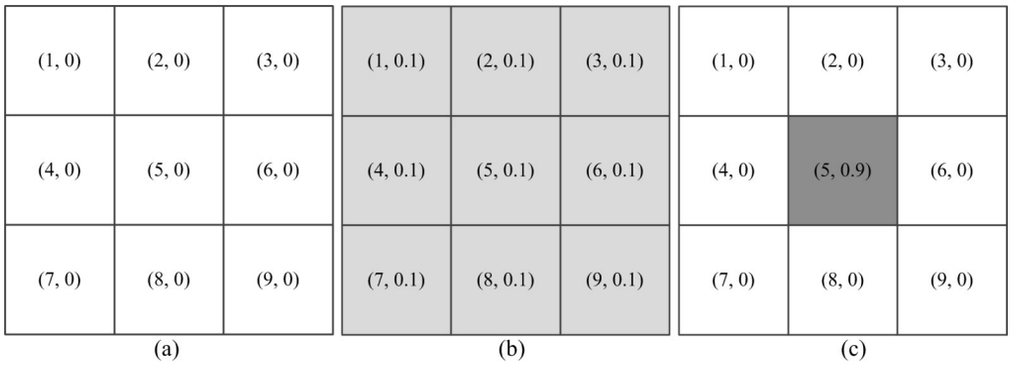


Fig. 3　The example of different error distributions: (a) the ground truth; (b) 1th predicted result with uniform error; (c) 2th predicted result with concentrated error.
图3　不同误差分布示例:(a)真值;(b) 误差均匀分布的预测结果;(c)误差聚集的预测结果

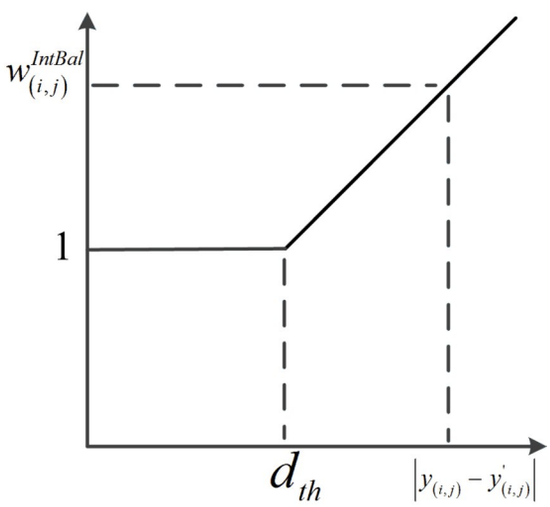


Fig. 4　The function of intensity weighting parameter.
图4　能量权重因子函数

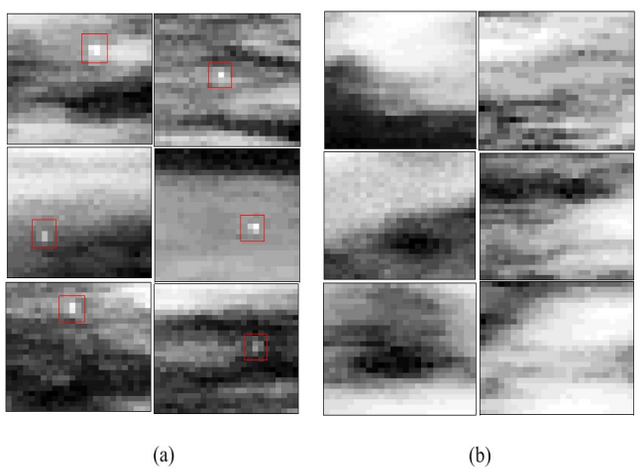


Fig. 5　The example of samples: (a) the target samples; (b) the background samples.
图5　样本示例:(a)目标样本,(b)背景样本

gets are generated by 2D Gaussian function as following [3, 14].

$$T(x,y) = I_p \exp\left\{ -\frac{1}{2}\left[ \frac{(x - x_c)^2}{\delta_x^2} + \frac{(y - y_c)^2}{\delta_y^2} \right] \right\}, \quad (7)$$

where $T(x, y)$ is target projection on image planar, $x$ and $y$ represent the spatial coordinates, $(x_c, y_c)$ represents the target center position, $I_p$denotes the peak intensity, $\delta_x^2$ and $\delta_y^2$ are the variance in row and column direction.

Furthermore, the observation of optical image embedded with dim point target can be obtained as following [3,14].

$$F(x,y) = T(x,y) + B(x,y) + C(x,y) \quad , \quad (8)$$

where $F(x, y)$ represents the output image, $B(x, y)$ denotes cloud background, $C(x, y)$ is noise.

For the background samples, the ground truth is a fully zero-value image. If there are target pixels and background pixels at the same time in the training sample, the ground truth only contains the gray values of target pixels, and the others are set to zeros. The preprocessing before entering the network is necessary for all training and testing samples. In this work, the preprocessing is carried out by de-averaging and normalization. It should be noted that the same average is used for all samples, because the sequences are obtained by same sensor. As a result, the absolute intensity of point targets can be better preserves to support distinction. Besides, $d_{th}$ = 0.2 in these experiments.

In the first experiment, the point targets are detected by the proposed method, Lin's method [22], Max-Mean filter [11], TopHat filter [10] and Spatial-Temporal Accumulative Difference method (STDA) [12], respectively. The proposed method and Lin's method belong to deep learning based solution, and the others are traditional methods based on hand-crafted filters. Besides, the STDA is a classical method based on the spatial-temporal fusion. The experiment is performed to validate the detection performance of the proposed method compared with existing methods. It illustrates the proposed method can significantly improve point target detection performance in heavy clutters.

In the second experiment, the availability of the proposed method with different input size is proved. It dem-

onstrates the advantage inheriting from fully convolution.

In the third experiment, the detection performance of the proposed method is evaluated under different conditions including original signal-to-clutter ratio (SCR) and jitter of sensor.

In order to measure the performance, the following evaluation metrics are introduced.

To measure the ability of removing background, the background suppression factor (BSF) is introduced[30]. It can be computed by

$$BSF = \frac{\sigma_{in}}{\sigma_{out}}, \tag{9}$$

where $\sigma_{in}$ represents the standard deviation of original image, and $\sigma_{out}$ denotes the one of output image, respectively.

The quality of image can also be indicated by SCR. It is defined as

$$SCR = \frac{\mu_t}{\sigma_b}, \tag{10}$$

where $\mu_t$ is the target intensity without background, and $\sigma_b$ represents the standard deviation of local background region[22].

The comprehensive detection result is evaluated by Receiver Operating Characteristic (ROC). The ROC curve can describe the detection result by a function with the probability of detection ($p_d$) and the probability of false alarms ($p_f$)[22, 29, 30]. $p_d$ is defined by

$$p_d = \frac{N_d}{N_t}, \tag{11}$$

where $N_d$ denotes the number of detection reports from true point targets, and $N_t$ denotes the number of true targets. $p_f$ is defined by

$$P_f = \frac{N_f}{N_i}, \tag{12}$$

where $N_f$ represents the number of detection reports from false alarms, and $N_i$ is the number of pixels of all testing images.

The simulation environment in this work is shown in the Table 2.

**Table 2 The simulation environment.**
**表2 仿真环境参数**

| Item | Parameter |
|---|---|
| CPU | Intel i7, 2.8GHz×12 |
| GPU | Nvidia-1080Ti |
| RAM | 64GB |
| System | Ubuntu 18.04 |
| Disk | 2TB |
| Software | Pytorch 1.1 |
| Language | Python 3.6 |

### 3.2 Comparison with other methods

In this simulation, 10000 training samples are obtained based on 10 real background sequences. Meanwhile, 10000 testing samples are obtained from another 10 real background sequences. The point taregets in samples are randomly added based on point target model. Mean SCR of original image is about 6. The jitter of sensor is simulated as random Gaussian distribution with $\sigma$ = 0.2. The size of infrared image is 25 × 25 pixels,

Fig. 6 and Fig. 7 show the results of different methods for background images. In these figures, (a) is the original images, and (b) shows the result of the proposed method. Subsequently, (c)-(f) are the results of Lin's method, Max-Mean, TopHat and STDA, respectively. The background in Fig. 6 (a) is smooth, thus these methods can obtain good results. There are many clutter signals in Fig. 7 (a), and they cannot be suppressed by traditional spatial methods (see Fig. 7 (c)-(e), where the clutter signals after processing are even over 2). As shown in Fig. 7 (f), benefiting from the fusion of spatial information and temporal information, the STDA achieves better result than traditional spatial methods. However, the max value of the output of the proposed method is less than 0.05, and the clutter signals can be rejected more easily.

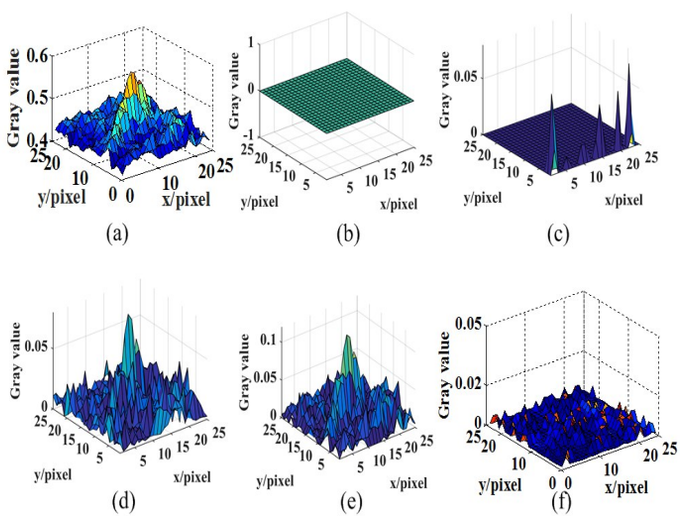


Fig. 6 The original image and results of different methods for 1th Background.: (a) the original input; (b) the result of our method; (c) the result of Lin's method; (d) the result of Max-Mean; (e) the result of TopHat; (f) the result of STDA.
图6 在第1组背景中不同原始图像和不同方法的处理结果: (a)原始输入; (b)本文处理结果; (c)Lin方法处理结果; (d) Max-Mean处理结果; (e)TopHat处理结果; (f) STDA处理结果

Fig. 8 and Fig. 9 show the results for target samples, in which the targets are marked by white circles. The target in Fig. 8 is stronger than background in the original image, and the target can be detected by the above methods. However, the target in Fig. 9 is far weaker than many clutters in the original image (the target is about 1, and some clutters are about 3). The weak point target cannot be found in the results of Max-Mean and TopHat. Those traditional spatial methods fail to catch the point target. The Lin's method and STDA can well

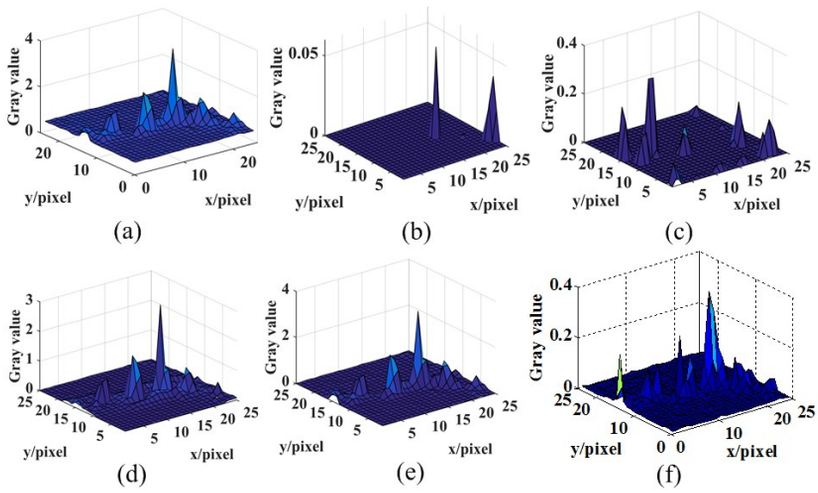


Fig. 7　The original image and results of different methods for 2th Background. : (a) the original input; (b) the result of our method; (c) the result of Lin's method; (d) the result of Max-Mean; (e) the result of TopHat; (f) the result of STDA.
图 7　在第 2 组背景中不同原始图像和不同方法的处理结果: (a)原始输入; (b)本文处理结果; (c)Lin 方法处理结果; (d)Max-Mean 处理结果; (e)TopHat 处理结果; (f) STDA 处理结果

extract target intensity, but many clutters are still kept. So, the performance of Lin' method and STDA greatly degrade under complex background. However, not only the weak target can be detected, but also the clutters can be well suppressed by the proposed method as shown in Fig. 9 (b). The result of our method has fewer clutters compared to the other methods, which is important to keep lower false-alarm rates under the same probability of detection.

The original image and standard deviation in the time domain of Target 2 are shown in Fig. 10. In fact, Target 2 is very weak and hard to be detected just based on spatial feature or temporal feature. The proposed method can extract spatial-temporal feature to suppress background, and the fusion of spatial-temporal feature is automatically achieved.

In order to intuitively show the background suppression performance, the comparisons of SCR and BSF for two point targets are listed in Table 3 and Table 4. The SCR is computed based on the output images of different methods. It is proved that the proposed method can enhance the ability of background suppression, which is very import for detection by threshold segmentation. More specifically, the mean SCRs of Max-Mean, TopHat, Lin's method, STDA and our method are 12.1859, 11.5509, 14.3741, 18.2125 and 19.7507, respectively. The BSFs of those methods are 1.2498, 1.0821, 2.6755, 3.4656 and 4.2671, respectively. It is clear that the proposed method can obtain the best background suppression.

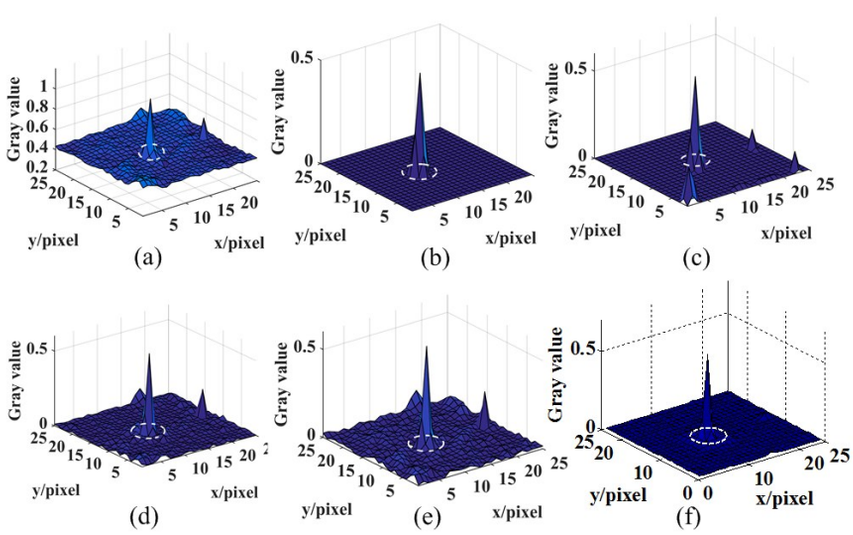


Fig. 8　The original image and results of different methods for Target 1: (a) the original input; (b) the result of our method; (c) the result of Lin's method; (d) the result of Max-Mean; (e) the result of TopHat; (f) the result of STDA.
图 8　目标 1 的原始图像和不同方法处理结果: (a)原始输入; (b)本文处理结果; (c)Lin 方法处理结果; (d)Max-Mean 处理结果; (e)TopHat 处理结果; (f) STDA 处理结果

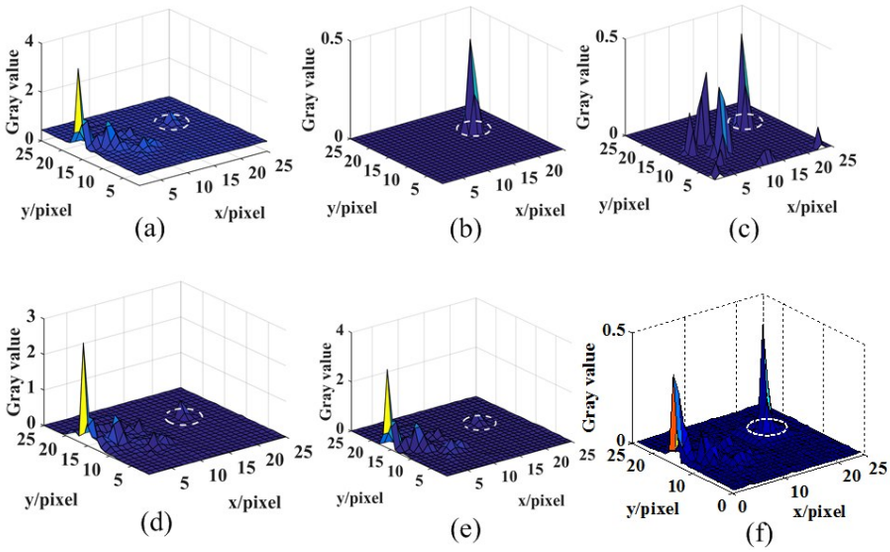


Fig. 9 The original image and results of different methods for Target 2: (a) the original input; (b)the result of our method; (c) the result of Lin's method; (d) the result of Max-Mean; (e) the result of TopHat; (f) the result of STDA.
图9 目标2的原始图像和不同方法处理结果:(a)原始输入;(b)本文处理结果;(c)Lin方法处理结果;(d)Max-Mean处理结果;(e)TopHat处理结果;(f) STDA处理结果

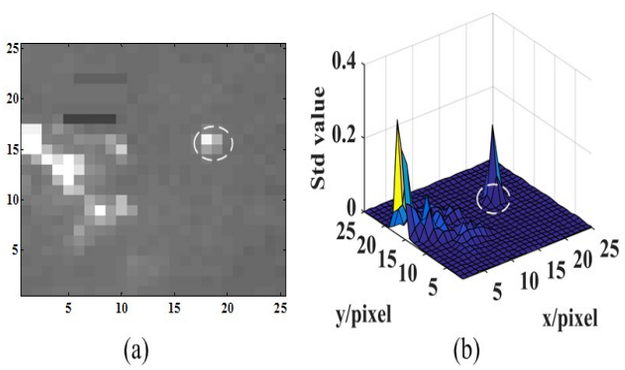


Fig. 10 The display of Target 2: (a) the original gray image; (b) the standard deviation in the time domain.
图 10 目标2展示:(a)原始灰度图像;(b)时域标准差分布

**Table 3 Background suppression comparison by SCR in output.**
**表3 不同SCR条件下的背景抑制比较**

| | Proposed | Lin's | Max-Mean | TopHat | STDA |
|---|---|---|---|---|---|
| Target 1 | 20. 4061 | 16. 8253 | 20. 9308 | 20. 0394 | 19. . 8378 |
| Target 2 | 19. 0953 | 11. 9228 | 3. 4410 | 3. 0623 | 16. 5872 |
| Mean | 19. 7507 | 14. 3741 | 12. 1859 | 11. 5509 | 18. 2125 |

**Table 4 Background suppression comparison by BSF.**
**表4 背景抑制能力的BSF比较**

| | Proposed | Lin's | Max-Mean | TopHat | STAD |
|---|---|---|---|---|---|
| Target 1 | 1. 3527 | 1. 0293 | 1. 2477 | 1. 1210 | 1. 3375 |
| Target 2 | 7. 1815 | 4. 3218 | 1. 2520 | 1. 0431 | 5. 5936 |
| Mean | 4. 2671 | 2. 6755 | 1. 2498 | 1. 0821 | 3. 4656 |

The ROC curves of different methods are depicted in Fig. 11. The methods based on spatial-temporal fusion (including the proposed method and STDA method) demonstrate better performance than the spatial methods (including Max-Mean, Tophat and Lin's method). The proposed method shows the best detection performance with low false alarm probability and high detection probability.

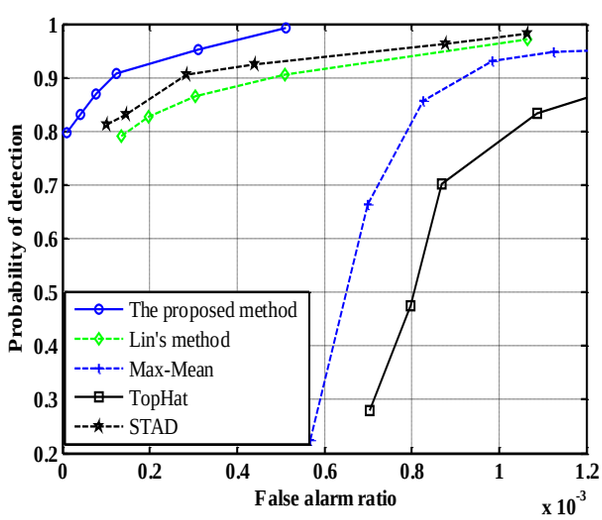


Fig. 11 The ROC curves of different methods.
图11 不同方法的ROC曲线

For IRST application, the high real time is required. The comparison of average runtime is listed in Table 5. More time is needed by the proposed method than Lin's method, because infrared sequences are put

into the network to extract the spatial-temporal features in this work. However, the proposed method is still faster than Max-Mean, TopHat and STDA, and can meet the requirement of real time.

**Table 5　Average runtime comparison.**
**表5　平均计算时间比较**

| | Proposed | Lin's | Max-Mean | TopHat | STDA |
|---|---|---|---|---|---|
| Average runtime(s)/sample | $5.84\times10^{-4}$ | $3.38\times10^{-4}$ | $3.30\times10^{-3}$ | $1.33\times10^{-2}$ | $4.51\times10^{-3}$ |

### 3.3　Evaluation with different input size

In the above experiment, the input size is set to 25 × 25pixels for convenience. But, it is not fixed. Benefiting from fully convolution, each pixel can be processed using the same parameters at the same time. As a result, though the proposed network is trained with the input size25 × 25, input of arbitrary size can be put into the network and correspondingly-sized output can be obtained at once. On the contrary, the patch-wise method is fixed.

The image of Fig. 9 (a) is extended to 35 × 35 pixels and 45 × 45 pixels as shown in Fig. 12 (a) and Fig. 12 (c), respectively. The position of target is changed to (25, 23) and (34, 28) from (16, 18). Meanwhile, the results are shown in Fig. 12 (b) and Fig. 12 (d), respectively. Comparison shows that results in three cases are in agreement with each other. So, under sufficient resources, the proposed network is applicable to images with different sizes. The important advantages of the proposed method are shift-invariant and position invariant.

### 3.4　Comparison under different conditions

In order to illustrate the detection performance of the proposed method under different conditions, the infrared sequences with different jitters and original SCRs are tested.

In Fig. 13, the targets are added based on mean original SCR 6, the ROC curves under different jitters are shown. The standard deviations of inter-frame jitters are set to 0.1, 0.2, and 0.5, respectively. Along with the increase of jitter, the probability of detection decreases gradually. The probability of detection changes from 98% to 77%, when the false alarm ratio is $10^{-4}$. Thus, it is very important for the proposed method that keeping sensor motionless.

The detection performances of the proposed method for point targets with different original SCRs are analyzed, and jitter is fixed to 0.2. Fig. 14 shows ROC curves under different SCRs, which are set to 4, 6 and 8, respectively. Though some targets may be missed, the proposed method can achieve fairly high probability of detection. For example, the probability of detection can reach about 90% at the false alarm ratio is $10^{-4}$, when the mean original SCR is 6. It should be noted that the SCRs of some targets in this test sets are lower than mean SCR.

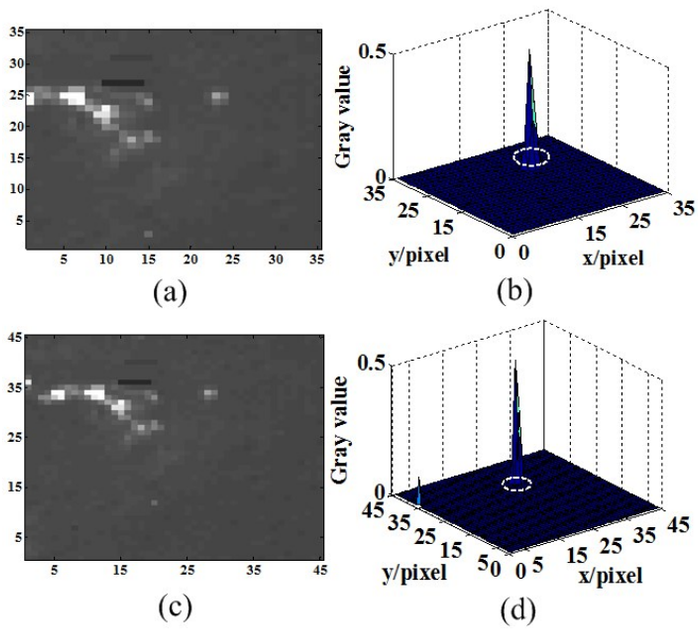


Fig. 12　The result of different input size: (a) the input image with 35×35 pixels; (b)the result of image with35×35 pixels; (c) the input image with 45×45 pixels; (b) the result of image with45×45 pixels.
图12　不同尺度输入测试结果:(a)输入35×35图像;(b)35×35图像对应的输出;(c)输入45×45图像;(d)45×45图像对应的输出

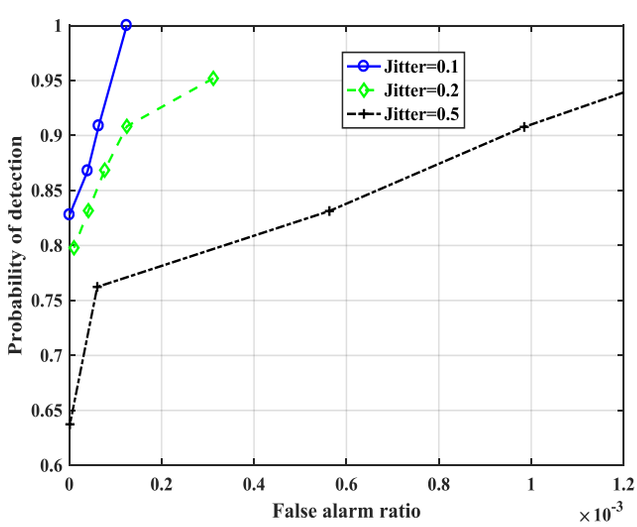


Fig. 13　The ROC curves with different jitters.
图13　不同抖动条件下的ROC曲线

## 4　Conclusions

In IRST, point detection is still a great challenge for some reasons. Traditional methods can't robustly and intelligently detect point targets in complex background. In this work, a deep spatial-temporal convolution neural network is proposed to address this problem. The network is built based on fully convolution without pooling layer and fully connected layer, factorized 3D convolution and multi-weighted loss function are adopted to en-

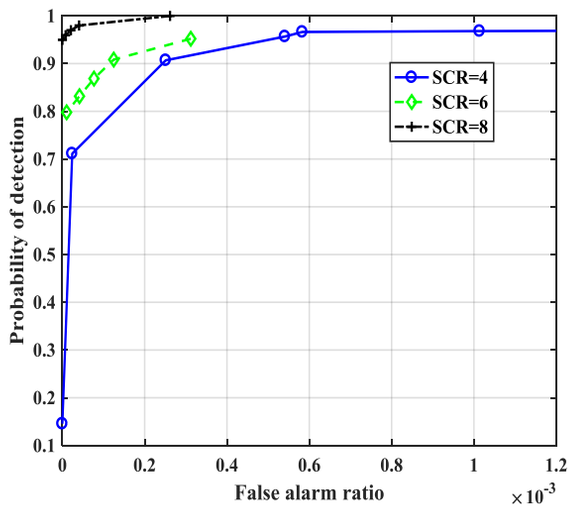


Fig. 14 The ROC curves with different mean original SCRs.
图14 不同均值SCR下的ROC曲线

hance the performance. The proposed method is compared to other four methods, including traditional methods (e. g. , Max-Mean filter TopHat filter and Spatial-Temporal Accumulative Difference method) and deep learning based method (e. g. , Lin's method). The detection performance is evaluated by different metrics, such as signal-to-clutter ratio, background suppression factor. Meanwhile, ROC curves are drew to confirm the robustness of the proposed approach. Additionally, the comparison under different conditions is carried out for the proposed method, and the affections of original SCR and sensor's jitter are demonstrated in detail. Consequently, the deep spatial-temporal convolution neural network can effectively detect point targets using less runtime.

## References


[1] Junhwan R, Sungho K. Small infrared target detection by data-driven proposal and deep learning-based classification [C]. Proc. of SPIE on infrared technology and application, **2018**: 10621J-10624J.

[2] Sui X, Chen Q, Bai L. Detection algorithm of targets for infrared search system based on area infrared focal plane array under complicated background [J]. *Optik*. 2012, **123**: 235-239.

[3] Zhao J, Tang Z, Yang J, *et al*. Infrared small target detection using sparse representation [J]. *Journal of systems engineering and electronics*. 2011, **22**(6): 897-904.

[4] Li M, Lin Z, Long Y, *et al*. Joint detection and tracking of size-varying infrared targets based on block-wise sparse decomposition [J]. Infrared Physics and Technology. 2016, **76**: 131-138.

[5] Gao J, Lin Z, An W. Infrared small target detection using a temporal variance and spatial patch contrast filter [J]. *IEEE Access*. 2019, **7**: 32217-32226.

[6] Zhang W, Cong M, Wang L. Algorithms for optical weak small targets detection and tracking: Review [C]. International conference on neural networks and signal, 2003: 643-647.

[7] Chen C, Li H, Wei Y, Xia T, *et al*. A local contrast method for small infrared target detection [J]. *IEEE transactions on geoscience and remote sensing*. 2014, **52**: 574-581.

[8] Warren R. Detection of distant airborne targets in cluttered backgrounds in infrared image sequences [D]. University of south australia, 2002.

[9] Barnett J. Statistical analysis of median subtraction filtering with application to point target detection in infrared backgrounds [C], Proc. of SPIE on infrared system and components III, 1989: 10-18.

[10] Tom V, Peli T, Leung M, Bondaryk J. Morphology-based algorithm for point target detection in infrared backgrounds [C]. Proc. of SPIE on signal and data processing of small targets, 1993: 2-11.

[11] Deshpande S D, Meng H E, Venkateswarlu R, *et al*. Max-mean and max-median filters for detection of small targets [C]. Proc. of SPIE on international symposium on optical science, engineering, and instrumentation, 1999: 74-83.

[12] Yu Q, Huang S, Zhao W, *et al*. A fusion detection algorithm of small infrared target based on spatial-temporal accumulative difference [J].(Journal of projectiles, rockets, missiles and guidance 于强, 黄树彩, 赵炜, 等. 一种基于时空域累积差分的红外小目标融合检测算法. **弹箭与制导学报**)), 2014, **34**(6): 181-189.

[13] Schmidt W. Modified matched filter for cloud clutter suppression [J]. *IEEE transactions on pattern analysis and machine intelligence*. 1990, **12**: 594-600.

[14] Gao C, Meng, D, Yang Y, *et al*. Infrared patch-image model for small target detection in a single image [J]. *IEEE transactions on image processing*. 2013, **22**(12): 5996-5009.

[15] Girshick R, Donahue J, Darrell T, Region-based convolutional networks for accurate object detection and segmentation [J]. *IEEE Transactions on pattern analysis and machine intelligence*. 2016, **38**: 142-158.

[16] Girshick R. Fast R-CNN [C]. Proc. of IEEE on computer vision, 2015: 1440-1448.

[17] Ren S, He K, Girshick R, *et al*. Faster RCNN: Towards real-time object detection with region proposal networks [C]. Advances in neural information processing systems, 2015: 91-99.

[18] He K, Gkioxari G, Dollar P. Mask R-CNN [C]. Proc. of IEEE on computer vision, 2017: 2980-2988.

[19] Redmon J, Divvala S, Gishick R. You only look once: Unified, real-time object detection [C]. Proc. of IEEE on computer vision and pattern recognition, 2016: 779-788.

[20] Joseph R, Farhadi A. YOLO3: An incremental improvement [J]. *ArXiv preprint*. 2018: 1-6.

[21] Liu M, Du H, Zhao Y, *et al*. Image small target detection based on deep learning with SNR controlled sample generation [C], Proc. of CSMA, 2017: 211-220.

[22] Lin L, Wang S, Tang Z. Using deep learning to detect small targets in infrared oversampling images [J], *Journal of Systems Engineering and Electronics*. 2018, **5**: 947-952.

[23] Ian G, Yoshua B, Aaron C. Deep learning [M]. MIT Press, 2017.

[24] Wang M, Liu B, Hassan F. Factorized convolutional neural networks [C]. Proc. of IEEE on computer vision workshop, 2017: 1-10.

[25] Sun L, Jia K, Yeung D, *et al*. Human action recognition using factorized spatio-temporal convolutional networks [C], Proc. of IEEE on computer vision, 2015: 4597-4605.

[26] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. *ArXiv preprint*. 2014: 1409-1556.

[27] Long J, Evan S, Trevor D. Fully convolutional networks for semantic segmentation [C]. Proc. of IEEE on computer vision and pattern recognition, 2015: 1-10.

[28] Kim S. Double layered-background removal filter for detecting small infrared targets in heterogenous backgrounds [J]. *Journal of infrared, millimeter and terahertz waves*. 2011, **32**: 79-101.

[29] Guo C, Deyu M, Yi Y, *et al*. Infrared patch-image model for small target detection in a single image [J]. *IEEE transactions on image processing*. 2013, **22**: 4996-5009.

[30] He Y, Li M, Zhang J, *et al*. Small infrared target detection based on low-rank and sparse representation [J]. *Infrared physics & technology*. 2015, **68**: 98-109.