IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Dim and Small Target Detection in Multi-Frame Sequence Using Bi-Conv-LSTM and 3D-Conv Structure

Xin Liu [1], Xiaoyan Li [1,2], and Liyuan Li [1], Xiaofeng Su [1,3], Fansheng Chen [1,2,3]

[1]University of Chinese Academy of Science, Beijing 100049, China
[2]Hangzhou Institute for Advanced Study, University of Chinese Academy of Science, Hangzhou 310024, China
[3]Key Laboratory of Intelligent Infrared Perception, Chinese Academy of Sciences, Shanghai 200083, China

Corresponding author: : Fansheng Chen (cfs@mail.sitp.ac.cn).

This work was supported in part by the CASEarth Minisatellite Thermal Infrared Spectrometer Project under Grant XDA19010102 and in part by the National Natural Science Foundation of China under Grant 61975222.

ABSTRACT Infrared dim and small target detection is widely used in military and civil fields. Traditional methods in that application rely on the local contrast between the target and background for single-frame detection. On the other hand, those algorithms depend on the motion model with fixed parameters for multi-frame association. For the great similarity of gray value and the dynamic changes of motion model parameters in the condition of low SNR and strong clutter, those methods possess weak robustness, low detection probability, and high false alarm rate. In this paper, an infrared video sequences encoding and decoding model based on Bidirectional Convolutional Long Short-Term Memory structure (Bi-Conv-LSTM) and 3D Convolutional structure (3D-Conv) is proposed, addressing the problem of high similarity and dynamic changes of parameters. For solving the problem of dynamic change in parameters, Bi-Conv-LSTM structure is used to learn the motion model of targets. And for the problem of low local contrast, 3D-Conv structure is adopted to extend receptive field in the time dimension. In order to improve the precision of detection, the Decoding part is divided into two different full connection with distinctive active function. Simulation results show that the trajectory detection accuracy of the proposed model is more than 90% under the condition of low SNR and maneuvering motion, which is better than traditional method with 80% in DB-TBD 20% in others. Real data experiment illustrate that that our proposed method can detect small infrared targets with a low false alarm rate and high detection probability.

INDEX TERMS Deep Learning (DL), Neural Network (NN), dim and small target detection, Long Short-Term Memory(LSTM), 3D Convolutional.

I. INTRODUCTION

With the advantages of strong concealment and wide field of view, space-based infrared detection system plays a great role in the military and civil fields, and has been widely concerned by academia and industry. Aiming to find targets in advance and make decisions, the detection of dim and small target is one of the important research directions in that system, and has great significance in space defense, satellite search and rescue, national security and other fields. Due to the small area, low intensity and lack of texture of the target detected by this platform, and the existence of various noises in the imaging system, dim and small target detection is a great challenge [1]. In recent years, many methods have been proposed.

The detection algorithms can be divided into single-frame-based, multi-frame-based, and neural-network-based.

A. Single-frame-based

Assuming that the original infrared image is composed of background image with low rank, target image of sparse, and noise image, Gao et al. migrates the study of Low Rank (Low Rank) [2 - 4] and convert the target detection problem into a Sparse and Low Rank matrix recovery problem under constraint conditions. Then the IPI model [2] is proposed by Gao C, which is effectively solved using accelerating proximal gradient descent (APG) or alternating direction multiplier method (ADMM). However, the model requires strict background low-rank assumptions, so it is lack of robustness in practical application. Moreover, the solution of the model requires multiple iterative operations, which makes it difficult to guarantee the real-time performance of the algorithm. According to the human visual system (HVS), Chen et al. proposes a detection algorithm [5] based on the local contrast (LCM). This method enhances the small target using human eyes perceive mechanism: Because the strong contrast area can be highlighted by the human eye, and the area which exist targets meet the conditions, targets can be enhanced through calculating local contrast. Many improved LCM algorithms [6, 7] are proposed. But, all of these algorithms require that the gray value of the target satisfies the local maximum hypothesis. With the similar gray values between targets and its neighborhoods in the condition of low SNR, LCM and others cannot enhance the target.

B. Multi-frame-based

In order to address the problems of dim target detection under low SNR, multi-frame detection algorithms are proposed. A typical method is called Tracking Before Detection algorithm (TBD). Using tracking algorithms with the fixed-parameters motion model, false detection are eliminated by data association between continuous images in TBD. Popular methods of tracking include particle filtering, Kalman filtering, random finite set (RFS) and other algorithms [8]. To introduce the time dimension information, a time variance filter (TVF) algorithm [9] for multi-frame sequences is proposed by Lvping yue et al., which focuses on the intensity distribution of grayscale values of each pixel in time and analyze the characteristics respectively.

But it would cause missed detection and a large number of false alarms when the targets have not strong gray value. Wang J et al. use dynamic programming algorithm (DP-TBD) to carry out path integral for the target intensity, and use multi-frame accumulation to carry out noise reduction processing and target enhancement at the same time [10, 11]. For using the value function to carry out the optimal accumulation at each moment to search for the global strategy combination, DP-TBD assumes that the gray value of the target on the full path maintains maximum value. The global strategy combination maximizes the value function by searching the path of the target movement. Therefore, the key point of this method lies in the selection of the value function. This algorithm has high detection accuracy, but its practical application is limited due to the huge computation amount and the harsh assumptions. Liu et al. extends the IPI model of two-dimensional space to three-dimensional space [12], but this method is also subjected to low-rank hypothesis. Peng et al. proposes STLDM algorithm [21]. By extending LCM of a single frame to a three-dimensional LCM, the local contrast of the space-time joint can be calculated. This method effectively uses the information of time dimension.

C. Neural-network-based

There are relatively few researches using neural network to detect dim and small target in video sequences [16 - 18]. Due to the small proportion of target area, lack of rich texture information and low target intensity, target detection algorithms in the general field, such as Fast-RCNN, YOLO, SSD, etc. [14, 15], are not suitable for such application scenarios. Shi M et al. [17] proposes a method for a single frame image using semantic segmentation, better background suppress factor and SNR gain can be obtained by segmenting the image pixel by pixel. In order to extract spatial-temporal information at the same time, Sinn Y U et al. [16] uses 3D-Conv to conduct time-space joint detection of multi-frame images. This method introduces 3D-Conv structure to process multi-frame infrared images for the first time. For the relatively simple processing of video sequences in this method, there are problems such as imperfect extraction of image time-sequences information. Due to the difficulty in obtaining images of space-based infrared dim and small target, data-driven deep learning methods cannot be rapidly developed. In order to make up for the lack of data in the field, Young et al. [19] use the ASSET system to simulate image data sets.

Neural network can effectively utilize the key features hidden in the video sequences to ensure the robustness of the algorithm under the condition that the target gray level is similar to the background gray level. For that, we propose a Encoding-Decoding model of infrared video sequences is proposed: Bi-Conv-LSTM structure is used for addressing the problem of dynamic change of parameters and 3D-Conv structure is used for encoding video sequences into high-dimensional feature vector, and two-way full connection (FC) structure is used for calculating the target positions and confidence simultaneously.

Our main contribution of the thesis is in the fellow:

- We propose a Encoding-Decoding model of infrared video sequences which contains Bi-Conv-LSTM and 3D-Conv structure. Experiments show that our model have more effective than traditional method.
- Simulation and real datasets are used to verify the validity of the model;

In real data experiments, our method achieve the best performance.

This paper is organized as follows: Section Ⅱ introduces the proposed Encode-Decoding model and the design of the loss function. Section Ⅲ introduces experiments. In this section ,we discuss the convergence of proposed model, the detection performance in different SNR and motion models. Section Ⅳ is summary.

II. Method

A. TIME SLIDING WINDOW

The First Input First Output (FIFO) queue is established in this paper for processing continuous-time images in video sequence, and the method of time sliding window is adopted. The diagram is shown in Fig 1. When the new frame data is updated, the image data of the first frame is deleted from the FIFO, and the new image data is added at the tail of the FIFO to maintain the time continuity of the video sequences.

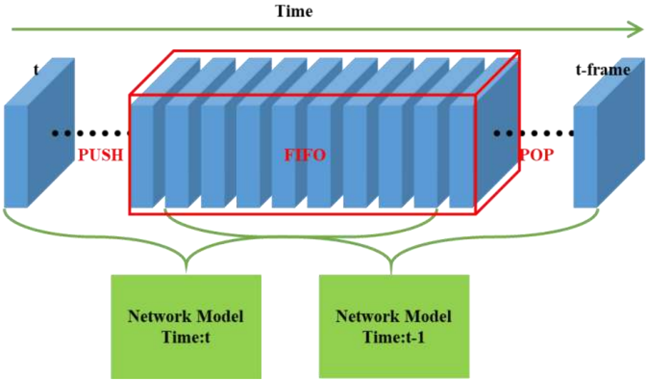

FIGURE 1. Diagram of Time sliding window FIFO queue

B. NEURAL NETWORK MODEL

Inspired by the general target detection models, a $Patch \times Patch$ spatial window is used for detecting small targets. In the high orbit remote sensing image, the infrared target occupies a small area and too large size of spatial window should not be used. Therefore, the Patch size is selected as 64 pixels in this paper. The overall structure of the network model is shown in Fig 2. For processing video sequences, an infrared video sequences Encoding-Decoding structure is proposed in this paper. The model consists of two parts: The Encode part based on the Bi-Conv-LSTM structure and the 3D-Conv structure, and the Decoding part of the fully connected structure. In this model, Bi-Conv-LSTM structure is first used for processing video sequences inputs. Then the outputs of each time are concated as input of 3D-Conv structure. The concated data is called for Concatenate Matrix (CM). After the convolution of 3D-Conv structure pooling operation, Concatenate Matrix is squeezed into high-level feature vector, which contains in the location of the object and probability. The depth spatial-temporal feature extraction is completed in Encoding part. Then the Decoding part decodes the target information through the two-channel fully connected structure to obtain the target position and probability.

Since the value of the probability of the existence of the target ranges from 0 to 1, the sigmoid activation function is used in the full connection structure for calculating the probability. It would lose the regression accuracy in the experiment using sigmoid activation function in the full connection of target position regression because the value of the target position ranges from 0 to Patch, so ReLU activation function is used in the full connection structure for calculating the position information.

Traditional multi-frame detection algorithms need to assume the motion of the target as fixed parameters in advance, and the algorithms use a certain motion model to carry out target correlation matching among multiple frames. But in the actual situation, the assumed motion model deviated from the real motion because the information such as the target's position, speed and motion direction cannot be known in advance. It reduces the detection accuracy. In addition, in order to cover more motion models, those algorithms need to pay extra calculation cost. Therefore, traditional algorithms need to balance between calculation cost and detection accuracy.

For addressing the problems existing in the traditional multi-frame detection algorithms, we use the Bi-Conv-LSTM structure to extract the temporal features and automatically fit the motion model of the target in the data set. Having short-term dependence and long-term dependence, motion models can be learned by LSTM structure.

Ordinary LSTM structure processes sequences according to the order of time, so the information at the next moment can only be predicted based on the information at the previous moment. Future information is introduced by extending a backward LSTM in Bidirectional LSTM structure. In motion models, the introduction of future information can make the current prediction of information more accurate, so the bidirectional structure is used. The Bi-Conv-LSTM structure used in this paper is shown in Fig 3. It is a variant of LSTM

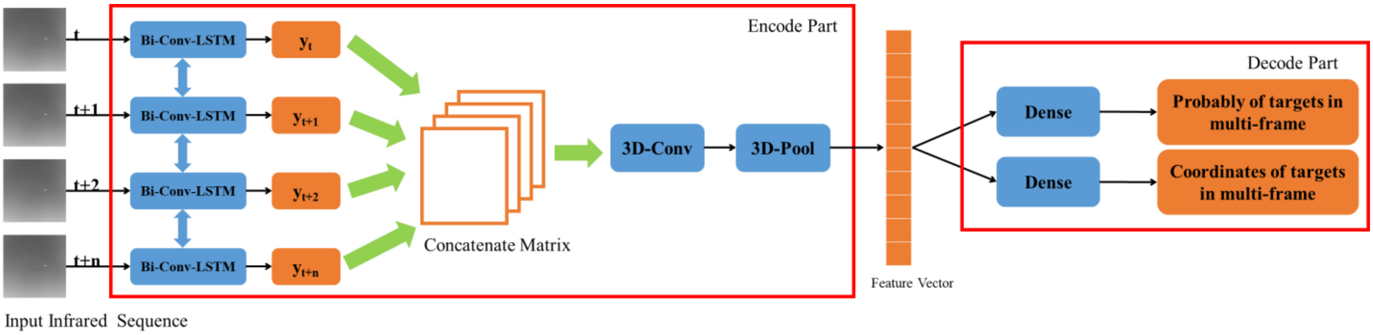

FIGURE 2. Overview of neural network structure

structure, and is used to process video sequences. After being processed by the Bi-Conv-LSTM structure, the images at each moment in the figure will flow into the next moment and are processed together with the images at the next moment. The figure contains two-time links, one processing forward along the time axis and the other processing backward along the time axis. The output of Bi-Conv-LSTM structure is the superposition of two links. Stacking the outputs at all times to obtain the Concatenate Matrix processed at the first level.

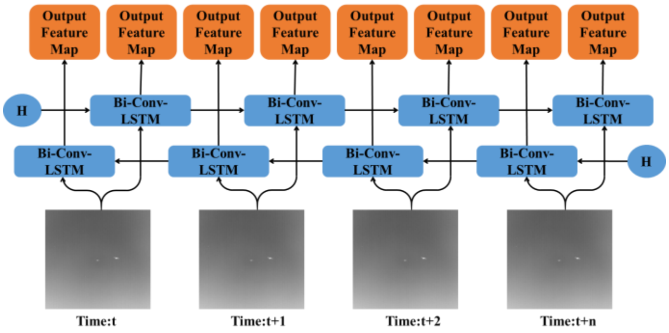

FIGURE 3. **Schematic diagram of Bi-Conv-LSTM in small target detection**

Concatenate Matrix is obtained through the first-level processing of Bi-Conv-LSTM structure, and then the second-level feature extraction of Concatenate Matrix is carried out by 3D-Conv structure. In the infrared video sequences, the continuous movement of dim and small targets is shown in Fig. 4(a). The traditional single frame detection algorithm or multi-frame detection algorithm can only detect targets with high local contrast. And the height of target and background are similar as shown in figure 4 (a), it is unable to identify the location of the target using the two-dimensional space only. But the target is clearly using the time dimension. Therefore, this paper introduces 3D-Conv structure to carry out feature mapping on Concatenate Matrix to fully extract spatial features.

TABLE I
NUMBER OF NEURAL NETWORK PARAMETER

Layer	Parameter
bidirectional_1 (Bidirectional)	408
Conv3d_1 (Conv3D)	1002
Conv3d_2 (Conv3D)	2004
Conv3d_3 (Conv3D)	8008
Conv3d_4 (Conv3D)	32016
Conv3d_5 (Conv3D)	128032
Conv3d_6 (Conv3D)	320040
dense1_1 (Dense)	2624
dense1_2 (Dense)	2624
dense2_1 (Dense)	2080
dense2_2 (Dense)	2080
dense3_1 (Dense)	1056
dense3_2 (Dense)	1056
outputs1 (Dense)	660
outputs2 (Dense)	330
total	504,020

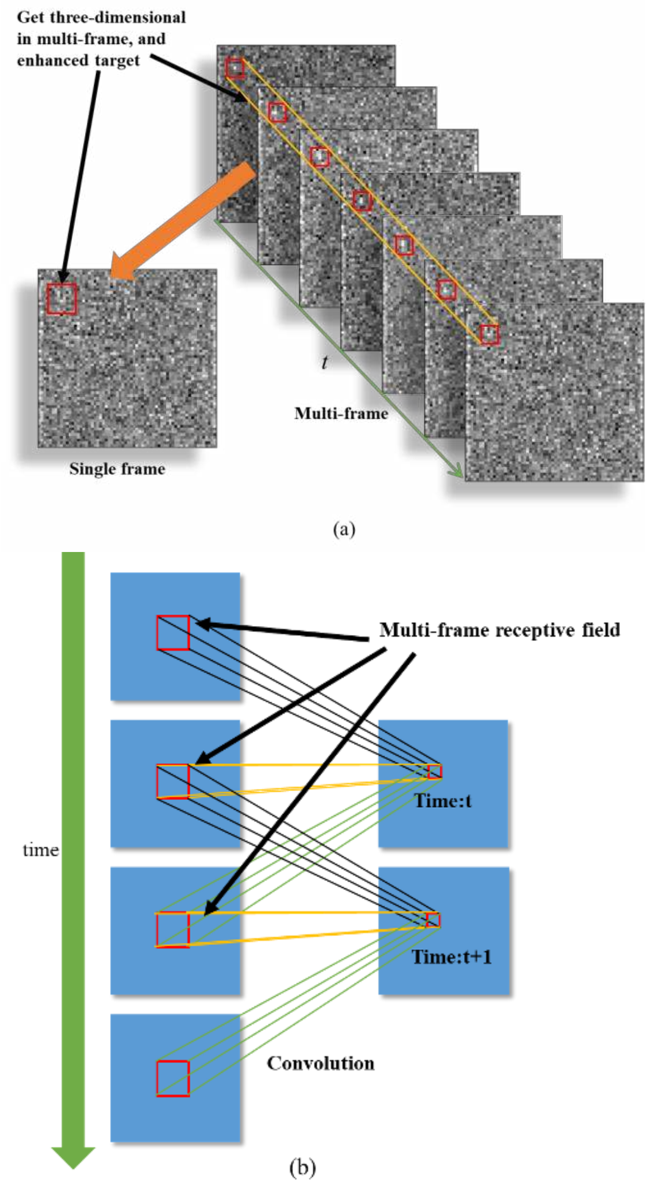
Get three-dimensional in multi-frame, and enhanced target
t
Multi-frame
Single frame
(a)
Multi-frame receptive field
time
Time:t
Time:t+1
Convolution
(b)

FIGURE 4. **(a) Multi-frame continuous motion diagram of a small target (b) Diagram of 3D Convolutional receptive field structure**

3D-Conv structure [20] is shown in Fig. 4(b). The difference between 3D-Conv structure and 2D-Conv structure is that 3D-Conv has time domain receptor field and can combine video sequences context information. So, 3D Convolution has great advantages in video classification, action recognition and human behavior recognition [20].

In order to simplify the calculation, the convolution kernel is designed to share weights in this paper, that is, the same convolution kernel is used on the same feature map.

The data flow in the network structure is shown in Fig 5. ReLU activation function is uniformly used in structures outside the full connection layer, because there are structures that use ReLU activation function in the full connection layer. If sigmoid is used, the input values of the full connection layer will be too close to 0, thus causing the death of neurons.

The number of parameters for the final model are shown in Table I.

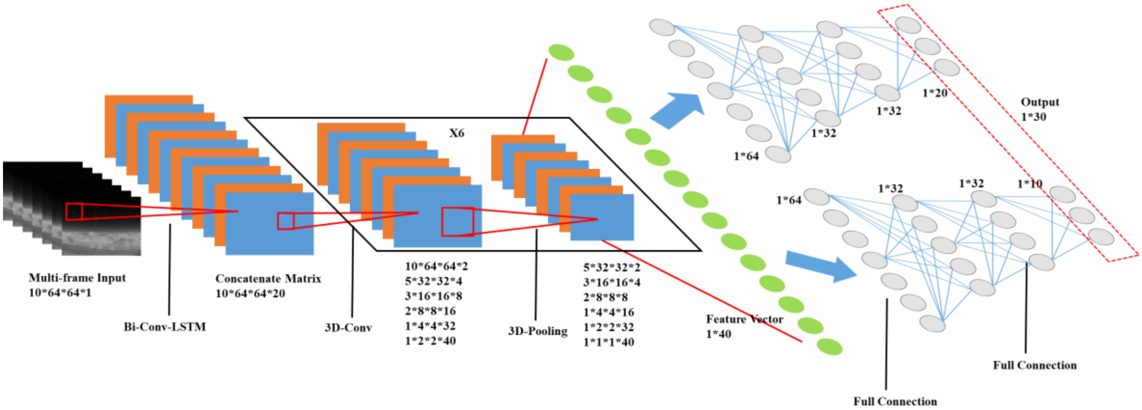

FIGURE 5. **Schematic diagram of data flow in neural network structure**

C. LOSS FUNCTION

The model proposed in this paper can calculate both the probability of the presence of the target and the target position at the same time, so the loss function of the model need to contain two parts:

Equation (1) represents the existence probability loss of the model, which is represented by *catLoss*. As for the confidence of the presence of the target, we consider it a classification problem. Here the binary cross entropy is used in the classification loss function.

Equation (2) represents the regression loss of the model, which is represented by *rLoss*. Its main function is to guide the network correctly calculating the coordinate position of dim and small target. Inspired by the SSD network, the Smooth L1 loss function is used, which can reduce the value of the loss function when the regression error is large, so as to avoid the gradient explosion problem. When the regression error is small, the loss value is increased appropriately to improve the accuracy of the network.

$$catLoss = \sum_i -y_i^p \log(\hat{y}_i^p) \tag{1}$$

$$rLoss = \sum_i \hat{y}_i^p \sum_{m\in\{x,y\}} Smooth_{L1}(\hat{y}_i^m - y_i^m) \tag{2}$$

$$Loss = \alpha * catLoss + \beta * rLoss \tag{3}$$

The weight factors α and β in Equation (3) are used to weight the loss values of the two parts. Since the value of target position ranges from 0 to Patch during training, and the value of target existence probability ranges from 0 to 1, the loss value of target position regression will be much higher than the that of classification. If the loss value of the two parts is not balanced in training, the loss of one part will be dominant in training, while the loss of the other part will produce the phenomenon of gradient disappearance. In order to make the two parts at roughly the same level, we set weights for the two parts respectively.

III. Experiments

In this section, the evaluation index is firstly selected. And then generation method of the data set based on the point spread function is briefly described. Finally, the simulation data set with noise is used for experiments, and the comparison with current mainstream multi-frame detection algorithms is made to illustrates the advantages of the algorithm in processing low SNR target detection.

Equation (4) represents the image's global signal-to-noise ratio (SNR). In the formula, I_k represents the intensity of the point target, and represents the gray value of the center point of the target in the image. σ represents the variance of Gaussian noise added to the image.

$$SNR = \frac{I_k}{\sigma} \tag{4}$$

In order to quantitatively describe performance, the mean absolute trajectory error (MAE) is defined, and the formula is as follows. The position of the target is described by horizontal and vertical coordinates, so the state space form of the target is a one-dimensional column vector. p_i^m represents the value of the m-th dimension of the predicted results, and $\hat{p}_i^m$ represents the true value of the m-th dimension of the ground trues.

$$MAE = \frac{\sum_{i=1}^{n} \sum_{m\in\{x,y\}} |p_i^m - \hat{p}_i^m|}{n} \tag{5}$$

An important index in the performance of dim and small target detection algorithms is the ability of trajectories detect, which reflects whether the algorithm model has sufficient detection accuracy for a continuously appearing target. Therefore, the concept of trajectory detection accuracy is defined in this paper, and the formula is shown in (6).

$$F_D = \frac{\sum_{p\in\{P||p-\hat{p}|<2\}} p}{\|Track\|} \tag{6}$$

If the difference between the detected target and the real target is less than 2 pixels, the detection is judged to be true detection. The ratio of the number of all true detections to the actual track length is defined as the trajectory detection accuracy, and this value ranges from 0 to 1.

A. SIMULATION EXPERIMENT

Due to the influence of light diffraction effect and point spread, the target dose not present a point on the image, but forms a light spot on the image in the remote imaging system. In this

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110395, IEEE Access

paper, the point spread function (PSF) is considered as two-dimensional Gaussian function approximately. At present, the research in the case of minimum target detection are insufficient, so this paper mainly discusses the problem of detection of minimum target size in the range of 3 × 3.

The point spread function approximated to the two-dimensional Gaussian distribution formula is shown in Equation (7). Where Δx and Δy denote pixel size, which are set as 1 in this paper; i and j represent adjacent pixel indexes; Σ represents the ambiguity coefficient of the imaging system.

$$SPF = \frac{I_k \times \Delta x \times \Delta y \times e^{-\left(\frac{(i\times\Delta x)^2+(j\times\Delta y)^2}{2\sigma^2}\right)}}{2\pi\Sigma^2} \tag{7}$$

In order to control the size of the target, the size of the Gaussian template is set as 3×3, the weight of the center point of the template is set as 1, and the weights of the surrounding pixels decrease according to the two-dimensional Gaussian distribution function. The improved point spread function is shown in Equation (8). In order to make the target energy relatively concentrated in the template, the value of Σ is selected as 0.7 in this paper. The template image is shown on the left in Figure 6.

$$SPF = I_k e^{-\left(\frac{i^2+j^2}{2\Sigma^2}\right)} \tag{8}$$

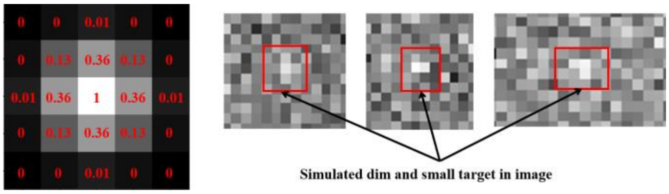

FIGURE 6. Gaussian template and generated dim and small targets.

Then, random noise is added to the image. In order to quantitatively control the SNR of images, the corresponding Gaussian noise with the mean of 0 and the variance of σ are added according to the definition of the global SNR of the image. The three images on the right of Fig 7 shows the dim targets generated by this method, which visually and intuitively resemble the real targets. The data generation flow chart is shown in Fig 7.

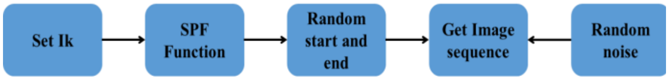

FIGURE 7. Flow chart for generating simulation data.

The above method is used for generating a set of 64*64 size simulation images for experiment. Multiple groups of single target simulation data are randomly generated in the experiment, and each group of data contained 100 frames of continuous images. The starting and ending positions of the target are randomly generated, and the target velocity is determined by the formula below.

$$v_m = \frac{\hat{p}_{frame}^m - \hat{p}_0^m}{frame}, m \in \{x, y\} \tag{9}$$

In Equation (9), $\hat{p}_{frame}^m$ represents the random target termination position. $\hat{p}_0^m$ represents the random target starting position. And $frame$ represents the number of frames of a group of data. Here, 100 is selected. In this case, the target speed is a random number range from 0FPS to 0.5fps. The contiguous image data generated in this way are shown in Fig 8.

Firstly, a convergence experiment is designed to verify the convergence and generalization performance of the proposed model. Then the target detection performance experiment is designed to test the detection performance of the model, and the effectiveness of the model proposed in this paper is illustrated. Finally, the detection performance experiments under different motion models are designed to test the processing ability of the model to different motion models, which indicates that the model has the ability to deal with maneuvering moving targets.

B. CONVERGENCE EXPERIMENT

In order to test the convergence and generalization performance of the model proposed in this paper, training data and test data are generated randomly. The training data include 200 sets of positive category samples with single target and linear motion and 100 sets of Gaussian noise data without target. The test data consist of 100 set of single target data.

The experimental environment of this paper is AMD Ryzen 5 2600X processor, 24G RAM, NVIDIA GeForce GTX 1050 TI graphics card, and 4G video memory. Windows 10 operating system using Keras deep learning framework to complete the design of the network.

We set the hyperparameter α and β in Loss function as 0.1 and 10. As for the $Patch$ in window size is about 64.

In terms of iterative training algorithm, Adam delta algorithm with fast convergence speed is selected. This algorithm has the characteristics of self-adaptive learning rate, and the Batch of each training is set to 10. After 500 epoch training, the network model converged. The training error curve is showed in Fig 9(c).

In the performance testing experiment, since the network can output the positions of multiple frames at the same time, we only select the output of the last frame as the actual detection result to compare with the ground trues considering that the recursive processing method is usually used in the practical application. The calculation formulas of the performance index are shown as above.

Fig. 9(a) shows the average absolute trajectory error on the test data set. The error between the target position calculated by the model and the real position is kept within 1 pixel on most data. Fig. 9(b) shows the trajectory detection accuracy on the test data set. It can be seen from the figure that the detection

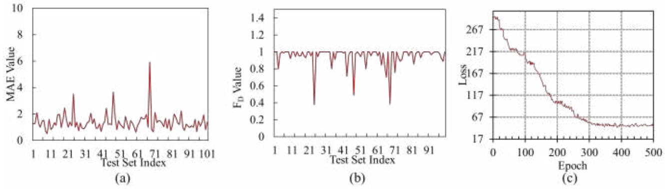

FIGURE 9. (a) Average Absolute Trajectory Error; (b) Trajectory Detection Accuracy; (c) Training error curve

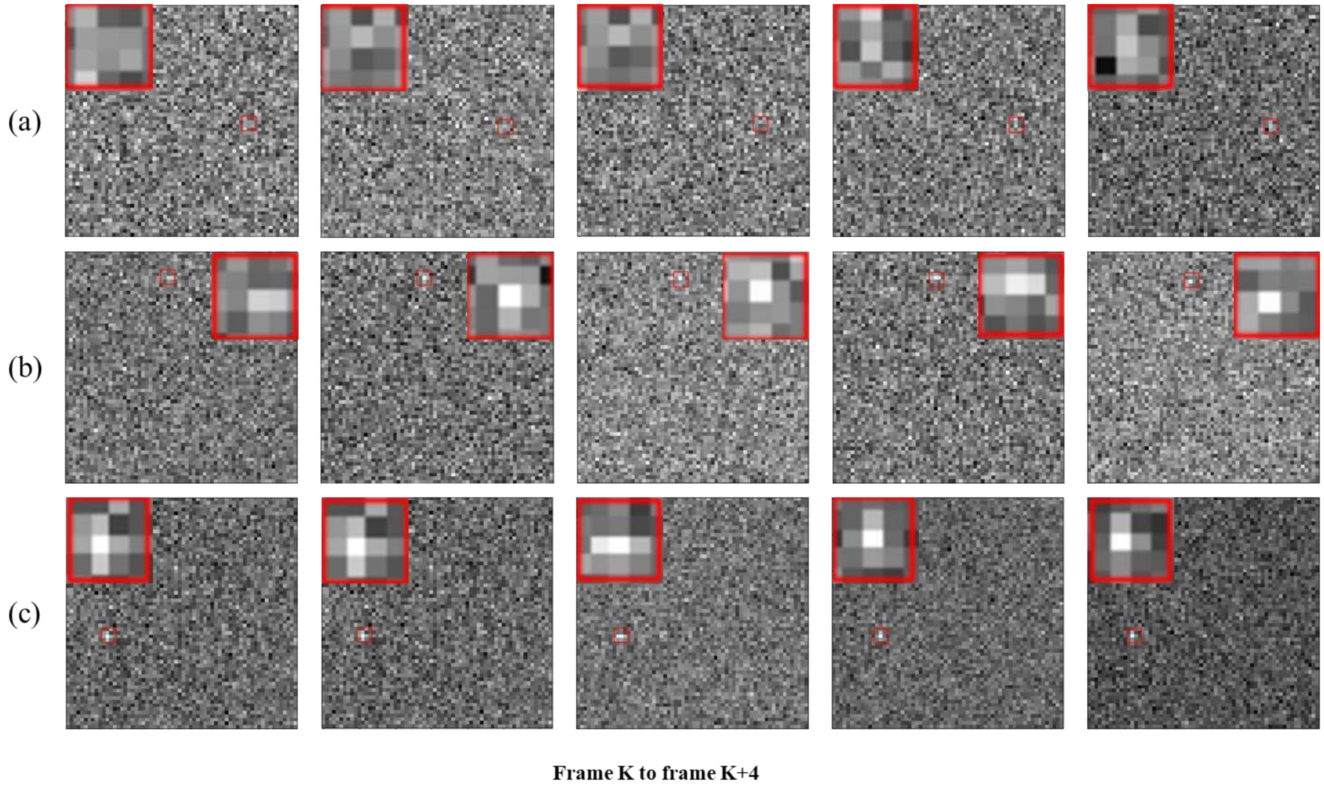

FIGURE 8. (a) Image sequence of simulation in SNR equal to 2 (b) Image sequence of simulation in SNR equal to 3 (c) Image sequence of simulation in SNR equal to 5.

accuracy on most data sets reaches 100% and fluctuates between 85% and 100% on average. Therefore, it can be considered that the model has reached the convergence state.

Meanwhile, the model proposed in this paper has good generalization performance, and its prediction results has high credibility. Table II shows the average values of MAE and FD in the test set. MAE represents a difference of 1.3866 pixels between the detection results and the actual detection results, and the accuracy of the target trajectory is 94.89%, which further indicates that the model can effectively detects the target position.

C. TARGET DETECTION PERFORMANCE EXPERIMENT

The detection performance of the algorithm is quantitatively analyzed by controlling the step change of the image signal-to-noise ratio (SNR). In this paper, the image signal-to-noise ratio (SNR) changes continuously from 2 to 6, with 0.1 as the step value. Three groups of data are simulated under each SNR, so as to compare the performance of each detection algorithm.

In order to reflect the effectiveness of the proposed model, we compare dynamic programming algorithm (DP-TBD) [11], time variance filter algorithm (TVF-TBD) [9], Kalman Filtering (KF-TBD) [8], Spatial-Temporal Local Difference (STLDM) algorithm [21] in this paper. The above algorithms cover the mainstream of dim and small target detection algorithm. The maximum threshold segmentation method is used at the same time as the benchmark.

Fig. 10 shows the detection performance comparison curve of the algorithm. The abscissa is the SNR of targets, and the ordinate is the trajectory detection accuracy (F_D). As can be seen from the figure, the performance of all algorithms increases with the increase of SNR. KF-TBD and TVF-TBD use adaptive threshold segmentation. The effect of these two

TABLE III
TRAJECTORY DETECTION ACCURACY IN DIFFERENT METHOD AND DIFFERENT SNR

SNR	DP-TBD	TVF-TBD	KF-TBD	Max-Value-baseline	STLDM	Our method
2	76.96%	16.13%	10.96%	17.20%	25.13%	**87.26%**
2.5	79.56%	36.33%	28.89%	32.40%	41.80%	**93.93%**
3	80.89%	48.87%	44.07%	51.67%	61.07%	**97.48%**
3.5	83.11%	66.60%	62.67%	69.33%	78.13%	**97.85%**
4	85.70%	67.07%	70.89%	83.80%	89.33%	**98.96%**
4.5	82.30%	75.47%	76.15%	92.93%	94.73%	**96.74%**
5	88.30%	81.80%	76.81%	97.53%	98.20%	**98.44%**
5.5	83.56%	82.27%	79.48%	99.13%	98.80%	**99.78%**
6	92.59%	87.00%	76.67%	99.33%	**99.67%**	97.78%
Average	83.66%	62.39%	58.51%	71.48%	76.32%	**96.47%**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110395, IEEE Access

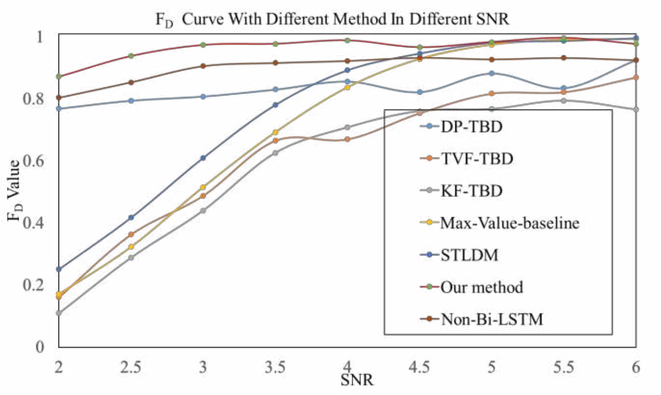

FIGURE 10. **Trajectory detection accuracy in different method.**

methods is similar to the directly used of maximum segmentation. Because these two methods rely on the difference of target gray and background gray, the results of adaptive threshold segmentation are consistent with the results of maximum segmentation. DP-TBD is superior to KF-TBD and TVF-TBD. This is because the dynamic programming algorithm accumulate the maximum grayscale value over time and enhance the targets, thus increasing the information. The proposed method is superior to the current algorithm in the whole SNR condition and achieves the SOTA. This is because the Bi-Conv-LSTM structure and 3D-Conv structure extract the depth features of the video sequence and can find the key features that are ignored in the traditional algorithms.

As for controlled experiment, we replace Bi-LSTM with ordinary LSTM structure. The result is shown in the Fig 10 with brown curve. It can be seen that Bi-LSTM is better than non-bidirectional structure.

Table III shows the average trajectory detection accuracy of the algorithms at SNR from 2 to 6. It can be seen that the model proposed in this paper achieves the best detection results.

We extract the real trajectory of small target movement and the model detection trajectory for comparison. Fig 11 shows the comparison results under four different SNR. As can be seen from the figure, the positions of dim and small targets detected by the model are distributed around the actual positions. The horizontal and vertical coordinates are within two-pixel deviations, and the overall trend is in line with the actual movement of the target. Results indicate that the proposed model can continuously detect the target within the allowable error range.

Table IV shows the average absolute trajectory errors under different SNR. When the SNR increased, the average absolute trajectory errors are significantly improved. The model can achieve an average absolute trajectory error of less than 1.5 pixels.

TABLE IV
AVERAGE ABSOLUTE TRAJECTORY ERROR IN DIFFERENT SNR

SNR	MAE
2-3	1.7802
3-4	1.4461
4-5	1.4344
5-6	1.2163
Average	1.4696

D. EXPERIMENTS UNDER DIFFERENT MOTION MODELS

As the motion model set in the above experiment and that of selected by Kalman filter and dynamic programming algorithm are both uniform liner motion, it is well matched with the actual situation, and these methods have a good effect. In order to verify the effectiveness of the model proposed in this paper under different motion models, the targets of maneuvering movement are generated in this section to observe the detection ability of the model.

Firstly, the motion model is established. It can be considered as a first-order Markov model. The state at time k is denoted as X_k, which contains the position and velocity information of the target and is a four-dimensional vector $[x \quad v_x \quad y \quad v_y]^T$. The state transfer matrix denotes as F, and the noise variance matrix of the transfer process is Q. The state transfer formula is as follows.

$$X_{k+1} = F \cdot X_k + Q \cdot rand \qquad (10)$$

$$F = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (11)$$

$$rand = \begin{bmatrix} \Delta v_x \\ \Delta v_y \end{bmatrix} \qquad (12)$$

$$Q = \begin{bmatrix} \frac{T{q_1}^2}{2} & q_1 & 0 & 0 \\ 0 & 0 & \frac{T{q_1}^2}{2} & q_1 \end{bmatrix}^{Tran} \qquad (13)$$

In this section, the same data generation method is used to generate target data of maneuvering movement combined with first-order Markov model. Fig. 12 shows the detection results of three groups of maneuvering targets when SNR is 3 and 5 respectively. The upper part of the figure represents the comparison figure of maneuvering motion detection when the SNR is 3, and the lower part represents the detection situation when SNR is 5. The model can be effectively applied to the maneuvering motion model. When the target rotates greatly, the horizontal and vertical coordinate errors are still guaranteed to be within two pixels. It is proved that the proposed model has the adaptive ability of the motion model

TABLE V
TRAJECTORY DETECTION ACCURACY AND AVERAGE ABSOLUTE TRAJECTORY ERROR IN DIFFERENT SNR AND MOVING MODEL

SNR	FD	MAE
SNR=5	80.49%	2.7478
	86.84%	3.0389
	97.78%	1.8572
SNR=3	85.06%	2.5854
	94.94%	1.7479
	98.89%	1.5261

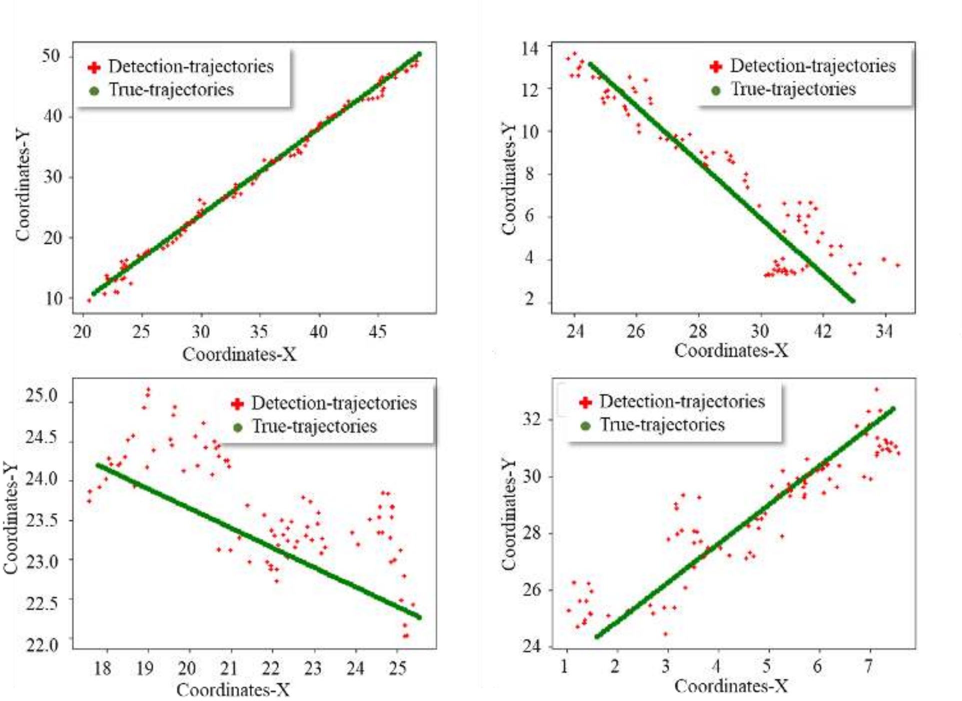

FIGURE 11. Comparison of detection results with actual trajectories in different SNR.

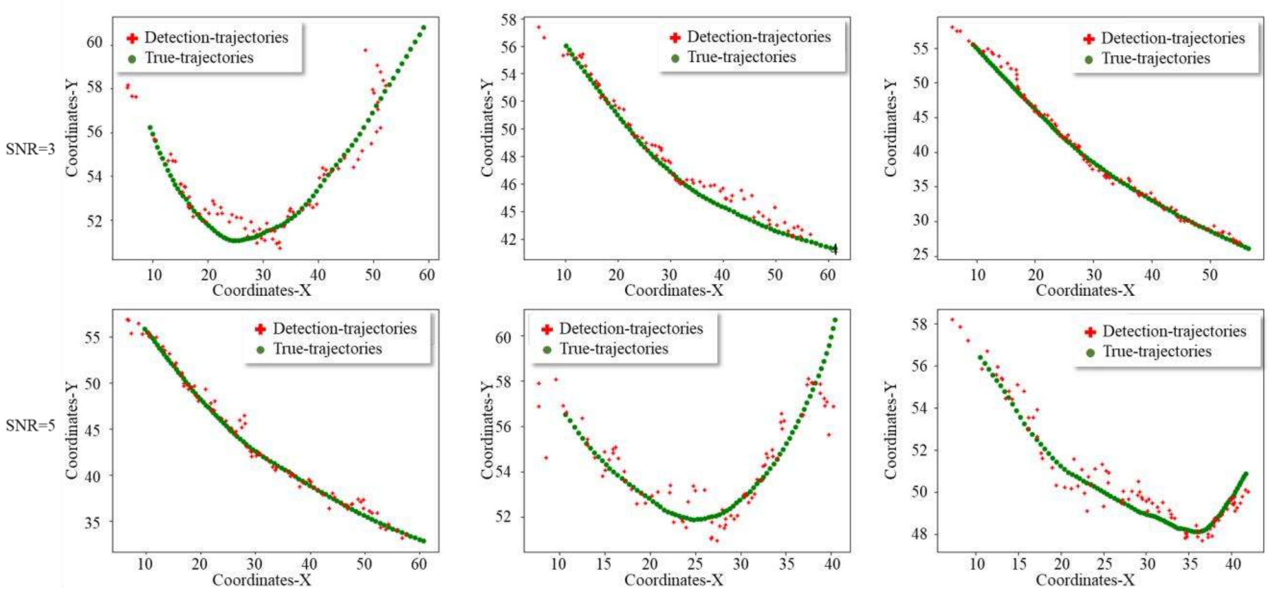

FIGURE 12. Comparison of detection results with actual trajectories in different SNR and moving model

without the need of priori modeling the motion model of the target.

Table V shows the trajectory detection accuracy and mean absolute trajectory error in the above cases. In the maneuverable motion, the target's motion pattern does not appear in the training set. But due to the effect of the Bi-Conv-LSTM network, the model can recognize the motion pattern adaptively. Results shows that the structure has a good ability of motion model generalization.

E. REAL DATA EXPERIMENT

In this section, we use one real infrared video sequences containing dim targets to evaluate the performance of the

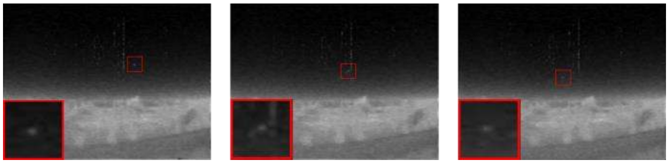

FIGURE 13. Image sequences in OTCBVS.

proposed algorithm. Sequence is from the "Plane Motion and Tracking" image sequence in the OTCBVS Dataset 05, which contains a single object (with very few pixels) moving from the top right to the bottom left in the field of view, the image size is 320*240 pixels, and the sequence is 760 frames.

The images in Sequence is shown as Fig 13.

Algorithms used for comparison include single-frame-based (IPI, HWLCM) and multi-frame-based (STLDM, TVF-TBD). The parameter of each algorithm are set in the following table.

TABLE VI
PARAMETER SETTING OF EACH ALGORITHM

Methed	PARAMETER SETTINGS
IPI	Patch size: 50x50, sliding step: 10, $\lambda = \dfrac{1}{\sqrt{50}}$, $\varepsilon = 10^{-7}$
HWLCM	K_T=5, K_B=9
STLDM	L=5
TVF-TBD	M=16,N=8

The ROC curve in two sequences are shown in Fig 14.

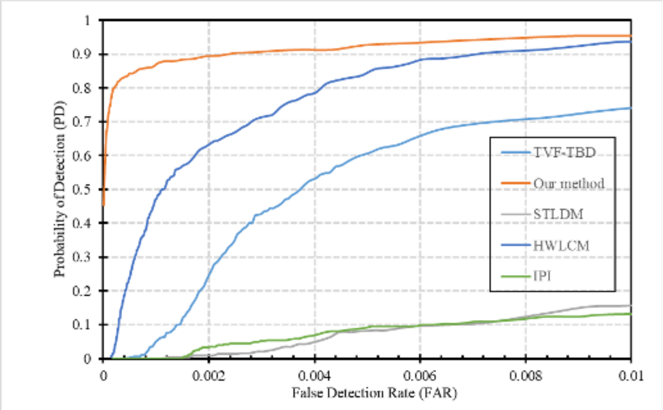

FIGURE 14. Trajectory detection accuracy in different method.

It can be seen that our method achieve the best performance compared to other methods in real inferred data.

IV. Conclusion

In this paper, a Encode-Decoding model based on neural network is proposed for targets detection with low visual salience and strong maneuvering in infrared video sequences. We use Bi-Conv-LSTM structure and 3D-Conv structure to encode the infrared video sequences, compress the video into feature vectors. And the full connection layer is used as the decoding structure to infer the target's confidence and position. Comparing with the traditional algorithm, the proposed model not only has stronger robustness under low SNR and low visual salience, but also can be applied to the maneuvering target with better applicability. The simulation results further illustrate the advantages of this model. This method can be applied in space-based infrared system or IRST, which provides research ideas for further exploration of the detection of dim and small targets with extremely low SNR, low visual contrast and strong maneuvering movement. Future work will focus on more complex backgrounds. By studying more realistic infrared data, background and noise suppression can be integrated into this model.

REFERENCES

[1] Eysa Raziye, Hamdulla Askar, “Issues on Infrared Dim Small Target Detection and Tracking” *2019 International Conference on Smart Grid and Electrical Automation (ICSGEA)*: 452-456.

[2] 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.

[3] Y. Sun, J. Yang. Meng, W. An , “Infrared Dim and Small Target Detection via Multiple Subspace Learning and Spatial-Temporal Patch-Tensor Model,” *IEEE Transactions on Geoscience and Remote Sensing*., pp. 1-16, Dec. 2020.

[4] H. Zhu, S. Liu, L. Deng, Y. Li and F. Xiao, “Infrared Small Target Detection via Low-Rank Tensor Completion With Top-Hat Regularization,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 2, pp. 1004-1016, Feb. 2020,

[5] Chen C L P , Li H , Wei Y , et al. “A Local Contrast Method for Small Infrared Target Detection,” *IEEE Transactions on Geoscience & Remote Sensing*, vol. 52, no. 1, pp. 574-581, Jan. 2014.

[6] H. Deng, X. Sun, M. Liu, C. Ye and X. Zhou, "Small Infrared Target Detection Based on Weighted Local Difference Measure," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 7, pp. 4204-4214, July 2016.

[7] X. Bai and Y. Bi, "Derivative Entropy-Based Contrast Measure for Infrared Small-Target Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2452-2466, April 2018.

[8] D. J. Salmond and H. Birch, "A particle filter for track-before-detect," *Proceedings of the 2001 American Control Conference. (Cat. No.01CH37148)*, Arlington, VA, USA, 2001, pp. 3755-3760 vol.5.

[9] LvPing-yue, Chang-qing, Lin, *et al.* “Dim small moving target detection and tracking method based on spatial-temporal joint processing model - ScienceDirect,” *Infrared Physics & Technology*, vol. 25 , pp. 1469-1483, 2019.

[10] J. Wang, W. Yi, L. Kong and J. Yang, "A computationally efficient recursive processing for multi-frame track-before-detect," *2015 IEEE Radar Conference (RadarCon)*, Arlington, VA, USA, 2015, pp. 0515-0520.

[11] W. Yi, Z. Fang, W. Li, R. Hoseinnezhad and L. Kong, "Multi-Frame Track-Before-Detect Algorithm for Maneuvering Target Tracking," in *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 4104-4118, April 2020.

[12] H. -K. Liu, L. Zhang and H. Huang, "Small Target Detection in Infrared Videos Based on Spatio-Temporal Tensor Model," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8689-8700, Dec. 2020.

[13] R. Girshick, "Fast R-CNN," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1440-1448.

[14] Redmon, J. and Farhadi, A., “YOLOv3: An Incremental Improvement”, *arXiv e-prints*, 2018.

[15] Y. U. Sinn, K. M. Hopkinson, B. J. Borghetti and B. J. Steward, "IR Small Target Detection And Prediction With ANNs Trained Using ASSET," *2019 IEEE Aerospace Conference*, Big Sky, MT, USA, 2019, pp. 1-11.

[16] Shi, M., Wang, H. “Infrared Dim and Small Target Detection Based on Denoising Autoencoder Network.” *Mobile Netw Appl,* vol. 25, pp. 1469–1483 , 2020.

[17] Y. Dai, Y. Wu, F. Zhou and K. Barnard, "Attentional Local Contrast Networks for Infrared Small Target Detection," in *IEEE Transactions on Geoscience and Remote Sensing*..

[18] Young S R , Steward B J , Gross K C . “Development and validation of the AFIT scene and sensor emulator for testing (ASSET),” *Infrared Imaging Systems: Design, Analysis, Modeling, & Testing XXVIII. International Society for Optics and Photonics*, 2017.

[19] SHI X,CHEN Z,WANG H,et al. “Convolutional LSTM network: a machine learning approach for precipitation nowcasting.” *Proceedings of the 28th International Conference on Neural Information Processing Systems*. 2015.

[20] Ji S , Xu W , Yang M , et al. “3D Convolutional Neural Networks for Human Action Recognition.” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, pp. 221-231, 2013.

[21] P. Du and A. Hamdulla, "Infrared Moving Small-Target Detection Using Spatial–Temporal Local Difference Measure," in *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 10, pp. 1817-1821, Oct. 2020.

XIN LIU received the B.S. degree in Wuhan University of Science and Technology, Wuhan, China, in 2017. He is currently pursuing the Ph.D. degree in electronic circuit and system at Shanghai Institute of Technical Physics of the Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China. He current research interests include Image processing, Machine learning, and high performance computing

XIAOYAN LI received the B.S. degree in Mechanism design, manufacturing and automatization from Northwest A&F University, Xi’an, China, in 2016. He is currently pursuing the Ph.D. degree in electronic circuit and system at Shanghai Institute of Technical Physics of the Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China. His current research interests include on-orbit accurate navigation and geometric calibration of remote sensing satellites.

LIYUAN LI received the B.S. degree in opto-electronics information science and engineering from Dalian University of Technology, Dalian, China, in 2018. She is currently pursuing the Ph.D. degree in physical electronics at Shanghai Institute of Technical Physics of the Chinese Academy of Sciences, Shanghai, China. Her current research interests include dim and small targets detection of IR through machine learning.

FANSHENG CHEN received the B.S degree in optoelectronic information engineering and Ph.D. degree in physical electronics from Shandong University, Jinan, China, in 2002 and Shanghai Institute of Technical Physics of the Chinese Academy of Sciences, Shanghai, China, in 2007, respectively. Since 2013, he has been a Professor with the Shanghai Institute of Technical Physics of the Chinese Academy of Sciences. His research interests include the design of spatial high resolution remote sensing and detection payloads, high-speed and low noise information acquisition technology, and infrared dim small target detection technology. Meanwhile, he has been committed to the research and development of the space infrared staring detection instruments, the high spatial and temporal resolution photoelectric payloads, and the application of infrared multi-spectral information acquisition technology in artificial intelligence, target recognition and other relative aspects.