The l0th International Conference on Control, Automation and Information Sciences
Xi'an, China - October 14-17, 2021

Target detection and tracking algorithm based on improved Mask RCNN and LMB

Zhuoqun Liu, Yingjie Deng, Feng Ma, Jinming Du, Chao Xiong, Moufa Hu, Luping Zhang, Xuhuan Ji
National Key Laboratory of Science and Technology on Automatic Target Recognition,
College of Electronic Science and Technology, National University of Defense Technology
Changsha,China
447824977@qq.com

2021 International Conference on Control, Automation and Information Sciences (ICCAIS) | 978-1-6654-4029-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCAIS52680.2021.9624519

Abstract—**This paper proposes an improved Mask RCNN and LMB algorithm for target detection and tracking in complex backgrounds, which follows the architecture of Detect-Before-Track. First, the novel algorithm adopts a two-stage neural network to improve the target positions' accuracy during the detection. Then, the label multi-Bernoulli filter, which is suitable for scenarios with an unknown number of targets and target intersection, is utilized to generate the multiple-target trajectories during the tracking stage. Experiments suggest the algorithm's effectiveness and superiority in complex backgrounds' infrared target detection and tracking.**

Keywords—Infrared target recognition, Detect-Before-Track, Mask RCNN, label multi-Bernoulli filter

I. Introduction

Infrared imaging systems are widely used in civil and military fields due to the unique advantages of excellent concealment, strong anti-interference ability, and practicality under all weather and time. As a significant part of target recognition, infrared target detection and tracking have become research hotspots in recent years[1]. Affected by complex background clutter interference, such as complicated clouds, sea surface, and so on, targets are easily submerged in background clutter and noise. What's worse， the occlusion due to small objects movement, ambiguity causing by infrared camera jitter and rotation, even low signal-to-noise (SNR) commonly raise challenges in infrared target recognition. As a whole, multiple-target detection and tracking are still challenges in reality.

At present, infrared target detection and tracking algorithms can be summarized into two categories[2]: Detect-Before-Track(DBT) and Track-Before-Detect(TBD). The former include the template matching method[3], wavelet transform method[4], and saliency segmentation method[5], etc. However, the segmented suspected target does not contain the real target under low SNR, which leads to the failure of the algorithm. The latter contain particle filter method[6], three-dimensional matching filter method[7], dynamic programming method[8], high-order correlation method[9], and so on. These algorithms make use of the motion characteristics to process multiple frame images. Unfortunately, they are unsatisfying because of their large storage capacity and complex operation.

Recently, the target detection algorithm based on deep learning has been broadly adopted in various target detection tasks for outstanding learning and representation ability. Detection algorithms based on deep learning firstly exploit a neural network to extract feature information from numerous data. Then they combine feature selection and feature classification into the same model to detect objects. Global optimization carrying out through end-to-end training enhances the discrimination of features and leads to superiority over the traditional methods in detection and location accuracy. The representative algorithms are YOLO series[10-13], Retina Net[14], Fast RCNN[15], Faster RCNN[16], Mask RCNN[17], etc.

Aiming at the problem of multi-target tracking, multi-target tracking methods based on Bayesian framework is the core of tracking theory, which raises many scholars' attention. Two types of research ideas form up to now. One is data association multi-target tracking while the other relies on random finite set(RFS) theory. The first kind considers the distinction between measurement and correlation matching between clutter and potential targets, causing the computational complexity especially in high target and clutter density environments. The second type studies target states' optimal and suboptimal estimation without data association between target and measurement. They include probability hypothesis density(PHD)[18], cardinalized PHD (CPHD)[19], multi-target-multi-Bernoulli(MeMBer)[20], and generalized labeled multi-Bernoulli(GLMB)[21]. GLMB retains all the information in multi-objective Bayesian filtering and assigns corresponding labels to each target state, which generates high tracking accuracy. Nevertheless， it suffers computational complexity for the introduction of the target state RFS. Later, B.N.Vo proposed Labelled Multi-Bernoulli(LMB) filtering to alleviate the computational complexity of GLMB filtering.

This paper suggests an improved Mask RCNN and LMB infrared target detection and tracking algorithm used for the public aerial vehicle data set of the third Sky Cup (Figure 1). The algorithm aims to solve the following real issues with great accuracy of target recognition.

(1)The situations not only have low SNR but also contain interferences, such as roads, trees, houses, motorcycles, pedestrians. The algorithm is able to correctly detect targets while reducing false alarms.

(2)The vehicle targets' size and motion are not consistent. At the same time, the pose change of Unmanned Aerial Vehicle flight makes the scenes more complex.

(3)The tracking trajectory continuity problem, which is caused by background occlusion, target birth and disappearance, and the intersection between targets in the driving process.

978-1-6654-4029-5/21/$31.00 ©2021 IEEE

Fig.1. Example diagram of public data set

II. THE SCHEME PLANNING OF INFRARED TARGET DETECTION AND TRACKING

The flow chart of the proposed algorithm is shown in Figure 3. The algorithm includes two parts, one is the target detection algorithm based on improved Mask RCNN, and the other is the target tracking based on LMB filter. Compared with the traditional detection algorithm, the neural network has some advantages in pattern recognition and information processing : due to the use of distributed storage, it has better fault tolerance and anti-interference or noise ability. A pattern is not stored in a fixed place, but distributed in the entire network, and is represented by an activation pattern composed of a large number of neurons. Therefore, it does not have a great impact on the overall situation when some information is lost. In addition, it also has the ability of adaptive self-learning. Through sample training, it constantly changes its network parameters according to the surrounding environment. It can not only deal with various changes of information, but also constantly changes itself while processing information.

The detection algorithm based on neural network algorithm can be divided into single-stage detection algorithm and two-stage detection algorithm. In order to improve the detection accuracy, we use two-stage detection algorithm. In the 2016 COCO Challenge, Mask RCNN has better detection performance in the mainstream two-stage detection algorithm. So we apply Mask RCNN to target detection algorithm.

In the multi-target tracking technology, the traditional multi-target tracking technology associates the measurement with the track, which not only has a large amount of calculation, but also has poor algorithm performance. Multi-target tracking filter based on random finite set theory such as PHD, CPHD, and MeMber filter. Compared with the traditional multi-target tracking technology, can reduce the amount of computation, but can not form a track of the target. The multi-target tracking method of labelled multi-Bernoulli filter realizes the real multi-target trajectory-level filtering, and inherits the excellent performance of multi-hypothesis tracking algorithm in environment with low signal-to-noise ratio, which can be used as the target tracking algorithm in this competition.

According to the actual scene, the specific steps of the algorithm we designed are as follows:(1) Image feature maps are obtained by using Backbone. (2) Get the target candidate box using the Region Proposal Network(RPN) network. (3) Use ROI-Align to get the target detection result box. (4) The improved LMB filter is used to obtain the target trajectory. The flow of detection and tracking is shown in Fig.2. Then we will introduce the detection and tracking scheme in detail.

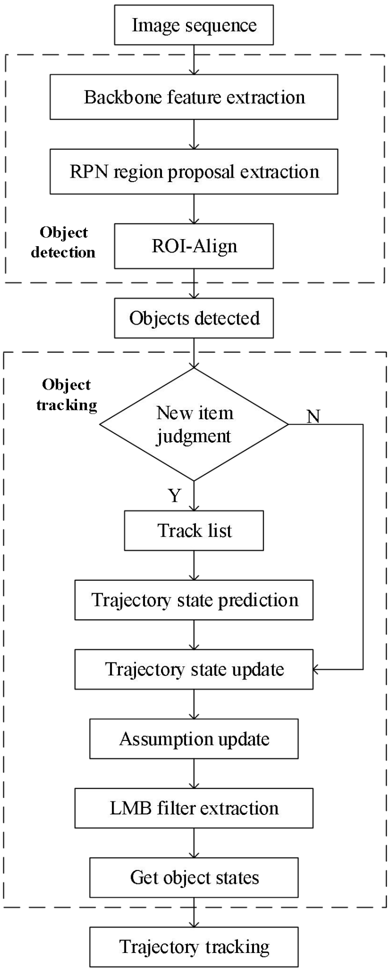

Fig.2. The flow chart of our algorithm

III. Target Detection Algorithm Based on Improved Mask RCNN

Mask RCNN is adept in detecting large targets with rich shape and texture information, like people and vehicles. When it is time for little infrared targets, the original algorithm hardly obtains satisfactory results from train infrared target sets and is time-consuming. So, we improve the Mask RCNN network as Fig.3.

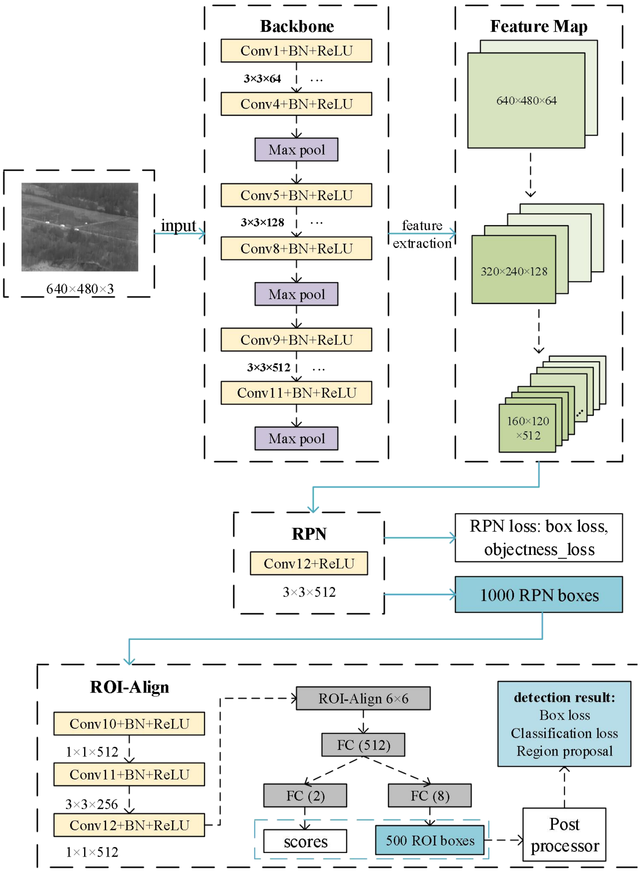

Fig.3. The block diagram of target detection

A. A more lightweight Backbone

Large targets contain more texture structure information and are proper for extracting features by a neural network with a more complex structure.

After analyzing most sizes of targets, we design a Backbone with 11 convolution-layer neural network. Each convolution layer is followed by a BN layer and a ReLU layer. The first four layers of convolution kernel are 3×3×64 while the middle four and the last three are 3×3×128 and 3×3×512, respectively. The output of the convolution kernel is put in a max pool whose kernel, step size, and filling step are 3,2,1, respectively.

Taking an infrared image with a size of 640×480×3 in the competition data as an example, we can get 512 feature maps with each map size of 160×120.

B. Smaller anchor box

Once generating feature maps from Backbone, the RPN network further extracts the target candidate box according to these feature maps. A critical point is to determine the anchor box size. Generally speaking, the designed anchor box should cover each target, unfortunately, a larger size means the anchor box covers multiple targets. It will increase the false alarm rate for detecting little size targets due to the existence of several adjacent objects. For this consideration, the final anchor box sizes are set as 5×5, 9×9, 13×13, 17×17 with corresponding scale factors are 0.5, 1, and 2. The kernel size of the RPN layer is 3×3×512， also followed by a ReLU layer. After the operation in the RPN layer, we obtain 1000 candidate boxes.

C. Different non-maximum suppression threshold

To ensure detection accuracy, the non-maximum suppression strategy is utilized twice in generating the candidate RPN box. When small targets are close to each other, these adjacent target areas are normal to produce too many detection boxes. A proper choice is suppressing repeated detection boxes by a little non-maximum threshold.

Consequently, anchor boxes containing samples are suppressed during the training of the model. Therefore, we miss the ideal training model because of the few positive samples used in the loss calculation.

This paper sets the thresholds as 0.8 and 0.2 to balance the results, respectively. The first-stage detection model chooses a high non-maximum threshold to ensure that sufficient positive samples are involved in training while a low non-maximum value using for ensuring detection accuracy in the second stage.

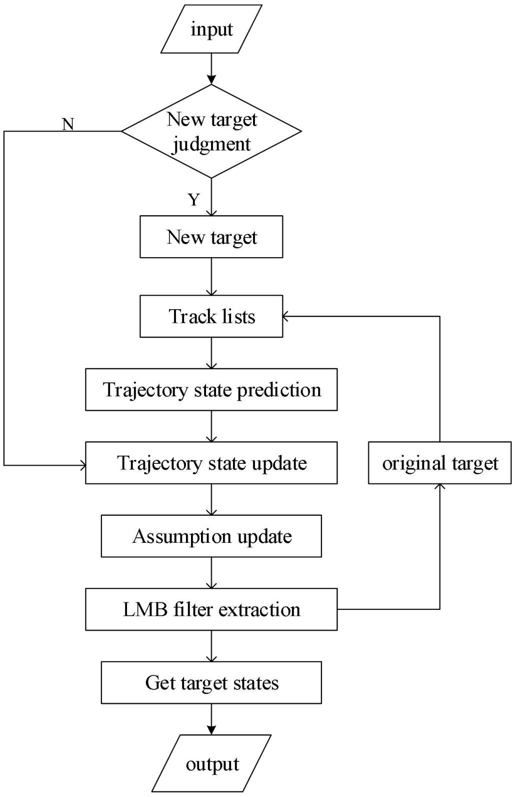

Fig.4. The flow chart of target tracking

D. Accurate classification

ROI-Align is more suitable for small target detection due to combining the bilinear difference method and avoiding pixel deviation causing by the deviation of calculation accuracy. In this step, 1000 candidate boxes are put into this part for more accurate classification. The module first uses three-layer convolution (Conv13 - Conv15) for further feature extraction and then connects ROI-Align. Next, after 512 full connection(FC) layers, 2 FCs are used to obtain binary classification scores, and 8 FCs are used to get the ROI frame with N = 500. Finally, the algorithm calculates positive ROI Boxs' loss and classification loss according to Smooth L1[23].

IV. Target Tracking Algorithm Based on LMB Filter

Target tracking, which needs to generate real target trajectory on the basis of detection, can be roughly divided into new target judgment and real target trajectory update. To solve the problem, involved in the emergence of new targets and the disappearance of old targets, and the change of target labels after the intersection of two targets, we suggest an improved labelled multi-Bernoulli filtering algorithm shown in Fig.4.

The RFS is a set of random variables, of which the number in the set is random. The RFS can be described as

$$X = \{x_1, x_2, \cdots x_N\} \tag{1}$$

where $x_i \in \mathbf{X}^m$ is the m-dimensional single target state, $\mathbf{X}^m$ is the target state space, and N is the number of targets.The multi-Bernoulli process can be regarded as several independent single-objective Bernoulli processes. The corresponding multi-Bernoulli stochastic finite set mathematical expressions are as follows

$$\pi(X) = \{r^i, p^i(\boldsymbol{x})\}_{i=1}^{N} \tag{2}$$

Where r^i and $p^i(\boldsymbol{x})$ are the probability and probability distribution function of the *i*-th target respectively.The random finite set of labels is to add a label to each element in the set. It may define $\mathbf{L}$ as the label value space, and function $\mathcal{L}$ describs the mapping on $\mathbf{X}^m \times \mathbf{L} \to \mathbf{L}$, satisfying $\mathcal{L}((\boldsymbol{x},l)) = l, l \in \mathbf{L}$, so $\mathcal{L}(X)$ is the label value set corresponding to X. The impulse function and indicator function in the set domain are expressed as follows

$$\delta_A(B) = \begin{cases} 1, A = B \\ 0, others \end{cases}, \quad 1_A(B) = \begin{cases} 1, B \subseteq A \\ 0, others \end{cases} \tag{3}$$

A and B in the formula represent set variables. The extended LMB RFS is composed of several LMB RFSs with different weights. The mathematical description is

$$\pi(X) = \Delta(X) \sum_{h \in \mathbf{H}} \omega_h(\mathcal{L}(X))[p_h]^X \tag{4}$$

In the formula ω_h, h, $\mathbf{H}$ are weight function, index, index space. $\Delta(X) = \delta_{|X|}(|\mathcal{L}(X)|)$ denotes the different labels of each point in the set.

When $p_h(x,l) = p_h(X), \omega_h(\mathbf{L}) = \prod_{i \in \mathbf{L}}(1 - r^i) \prod_{l \in \mathbf{L}} \frac{1(l) r^l}{1 - r^l}$, a priori form of LMB can be obtained

$$\pi(X) = \Delta(X) \sum_{h \in \mathbf{H}} \prod_{i \in \mathbf{L}} (1 - r^i) \prod_{l \in \mathbf{L}} \frac{1(l) r^l}{1 - r^l} [p_h]^X \tag{5}$$

In the formula, $\Gamma(X_{k-1})$, $\Upsilon(X_{k-1})$ and $X_{B,k}$ represent survival, derivative and natural newborn target states, respectively, independent of each other. We do not need to consider the $\Upsilon(X_{k-1})$ term in the above equation because of no derivative target in the data set.

The tracking part of the algorithm judges the new target item according to the measurement results obtained by target detection. If the input measurement is not related to the previous survival item, it is judged as the new target item and adds the new target item to the trajectory list. After that, the algorithm performs the steps of prediction, update and hypothesis update of track state, and extracts the tags of all targets as new survival items. If a measurement can be related to the survival item, only the track state update and hypothesis update process need to be performed until all the measurement values are input. Finally we can obtain required track information.

V. Experimental Results

To verify the validity of the proposed algorithm, we randomly selected two sets of 64 training sets. The detection and tracking results for training set-3 and training set-38 are shown in Fig.5. and Fig.6. In each frame, the blue box(Ground Truth), the red box(Meas) and the green box(Track) are the actual positions of the targets, the target positions detected by the algorithm and the target tracking positions respectively. It is pointed out that the size of the detection tracking box does not represent the actual size, but only to observe the effect better. Taking Figure 5(a) as an example, T1-1, T1-2, T1-3 and T1-4 are four targets in the scene respectively.

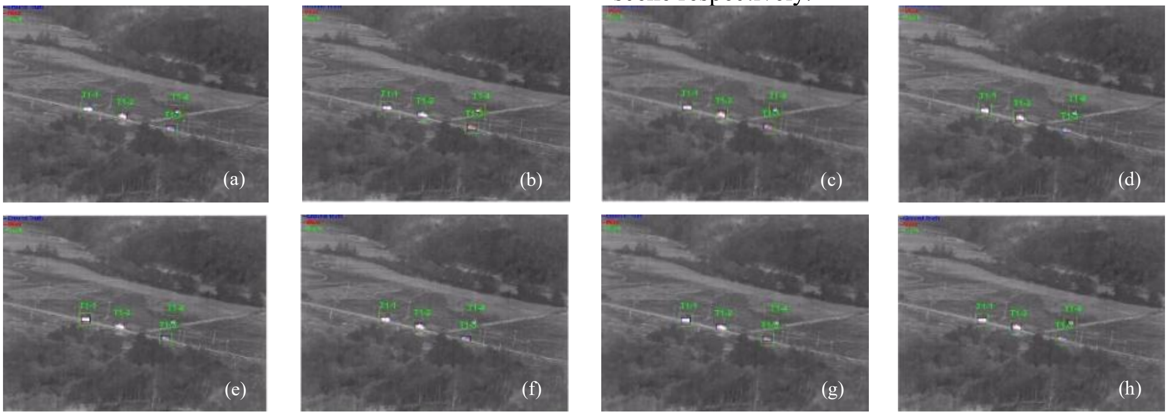

Fig.5. Training set -3 detection and tracking results (a-h are the images from 3rd frame to 10th of the original image)

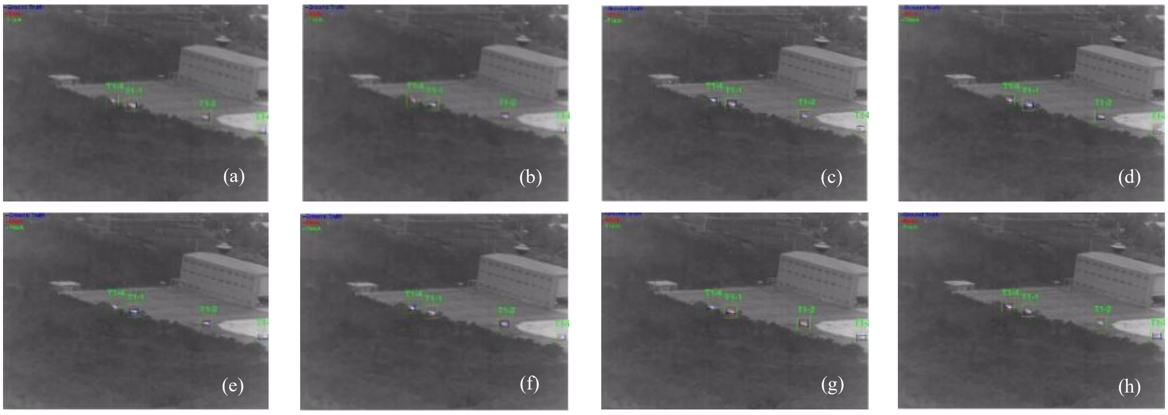

Fig.6. Training set -38 detection and tracking results (a-h are the images from 22nd frame to 29th of the original image)

From the results in Fig.5 and Fig.6, it can be seen that the proposed algorithm first correctly detects all the real targets in the scene. Through the tracking results of eight consecutive frames, the proposed algorithm can correctly track the target without frame loss and tracking errors between targets, proving the effectiveness of the our algorithm. The algorithm in this paper only selects some results for presentation due to space reasons, but it is also applicable to other training sets.

According to statistics, the training results of the improved Mask RCNN and LMB infrared target detection and tracking algorithm on the public data set are shown in Table Ⅰ. In the meanwhile, the algorithm obtained 11834 points in the online test of the third Sky Cup final, ranking second, which has reference value for infrared target detection and tracking in multiple background. And the FPS is 9.68 in the 1080Ti configuration computer, meaning that the algorithm also needs to further improve the operation speed to meet the real-time requirements.

TABLE I. DETECTION RESULTS ON PUBLIC DATASETS

Detection rate	**Miss rate**	**False alarm**
94.38%	5.62%	1%

VI. CONCLUSION

We propose a method of DBT for infrared target detection and tracking. An impoved Mask RCNN is proposed for target dection, in whicn a more lightweight backbone, small anchor boxes, different non-maximum suppression threshold are used. For target tracking based on LMB filter, this paper propose a new target item judgment module to solve the problem of the disappearance of the old target generated by the new target in the scene. Experiments and online tests suggest the effectiveness of the algorithm. It is worth noting that although the algorithm is based on the open data of this Sky Cup, its framework is general and can be used for reference to solve the problem of target recognition of similar infrared images.

REFERENCES

[1] Yin H. P. , Chen B. , Chai Y. , et al. ， “Vision-based Object Detection and Tracking: A Review，” Acta Automatica Sinica, 2016.

[2] Hou, W., et al., “Present State and Perspectives of Small Infrared Targets Detection Technology,” Infrared Technology 37.1(2015):1-10.

[3] R. M. Liu, X. L. Li, L. Han, “Track infrared point targets based on projection coefficient templates and non-linear correlation combined with Kalman prediction,” Infrarded Physics & Technology, 2013, 57:68-75.

[4] Boccignone G., Chianese A., Picariello A., “Small Target Detection Using Wavelets,” International Conference on Pattern Recognition. IEEE Computer Society, 1998:1776-1778.

[5] Yang Y., Yang J., “Pedestrian detection of infrared images based on saliency segmentation,” research studies of toyama mercantile marine college, 2013.

[6] Qian Kun，Zhou Huixin，Qin Hanlin，et al. ，“Guided filter and convolutional network based tracking for infrared dim moving target,” Infrared Physics & Technology，2017，85: 431－442.

[7] Deng Lizhen，Zhu Hu，Tao Chao，et al. ，“ Infrared moving point target detection based on spatial-temporal local contrast filter,” Infrared Physics & Technology， 2016，76: 168－173.

[8] Arnold J., Shaw S. W., Pasternack H., “Efficient target tracking using dynamic programming,” IEEE Transactions on Aerospace & Electronics Systems,1993,29(1):44-56.

[9] Liou R. J., Azimi-Sadjadi M .R., “Dim target detection using high order correlation method,” IEEE Transactions on Aerospace & Electronics Systems, 1993, 29(3):841-856.

[10] Redmon J., Divvala S., Girshick R., et al., “You Only Look Once: Unified, Real-Time Object Detection,” IEEE, 2016.

[11] Redmon J, Farhadi A, “YOLO9000: Better, Faster, Stronger,” IEEE Conference on Computer Vision & Pattern Recognition, IEEE, 2017:6517-6525.

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

[13] Bochkovskiy A, Wang C Y, Liao H, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” 2020.

[14] Lin T Y, Goyal P, Girshick R, et al, “Focal Loss for Dense Object Detection,” IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP (99):2999-3007.

[15] Girshick R, “Fast R-CNN,” Computer Science, 2015.

[16] Ren S., He K., Girshick R., et al., “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.

[17] Li Y., Chen Y., Wang N., et al., “Scale-Aware Trident Networks for Object Detection, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2019.

[18] Mahler R. P. S, “Multitarget Bayes filtering via first-order multitarget moments,” IEEE Transactions on Aerospace and Electronic systems, 2003, 39(4): 1152-1178.

[19] Mahler R. P. S., “PHD filters of higher order in target number,” IEEE Transactions on Aerospace and Electronic Systems, 2007, 43(4): 1523-1543.

[20] Mahler R. P. S., “Statistical multisource-multitarget information fusion,” Artech House, Inc., 2007.

[21] Vo B. N., Vo B. T., Phung D., “Labeled random finite sets and the Bayes multi-target tracking filter,” IEEE Transactions on Signal Processing, 2014, 62(24): 6554-6567.

[22] Du J., Lu H., Hu M., et al., “CNN‐based infrared dim small target detection algorithm using target‐oriented shallow‐deep features and effective small anchor,” IET Image Processing, 2021, 15.