Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A method of infrared multi small target detection based on MMNet

Bing zhou (zhoubingtougao2022@163.com)
Jiaozuo University
Bei Lu
Jiaozuo University
Zhigang Zhang
Jiaozuo University

Research Article

Keywords: Target tracking, Air target, Infrared multi-target, Data association, Multi feature fusion

Posted Date: December 5th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3348658/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

A method of infrared multi small target detection based on MMNet

Bing Zhou[1*], Bei Lu[1], Zhigang Zhang[2]
1.College of Information Engineering, Jiaozuo University, Jiaozuo, 454000, China,
2.College of Artificial Intelligence, Jiaozuo University, Jiaozuo, 454000, China,

Corresponding author: Bing Zhou (e-mail: zhoubingtougao2022@163.com)

ABSTRACT In practical scenarios, as the infrared detection range shrinks, the size of infrared small targets will dynamically increase, and commonly used infrared small target detection and tracking algorithms will not be able to continue to detect and track them stably. The MMNet (Multi task Matching Network) method is proposed to address the above questions. Firstly, a lightweight fully convolutional neural network based on codec architecture was designed to segment infrared images, achieve background suppression and target enhancement; Then, utilizing the significant features of infrared small targets to further suppress false alarms; Finally, an adaptive threshold method is used to separate small targets. In the network structure, multiple lower sampling layers are introduced to reduce computation and increase receptive fields. The testing of real infrared images shows that this algorithm outperforms typical infrared small target detection algorithms in terms of detection rate, false alarm rate, and operation time, and is suitable for infrared small target detection in complex backgrounds. For unreliable tracking results, a Kalman filter is used to predict the target position. The experimental results on the LSOTB-TIR (Large Scale Thermal Infrared Object Tracking Benchmark) infrared dataset show that the improved algorithm has better performance, with tracking accuracy and success rate improved by 5.7% and 4.2% respectively compared to MMNet.

INDEX TERMS Target tracking; Air target; Infrared multi-target; Data association; Multi feature fusion.

I. INTRODUCTION

Near infrared (NIR) optical positioning system is based on the theory of multi camera stereo vision to position the luminous targets in the NIR band in three-dimensional space. Because of its non-contact measurement and characteristics of working in the independent infrared band, it is widely used in surgical navigation and industrial measurement and other scenes [1]. The detection of NIR luminous targets in the field of view is a key technology in optical positioning. With the continuous development of imaging technology, the resolution and imaging rate of images are gradually improved. In the application scenarios that require high accuracy and high real-time, the system needs to be able to quickly process a large amount of image data to complete real-time detection of specific targets, which puts forward higher requirements for computing performance [2].

At present, the infrared multi-target tracking method is mainly divided into two modes: the first is the tracking method that does not rely on the detection results. The tracking process is to first mark the candidate box in the initial frame, and then extract the target features in the candidate box. According to the proposed features, use Kalman filter [3] and particle filter method [4] to find the target position in the next frame. However, in the infrared scene, the similarity between targets is high, and there is much noise in the background, so the accuracy of this method is low and it is easy to deviate from the target. With the rapid development of target detection technology, another multi-target tracking method based on detection results has become the main paradigm of current target tracking research [5]. The basic process is: the target detector first obtains the position of the target in each frame of the image through offline training, and then distributes the detection results to the existing tracks through data association technology to generate continuous tracks. According to the difference of data association algorithm, multi target tracking technology can be divided into offline tracking and online tracking. Typical offline tracking methods include conditional random field [6], min cost max flow network flow [7], maximum weight independent set [8], etc. On the one hand, this kind of method can effectively integrate the information of current frame, past frame and future frame, which is convenient to deal with trajectory drift, error matching and object occlusion; On the other hand, the use of future frames will inevitably lead to time delay and large amount of computation, which cannot meet the real-time requirements in infrared target tracking scenes. Therefore, it is less used in infrared multi-target tracking. Online tracking methods [9] include the Sort (Simple online and real-time tracking) method [10] based on linear Kalman filter. This method first uses the motion characteristics of the target to correlate the time sequence of the detection response of adjacent frames frame by frame, and then uses the Hungarian algorithm [11] to match the trajectory with the target. Because the method uses the motion characteristics of the target, it can also be applied to the infrared scene. It has been studied to improve the anti-jamming ability of data association algorithm by using multi feature fusion method [12]. Therefore, in infrared

multi-target tracking, it is proposed to use target appearance features, motion features and manual design features. Apparent features are used to describe the appearance changes of targets, which usually face problems such as illumination changes and target deformation. Common features include color histogram [13], gradient feature and depth feature [14-15]. The motion feature is used to describe the motion track of the target, which usually adopts linear motion [16] and nonlinear motion [17]. Manual design features include improved appearance model [18] and other methods; In order to improve the performance of the association algorithm, some researches have refined the data association into multi-layer computing [19-20]. Some scholars introduced Simaese Network feature and Kernel Correlation Filter (KCF) in single target tracking into multi-target feature association [20-21]. Both of them confirmed that the combination of multiple features and improved data association are conducive to improving the accuracy and stability of multi-target tracking. However, due to the limitation of network training data and the requirement of algorithm complexity [22], its application in air infrared multi-target tracking is limited.

Since 2013, deep learning technology has promoted the breakthrough development of visual object detection technology [23] In recent years, some scholars have begun to use depth learning technology to detect infrared targets [24], but the main research object is the medium scale infrared targets, and the research on small infrared targets is less Literature [25 - 26] uses two depth neural networks to suppress the background of infrared images and classify the regions of interest, respectively, so as to achieve infrared small target detection, but there is a situation of repeatedly calculating image features Literature [27] uses the full connection neural network and sliding window method to score the whole image, and then uses the constant false alarm rate algorithm to extract small infrared targets. Although only one deep neural network is used, its receptive field is limited by the size of the window, which is easy to generate a large number of false alarms in complex backgrounds.

In order to solve the problem of low SNR infrared detection, and adapt to the size change of infrared target to achieve stable tracking, this paper selects the DBT framework to design an adaptive infrared target size detection and tracking algorithm, which can sensitively perceive suspected targets through local SNR screening, and at the same time, based on the adaptive size segmentation degree of the target size in the scene, realize the detection of large size targets, Finally, the optimized Kalman model is used to predict the target motion to achieve tracking. The experiment shows that the method can effectively deal with the detection and tracking under the dynamic change of target size, and the processing speed meets the requirements of engineering applications. However, MMNet updates the target model online with fixed coefficients, and has no occlusion processing mechanism, so the tracking success rate is not high in some challenging environments. This paper proposes to dynamically set the update coefficient of the target model based on the peak sidelobe ratio, which can be learned faster when the tracking reliability is high, and can effectively avoid model pollution when challenged by occlusion. At the same time, Kalman filtering is used to predict the target position, which effectively solves the above problems and further improves the accuracy of infrared target tracking.

II. Method and principle

A.MMNet Trace Framework

Liu et al. proposed a multi task framework MMNet at the AAAI2020 meeting, which can be used to learn the discriminant and fine-grained correlation features of thermal infrared targets [8]. Figure 1 shows the MMNet [28] framework, which includes three categories of branches: sample classification branch, discriminant matching branch and fine-grained matching branch. The sample classification branch and the discriminant matching branch jointly distinguish different types of targets, while the fine-grained matching branch is used to distinguish similar targets.

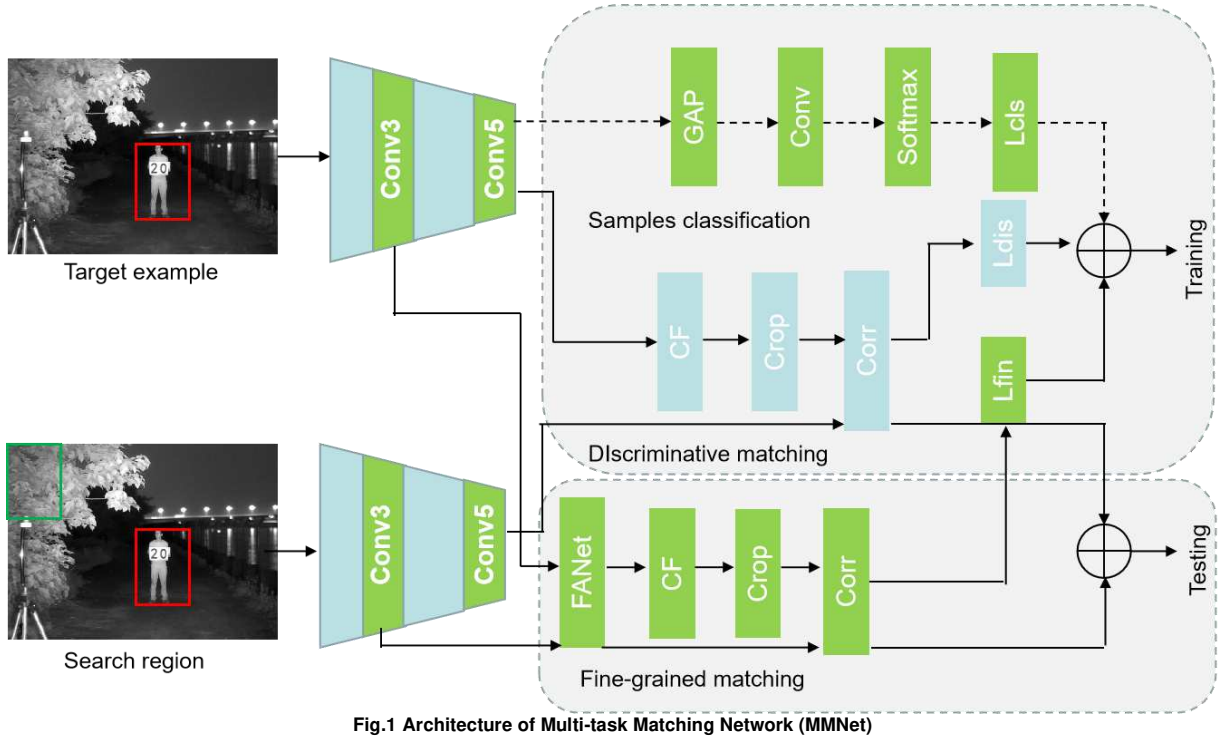

Fig.1 Architecture of Multi-task Matching Network (MMNet)

In Figure 1, convolution (convolution), GAP (global average pool), CF (correlation filter), Corr (correlation) and FANet (fine-grained sensing network) represent convolution, global average pool, correlation filter, cross-correlation and fine-grained sensing network respectively. The fine-grained perceptual network is the core of the fine-grained matching branch, which is used to learn the fine-grained correlation characteristics of thermal infrared images, the architecture of fine-grained perceptual network and the introduction of related modules.In order to distinguish different types of targets, MMNet uses CFNet [29] (correlation filter network) [9] as the basic framework to achieve discriminant matching. Compared with shallow features, deep features are more suitable for inter class discrimination because they contain more semantic information. Therefore, MNet uses the last convolution layer of AlexNet's network to achieve discriminant matching. Suppose a target template image Z ∈ R255 is given × two hundred and fifty-five × 3 and the corresponding search area image$Y \in R^{255\times255\times3}$, Then the discriminant matching can be expressed as：

$$f_{disk}(Z,Y) = g(\sigma(\phi_{conv5}(Z)),\phi_{conv5}(Y)) \quad (1)$$

Where : $\phi_{conv5}(.)$ Represents the last convolution layer feature of the extracted AlexNet network ; $g(.,.)$ Represents a cross correlation operator; σ $(\cdot)$ represents a differentiable correlation filter layer [30].

Distinguishing similar targets is the difficulty of infrared target tracking. MMNet solves this problem by obtaining fine-grained correlation features for fine-grained matching. Because the shallow features have higher resolution and contain more fine-grained feature information, MMNet uses shallow convolution features to learn the fine-grained correlation features [31]. Similar to discriminant matching, fine-grained matching can be described by the following formula.

$$f_{fin}(Z,Y) = g(\omega(\phi_{conv3}(Z))),\omega(\phi_{conv3}(Y))) \quad (2)$$

Where : $\phi_{conv3}(.)$ Represents the third layer convolution feature of the extracted AlexNet network; ω $(\cdot)$ represents the proposed fine-grained awareness network.

In order to effectively obtain the fine-grained correlation features mentioned in the previous section, MNet has designed a fine-grained awareness network FANet composed of the overall correlation module and pixel level correlation module. Figure 2 depicts the network architecture. Given a shallow convolution layer characteristic X ∈ RH × W × C. Then the fine-grained awareness network can be expressed as：

$$\omega(X) = f_c(\varphi_h(X),\varphi_p(X)) \quad (3)$$

Where : $\varphi_h(.)$ Indicates the overall related module ; $\varphi_p(.)$ Indicates pixel level correlation module ; $f_c(.,.)$ It is used to fuse the above two association features. In Figure 2, ⊙ represents the multiplication of broadcast elements; ⊗ Represents batch matrix multiplication; ⊕ Represents the addition of broadcast elements [32]。

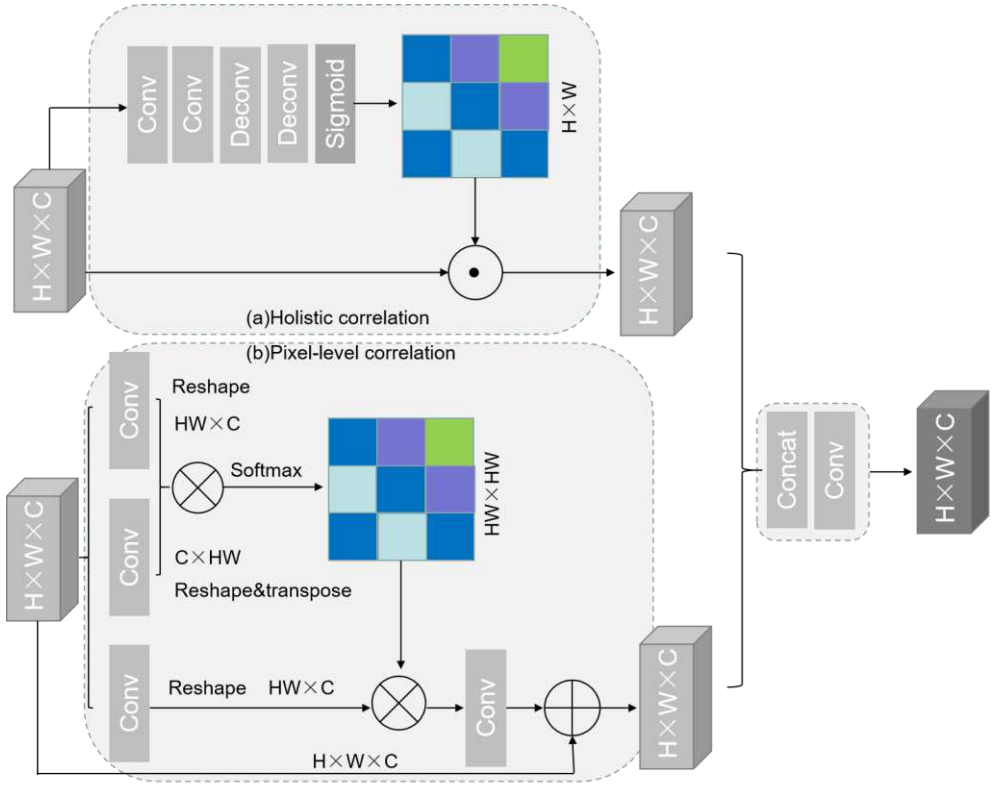

Fig.2 Architecture of Fine-grained Aware Network (FANet)

B.AC3 module with attention mechanism

Cao et al. [33] proposed a new instantiation method GCNet (Global Context Network) of the global context modeling framework in 2019, which can effectively model the global context and obtain the global context information. GCNet includes context modeling and transformation modules. In the context modeling module, first use convolution and SoftMax functions to obtain attention weights, and then pool attention to obtain global context characteristics. In the transformation module, the feature transformation is carried out through convolution, so that the model can establish the channel dependency; Finally, the broadcast addition method is used to aggregate the global context features to the features of the corresponding location to achieve feature fusion, so that the model gives greater weight to important semantic features.

Although GCNet can enable the model to adjust the channel weight of important feature information well, it does not have the ability to locate the global position feature information of the target in the face of street scenes with complex backgrounds. Therefore, inspired by the coordinate attention mechanism [34], this paper improves the GCNet attention module and proposes an attention module that is more suitable for infrared target detection in complex backgrounds. It can better obtain the position dependency and attention weight of the target in the global scope, and more accurately locate the target. Figure 3 shows the improved attention module. In the module, the global average pooling in different directions is introduced and coded separately to enhance the sensitivity of the model to target feature channels in different directions. In the module of acquiring target location information, the following steps are adopted to extract and fuse location features. The first step is to pool the input feature map horizontally and vertically. One direction is used to capture the remote dependency of target location features, and the other direction can retain accurate target location information. The second step is to splice these feature information into channels to aggregate the location remote dependencies in different spatial directions and generate independent location feature mapping. In the third step, the aggregated features with direction perception and location information are coded, and then the channel is cut to re divide them into location feature information in different directions and coded again. The fourth step is to fuse the direction sensing and position sensitive attention map with the original feature map, so as to embed the global position information of the target into the feature channel, so that the model can quickly locate the target. In the last step, the feature map embedded with position information is input into the context modeling and transformation module to adjust the channel weight. In order to make the attention module lightweight, the convolution is used in both the target location information acquisition module and the conversion module to reduce the number of channels in the feature map, which can significantly reduce the number of parameters. The improved attention mechanism introduces average pooling in different directions to embed the location information of the target in the feature channel, so that the model can more easily capture the long-distance dependency between targets, greatly improve the model's ability to locate targets, and effectively suppress the impact

of complex backgrounds on detection performance, which is very important for target detection tasks. It is worth mentioning that in the face of relatively conventional street scene target detection tasks, although there is less background feature information, the detection results are also inevitably interfered by the background information. Therefore, when performing target detection tasks in general scenes, it is also necessary to add global average pooling in different directions to the attention mechanism to enhance the model's ability to locate targets, Reduce the impact of background information on detection performance.

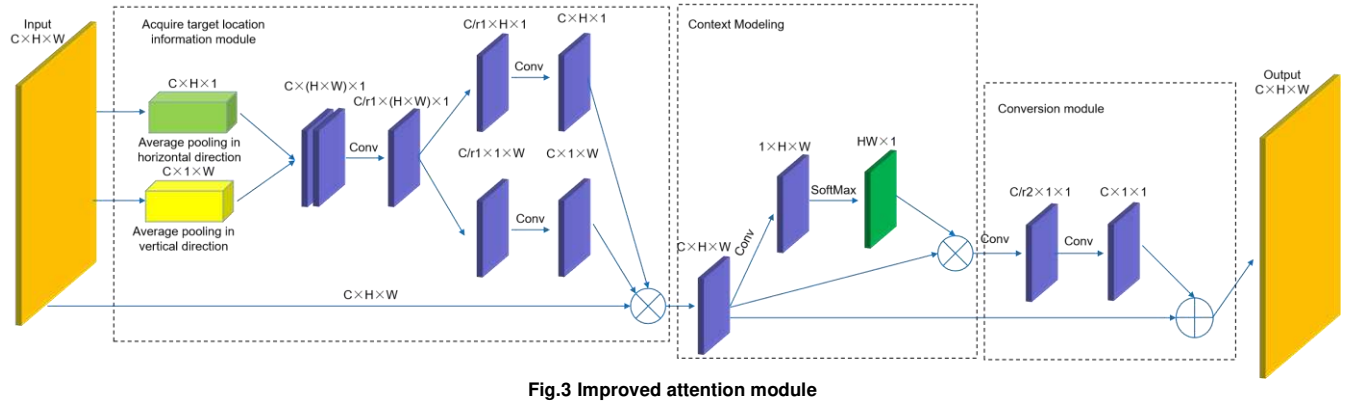

Fig.3 Improved attention module

In order to verify the effectiveness of the improved attention mechanism for street scene target detection task, this paper uses the improved attention module to improve YOLOv5 network. C3 residual block is the most important feature extraction module in YOLOv5 [35]. It is distributed in the model skeleton (Backbone) and neck (Neck), composed of multiple residual structures, and integrates feature information at different scales, which is related to the overall feature extraction ability of the model and plays a crucial role. In order to make the improved attention module better adjust the feature channel weight and reduce the parameter amount, this paper only adds the attention module to the convolution block at the end of the C3 residual block, and names the new module AC3. The specific introduction position of the attention module is shown in Figure 4, which is between the normalization and the SiLU activation function. Conv1, Conv2, Conv3, and Conv4 all represent ordinary convolution blocks, including standard convolution, normalization, and activation functions. BN represents batch normalization, Attention represents the Attention module, and SiLU represents the SiLU [24] activation function.

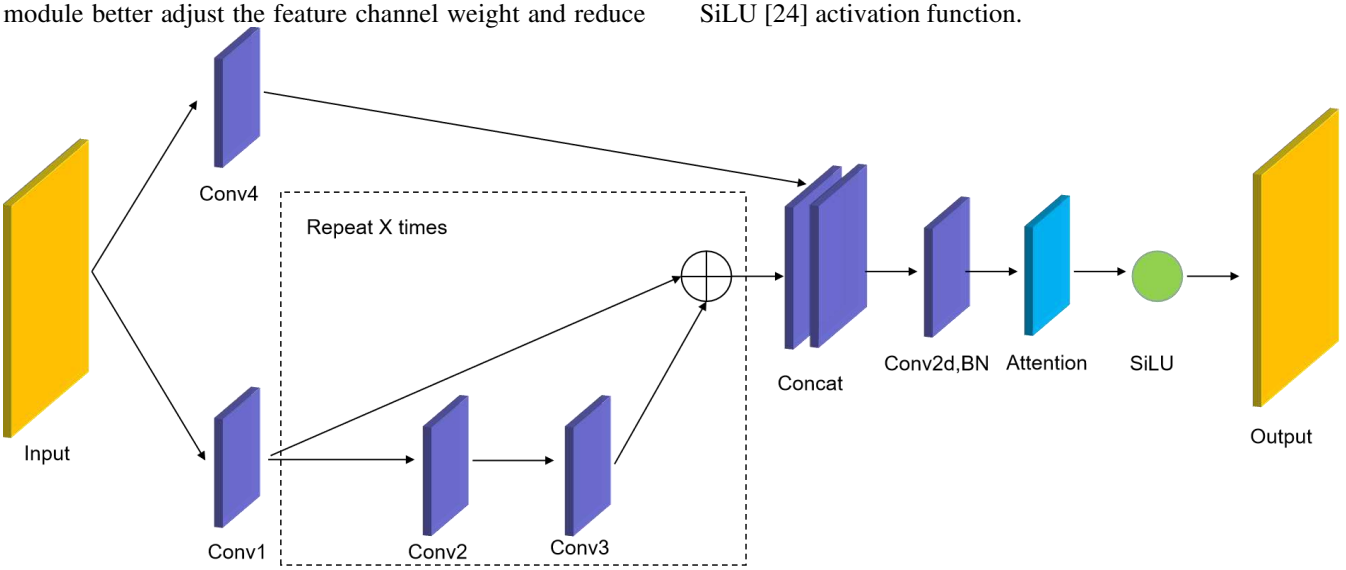

Fig.4 Attention C3 module

C.Adaptive detection and tracking algorithm for infrared target size change

In this paper, DBT detection and tracking framework is selected, and low SNR threshold filtering is used to establish suspect target database. Key features of suspect targets are extracted through region growth based on gray level, and the target database is sorted based on the SNR and other features of suspect target database to simplify the target database; For the problem that the target size changes dynamically and may be wrongly detected as the background, this paper decomposes the target that exceeds the local SNR detection window size through adaptive size segmentation, and then re detects and corrects the suspected target library. On this basis, this paper uses Kalman filtering to predict the short-term displacement of the target, and updates the target track database in real time through

key information matching between target frames, so as to effectively achieve target tracking. The algorithm flow chart is shown in Figure 5.

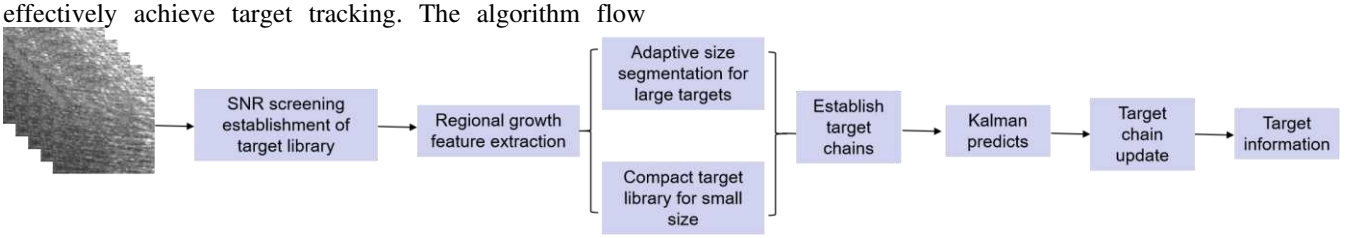

Fig.5 Our algorithm flow chart

In the pre-processing process, in order to avoid the reduction of efficiency caused by a large number of sorting operations, this paper selects mean filtering to initially suppress the impact of white noise, and in order to reduce the impact of blind elements in infrared data, the proportion of filter center core is appropriately reduced. The adjusted convolution core is shown in Formula (4).

$$core = \frac{1}{17}\begin{bmatrix} 2 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 2 \end{bmatrix} \tag{4}$$

The infrared small and weak targets in complex scenes are characterized by low contrast and low signal-to-noise ratio. It is not effective to detect targets in a single frame image by relying on commonly used shape, grayscale, gradient and other features. It can be observed that the performance of infrared small and weak targets in complex backgrounds tends to highlight the brightness and color features in local areas, and the noise changes violently [36]. Therefore, this paper refers to human visual attention mechanism, That is, human vision in complex scenes will be attracted by local areas with sharp changes in brightness, color, etc. The signal to noise ratio that can sensitively perceive the brightness changes in infrared scenes is selected as the feature basis for target preliminary screening.

First, the sliding window is used to traverse the image to extract the local SNR. The center of the window is used as the preselected target position, and other pixels in the window are used as the background by default. In order to ensure the calculation efficiency and avoid target undetected due to too large window, the sliding window size is set to 5 × 5. The target position is 5 pixels of the central area (3 pixels in a single direction). Then calculate the target signal-to-noise ratio, and judge whether the position is a suspected small target by comparing the brightness and other characteristics of the preselected position pixel under the local background, that is, the ratio between the difference between the target gray level and the background average gray level and the background gray level root mean square, as shown in Formula (5).

$$SNR = \frac{\mathrm{Gray}_{target} - Gray_{bg}}{MSE_{Gray}} \tag{5}$$

Where: Gray_{target} is the target gray scale; Gray_{bg} is the average gray level of the background; MSE_{Gray} is the root mean square of the background gray scale. There is also a crucial parameter selection in the SNR screening, namely the SNR threshold, which is related to the number of suspected targets segmented. The commonly used SNR threshold value is 2 or 3. For this paper, we choose to test under the sea surface looking down background. When the threshold value is 3, 178 suspected targets are extracted. When the threshold value is 2, 1267 suspected targets are extracted. The extraction of too many suspected targets brings about a decline in computational efficiency and an increase in false alarm rate. Therefore, this paper sets the SNR threshold value as 3, and on this basis, we sort the extracted SNR of suspected targets, Record the SNR and position information of the first 20, middle 20 and last 20 suspected targets, and add them to the target library [37], as shown in Figure 6.

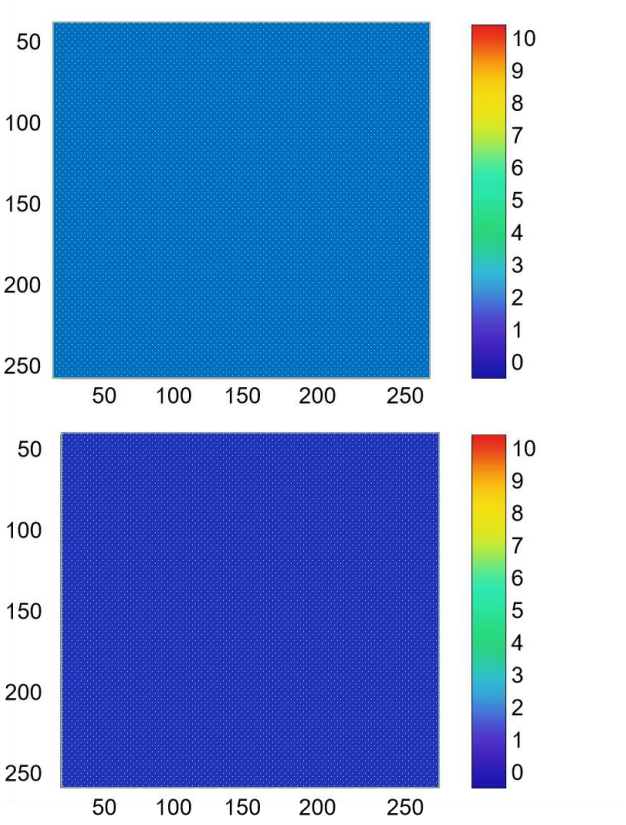

Fig 6 Original Signal to Noise Ratio (top) and Suspected Target after Segmentation (bottom)

Grayscale is a significant feature that distinguishes infrared targets from background. Therefore, after the initial screening target database is constructed through the signal-to-noise ratio, the key information of the target is further extracted with the help of gray area growth. Take the position

of the initial screening target centroid as the center, and select the size of 5 × 5. The default window edge is the background. In order to reduce the influence of false alarm, the gray threshold is set to 50% of the average gray level of the background and the maximum gray level interval, and the pixel whose gray level is greater than the gray level threshold in the window is extracted as the target. When the target position is at the edge of a certain direction of the window [38], expand the edge of the window in that direction. If the size of the target in that direction is larger than the maximum target size in the target library, discard the suspected target; If the size is not exceeded, re select the grayscale threshold for the expansion window, extract the target pixel position and its key features (size, grayscale, real signal-to-noise ratio, etc.), and add them to the suspected target database.

Different algorithms and systems have different boundary definitions for the size, energy strength and other characteristics of infrared targets. In order to explain the rationality of the definition of target size in this paper, we analyzed it from the theoretical level and verified it by experiments. Take 6 m first × The target plate of 6 m is input as the target size [39], and the image size of this data set is 256 × 256 is the detector specification, and the field of view angle matching is 10 ° × 10 ° (meeting the detection performance requirements of imaging size more than 1 pixel at a distance of 5 km).

The image size of targets at different distances calculated from far and near is shown in Table 1. When the target is 1.5 km or more away from the detection system, the image size changes more acutely with the distance change. The number of pixels per 500 m interval increases by more than 1 pixel, which has exceeded the category of weak point targets whose size is generally within the range of 1 to 3 pixels. Therefore, this paper defines the image as 5 × Those with more than 5 pixels are large-sized targets, which can not only contain weak point targets with the size of 1-3 pixels, but also reserve space for algorithm switching and downsampling with the growth of target imaging pixels [40].

TABLE 1 IMAGING PIXELS AT DIFFERENT DISTANCES

Distance/km	Imaging pixel dimensions	Pixel growth
5	1.76×1.76	-
4.5	1.94×1.94	0.18
4	2.22×2.22	0.28
3.5	2.51×2.51	0.29
3	2.93×2.93	0.42
2.5	3.52×3.52	0.59
2	4.4×4.4	0.88
1.5	5.9×5.9	1.5
1	8.8×8.8	2.1
0.5	17.5×17.5	8.7

In order to further illustrate the applicability of the size definition in different scenes, this paper takes the complex cloudy, sea antenna and sea surface overhead background data sets as the experimental input, and the detection gates are defined as 3 × 3、5 × 5、8 × 8 and 12 × 12, that is, the size target boundary, and the test results are shown in Table 2.

TABLE 2 EXPERIMENTAL RESULTS OF CORRESPONDING DETECTION GATES UNDER DIFFERENT SIZE DEFINITIONS

Gate size/1	Complex and cloudy background				Top view background of sea surface				Sea antenna background			
	Detection tracking rate/%	Undetected rate/%	False alarm rate/%	Frame rate/Hz	Detection tracking rate/%	Undetected rate/%	False alarm rate/%	Frame rate/Hz	Detection tracking rate/%	Undetected rate/%	False alarm rate/%	Frame rate/Hz
3×3	98	2	7.4	55	90	10	7	51	91.8	8.2	5	56
5×5	98.5	1.5	7	50	94.1	5.9	2	48	93.5	6.5	3	52
8×8	98	2	6	44	95	5	5	43	92	8	3.2	46
12×12	94.4	5.6	15	31	91.2	8.8	12	29	88.6	11.4	4	31

In the test scenario, due to the adoption of 3 × 3. The detection range of the size gate is small. When the SNR is initially screened in the full image detection, it is easily affected by the flicker point noise in the scene, resulting in a false alarm rate slightly higher than 5 × 5 Gate detection. At the same time, in the sea surface looking down multi-target scene, the size of a target that appears for the first time exceeds 3 × 3. The algorithm did not downsample in time, resulting in a high rate of missed detection. In addition 3 × Although the processing efficiency of 3 size gate detection is improved, it will lead to more frequent downsampling, resulting in limited frame rate increase. twelve × The size of 12 is firstly that the number of single wave gate processing pixels is increased to 144, which greatly affects its processing efficiency; Secondly, as the processing range of single wave gate becomes larger, weak point targets of 1~2 pixels will be mistakenly detected as flickering noise, resulting in missing detection; Finally, during the initial screening of the full picture signal-to-noise ratio, the sea clutter in the sea surface overhead background and small clouds moving with the background in the cloudy scene will be wrongly detected as targets, resulting in a high false alarm rate. eight × 8 Dimension detection performance and 5 × 5 The size is similar, but the processing frame rate decreases slightly due to the increase in the number of pixels processed by the single wave gate. Therefore, for the data set in this paper where the size of weak point target gradually increases from 1 to 3 pixels, the definition is greater than 5 × The size of 5 pixels is a large target, using 5 × 5 wave gate detection is appropriate; For other data sets, it is necessary to redefine the size and set the detection gate according to the initial size and change rule of the target and the actual scene noise.

IV.Experiment

A.Experimental environment

We implemented the proposed target tracking algorithm based on MMNet algorithm framework using Matlab2016b programming language. In order to verify the effectiveness of the algorithm proposed in this paper, experiments were carried out on a PC equipped with I7-10700 2.9 GHz CPU and TX-1080 GPU, and MMNet, HCF (Hierarchical

Collaborative Features) [12], HSSNet, MCCT (Multi Cue Correlation Filters for Robust Visual Tracking) [13], MLSSNet, CFNet SRDCF (Spatial Regulated Discriminatory CorrelationFilters) [14] and other seven advanced trackers were compared.

This algorithm uses Stochastic Gradient Descent (SGD) for network training. The training adopts the same mixed training mode as MMNet, that is, the thermal infrared image data set and the visible light image data set [15] are mixed as the final training set. In terms of test set selection, this paper selects LSOTB-TIR [16] as the test data set. LSOTB-TIR is a large-scale and highly diversified thermal infrared universal target tracking data set proposed in 2020, which includes four common application scenarios in the civil field: hand-held camera shooting scenarios, vehicle shooting scenarios, video surveillance shooting scenarios, and unmanned aerial vehicle shooting scenarios, covering five categories of targets: animals, pedestrians, vehicles, ships, and aircraft. Compared with other infrared data sets, LSOTB-TIR almost covers the target categories of interest in thermal infrared target tracking. In this paper, 25 typical infrared target video sequences are selected from the above four common application scenario datasets in civil fields for testing, and the average tracking speed of the proposed algorithm is about 19.2 FPS.

B.Convergence rate, thermodynamic diagram and loss analysis

In order to study the dynamic impact of different improvement methods on model detection results, Figure 7 shows the dynamic impact of different improvement methods on mAP (IoU=0.5) with the increase of training times. It can be seen intuitively from the figure that the detection results of any improved method are faster than that of the original YOLOv5s model. When the three improved methods are added at the same time, the detection effect is the best and the convergence speed is the fastest. This is due to the addition of SSPPF and attention module, which enhances the feature extraction ability of the model and makes the model more focused on target information. In addition, the reprogramming of BN also alleviates the performance degradation of the model. It can be seen from the figure that Strip YOLOs has achieved a good detection effect when training to about the 100th time, which can significantly reduce the number of training.

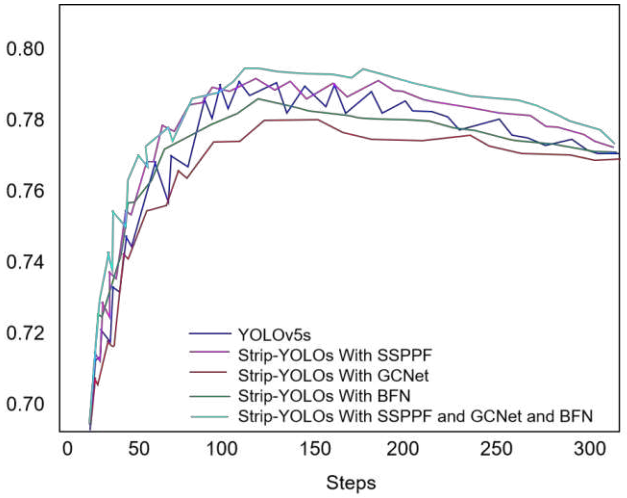

Fig.7 Dynamic influence of different improvement methods on mAP (IoU=0.5)

In order to more intuitively reflect the effectiveness of the improved method, inspired by the class activation mapping CAM (Class Activation Mapping) [28], this paper, based on the confidence level of the target, draws the thermal maps in the detection head used to detect targets of different scales (small target 80x80, medium target 40x40, large target 20x20). As shown in Figure 8, the first and second lines show the visualization results of small, medium and large target feature maps of YOLOv5s and Strip YOLOs detection heads. It can be seen from the visualization results that Strip YOLOs pays attention to more semantic information (such as pedestrians on the right and cars on the left in the figure). This is because the pooling core used in the SSPPF module is bar shaped and no longer a regular shape, which creates a global dependency between isolated objects in the image, and reduces the loss of pedestrian and bicycle features and the pollution of semantic information between objects, The feature information between targets is no longer jumbled together, which enhances the robustness of the model. In addition, the PPM in the MPM module can also extract local semantic information between dense targets (such as densely distributed pedestrians on the right side of the figure), which makes the model rich in more semantic information. Finally, the addition of attention mechanism also makes the model more accurate in positioning the target, and pays more attention to the feature information of the target. The above further proves the effectiveness of the improvement method in this paper.

Fig.8 The thermal maps of the feature maps for the detection of different size target in the detection head based on the confidence degree

The loss function is very important for the deep learning model. It reflects the internal relationship between the predicted value and the real value, and directly affects the quality of the model training. The smaller the loss function, the closer the predicted value is to the true value, and the better the detection performance of the model. The loss composition of Strip YOLO is the same as that of YOLOv5, including confidence loss, classification loss and boundary box positioning loss. Figure 9 shows the impact of different improvement methods on model loss. It can be seen from the figure that the improved Strip YOLOs model has less loss in the verification set than YOLOv5s. This proves that Strip YOLs extracts more semantic information, reduces the loss of feature information in the image, and its predicted value is closer to the real value.

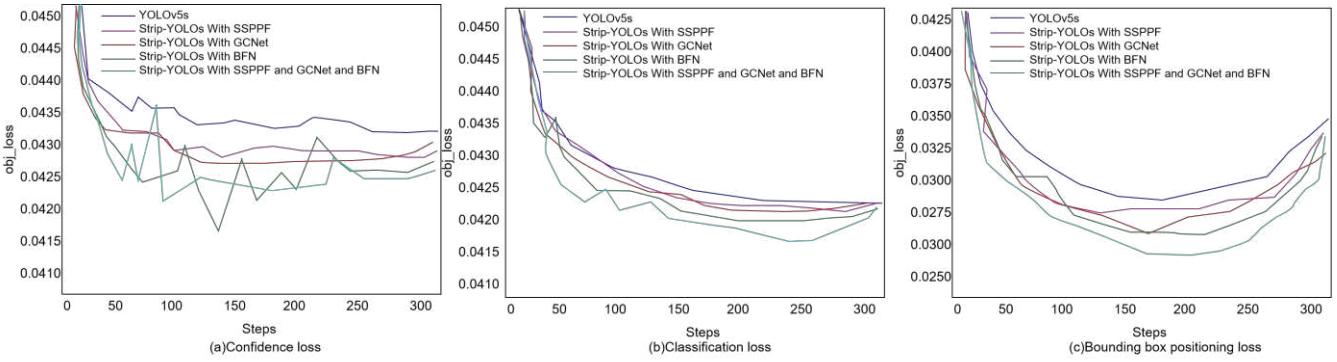

Fig.9 Loss comparison of different improved methods on FLIR verification set

C.Contrast experiment

In order to more objectively reflect the impact of different improved methods on model detection performance, this paper conducts a comparative experiment on Strip YOLOs on FLIR dataset. In this paper, the test results of YOLOv5s are taken as the baseline, and the effects of different improvement methods on the detection performance, convergence rate, thermodynamic diagram of the characteristic diagram of YOLOv5s are recorded and analyzed respectively. 3.2.1 Effectiveness of different improvement methods Table 3 shows the impact of different improvement methods on YOOv5s detection results, and Table 4 shows the impact of different improvement methods on the detection accuracy and recall rate of various categories of targets. 1) Effectiveness of SSPPF. In order to reduce the problem of feature pollution and feature loss caused by rule pooling operation for the model, this paper will use the SSPPF module reconstructed by strip pooling to join the backbone network, which reduces the feature loss of the model for long and narrow targets and the feature pollution between unrelated targets. At the same time, the model will establish a long-term dependency relationship in the whole office, and enrich the global relationship between targets. The first group of comparative experiments in Table 3 shows the comparison of detection performance between YOLOv5s and SSPPF module only. It can be seen that when the module is added to SSPPF module alone, the model's mAP (IoU=0.5) increases by 0.9% and F1 increases by 2.1%, which proves that the addition of SSPPF module can make the model better adapt to the changes of target shape, enrich more feature information and reduce feature loss, especially for narrow pedestrians and bicycles. The first group of comparative

experiments in Table 3 also verified this conclusion. It can be seen from the first and second rows in Table 4 that the recall rate of the model has increased significantly after the SSPPF module is added, especially for relatively narrow pedestrians (the recall rate has increased by 0.7%) and bicycles (the recall rate has increased by 3.6%). This again proves that the addition of SSPPF can effectively reduce the missed detection rate of the model for long and narrow targets. In addition, because PPM in MPM focuses important features on smaller feature maps, and uses addition to the original features for feature fusion, which increases the weight of target features, the accuracy of the model is also improved accordingly. 2) Effectiveness of AC3. In order to make the model locate the target more accurately and reduce the impact of complex background on the detection performance, this paper embeds the direction sensing information and global target location dependent features obtained by different direction pooling operations into the channel, which makes up for the lack of global target location information in GCNet, and enhances the model's ability to locate the target. The second group of comparative experiments in Table 2 shows the comparison of detection performance between the model without attention module and the model with attention module. It can be seen that the mAP (IoU=0.5) of the model with attention module is improved by 0.3% and F1 is improved by 1.6%. This is because the model is more sensitive to the location information of the target due to the addition of global direction pooling, and the spatial dependency between global target locations is established. The second group of comparative experiments in Table 3 once again supports this conclusion. From the experimental results in the first and third lines, we can see that after adding the AC3 module, the accuracy of the model is further improved (the average accuracy of the three categories is improved by 2.8%), which shows that the addition of the AC3 module can enable the regression box generated by the model to more accurately locate the target rather than the background. 3) Effectiveness of BN distribution re planning. In order to alleviate model degradation and improve detection performance, GN is used to re plan the distribution of normalization operations in the model, which blocks the accumulation of BN estimation offsets. The first and fourth rows in Table 2 show the detection performance comparison of YOLOv5s after the BN distribution is not re planned and the BN distribution is re planned. It can be seen that the mAP (IoU=0.5) of the model is improved by 0.8% and F1 is improved by 1.4% after the re planning, which fully proves that the re planning of BN effectively alleviates the performance degradation of the model and improves the detection performance of the model. The comparative experimental results in the first and third rows in Table 3 verify this result. It can be seen from the results that after the BN distribution is re planned, the accuracy and recall of the model have been greatly improved (the accuracy of bicycles and cars have been improved by 0.4% and 1.1% respectively, and the recall of pedestrians and bicycles have been improved by 1.7% and 3.8% respectively). It is worth mentioning that since only part of BN is replaced with GN after re planning, the detection speed and parameter quantity of the model are almost the same as those of the original model, which will not affect the detection speed of the model.

TABLE 3 EFFECTS OF DIFFERENT MODIFIED METHODS ON MODEL DETECTION PERFORMANCE

SSPPF	AC3	BFN	AP(IoU=0.5) People/%	Bicycle/%	Car/%	mAP(IoU=0.5)/%	F1/%	Time/ms	Params/M	FLOPs/B
			82.6	63.4	90.4	78.8	75.6	10.8	7.1	16.0
√			84.4	64.0	90.5	79.7	77.7	14.2	7.3	18.6
	√		83.3	63.6	90.4	79.1	77.2	16.8	7.9	16.6
		√	83.9	64.8	89.9	79.6	77.0	11.3	7.1	15.9
√	√	√	84.8	67.1	90.5	80.7	78.0	20.4	8.1	19.3

TABLE 4 THE IMPACT OF DIFFERENT IMPROVEMENT METHODS ON THE ACCURACY AND RECALL OF VARIOUS CATEGORIES OF TARGETS

SSPPF	AC3	BFN	P/% People/%	Bicycle/%	Car/%		R/% People/%	Bicycle/%	Car/%	R/%
			84.9	74.9	82.1	80.6	73.0	55.3	85.4	71.2
√			86.9	78.0	84.5	83.2	73.7	58.9	85.6	72.8
	√		85.8	78.9	85.6	83.4	73.6	56.6	85.0	71.8
		√	84.9	75.3	83.2	81.3	74.7	59.1	85.3	73.1
√	√	√	86.2	82.2	87.1	85.2	73.0	58.9	83.7	71.9

D.data set

Seven algorithms are tested on the test data set. The ROC curve is used to compare the detection performance of various algorithms. The results are shown in Figure 10 The ROC curve of our algorithm is located at the top left of other algorithms, which shows that its detection performance is better than other algorithms When the false alarm rate is 3.2 × At 10-6, the detection rate of this algorithm has exceeded 92%, while the detection rate of other algorithms is not higher than 84%. As the detection threshold decreases, the false alarm rate and detection rate of various algorithms gradually increase Among them, the detection rate of ADMD, MaxMedian, and FM - DNN algorithms has an obvious growth trend, but when the detection rate exceeds the algorithm in this paper, the false alarm rate has been greater than 6.9 × 10-5, more than 21 times of the algorithm in this paper 50 sheets 128 are selected in the test set × 128 infrared images. Then the algorithm in this paper and six comparison

algorithms were tested for multiple rounds of speed. Finally, the average time of each algorithm to process a single image was obtained, as shown in Table 5 It can be seen that the algorithm in this paper has the shortest operation time.

TABLE 5 AVERAGE RUN TIME OF DIFFERENT ALGORITHMS

Method	ADMD	IPIM	LIG	MaxMedian	FMDNN	FCN	S-FCN
Times/s	0.996	0.866	0.284	0.219	0.066	0.164	0.033

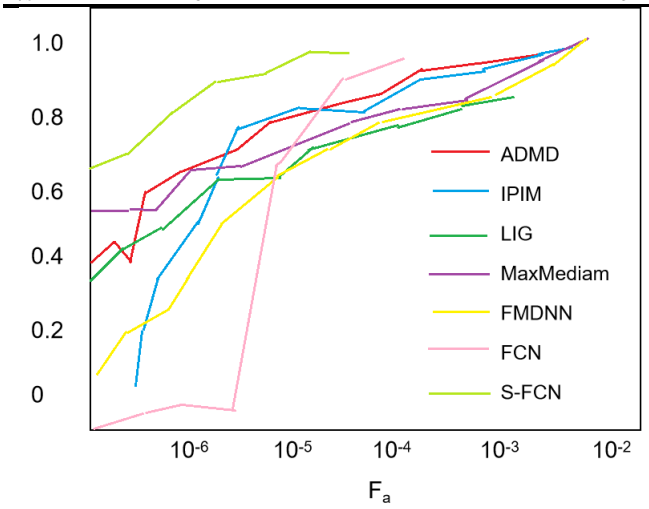

Fig. 10 ROC curves of different algorithms

V.Conclusion

In this paper, an idea based on hierarchical data association is proposed to solve the problem of air infrared multi-target tracking. First of all, based on the characteristics of infrared scene, this paper proposes to construct the similarity of three features: appearance, motion and scale. Then a hierarchical data association method is proposed. The experiment shows that Strip YOLO algorithm has excellent performance on FLIR dataset, especially for long and narrow targets in infrared images. The comparison experiment verifies the effectiveness and rationality of various improvement methods. In addition, Strip YOLO is also lightweight. The improved module only adds a few parameters, ensuring the ability of real-time detection. In the future, we will continue to optimize the main feature extraction part of the model, improve the feature extraction ability of the model for small targets or targets that are difficult to detect, and at the same time ensure the real-time detection ability of the model. Experiments show that the algorithm in this paper can reach the capacity level of the traditional DBT algorithm in small and weak target scenes, and the processing speed has been improved slightly. For the scene where the size of the target selected in this paper changes dynamically and there are large targets, The detection and tracking rate of this algorithm is improved by about 10% compared with the traditional algorithm.

Funding Statement

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Ethical Statement

This manuscript does not involve human or animal ethics.

Conflict Of Interest (COI) Statement

The authors have no relevant financial or non-financial interests to disclose.

Funding Statement

This work was supported by Training plan for young backbone teachers in Henan Province （No.2018GGJS267）.

Ethical Statement

This manuscript does not involve human or animal ethics.

Conflict Of Interest (COI) Statement

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Du Ziwei, Zhou Heng, Li Chengyang, Li Zhongbo, Xie Yongqiang, Dong Yuchen, Qi Jin. Overview of Small Target Detection Algorithms for Deep Convolution Neural Network [J]. Computer Science, 2022,49 (12): 205-218

[2] Moradi S, Moallem P, Sabahi M F. A false-alarm aware methodology to develop robust and efficient multi-scale infrared small target detection algorithm[J]. Infrared Physics & Technology, 2018, 89: 387-397.

[3] Guo Mingtao Research and application of target detection model based on improved YOLOv3 [D]. Nanjing University of Posts and Telecommunications, 2022.

[4] Su J, Xing M, Wang G, et al. High-speed multi-target detection with narrowband radar[J]. IET radar, sonar & navigation, 2010, 4(4): 595-603.

[5] Cheng Wenshao, Fan Qiang, Zou Erbo. Design of target detection hardware accelerator based on convolutional neural network [J]. Optics and Optoelectronics Technology, 2022,20 (05): 108-114.

[6] Wang Jiachen, Wu Yifeng. Target detection method based on complex value depth neural network [J]. Signal Processing, 2022,38 (10): 2021-2029.

[7] Mendes A, Bento L C, Nunes U. Multi-target detection and tracking with a laser scanner[C]//IEEE Intelligent Vehicles Symposium, 2004. IEEE, 2004: 796-801.

[8] Guo Wenjuan, Feng Quan, Li Xiangzhou. Research progress of convolution neural network model based on crop disease detection and recognition [J]. China Agricultural Machinery Chemistry Journal, 2022,43 (10): 157-166.

[9] Yang Yi, Fu Zefeng, Gao Youjin, Cui Kefei, Wang Keping. Video target detection of fully mechanized mining face based on deep neural network [J]. Industrial and Mining Automation, 2022,48 (08): 33-42.

[10] Yang Liu, Li Zhiyu, Sun Dai, Li Yuexing. Research on Radar Target Detection Algorithm Based on Neural Network and Opencv [J]. Mathematical Practice and Cognition, 2022,52 (08): 104-114.

[11] Harkat H, Ruano A E, Ruano M G, et al. GPR target detection using a neural network classifier designed by a multi-objective genetic algorithm[J]. Applied Soft Computing, 2019, 79: 310-325.

[12] Gao Jie. An Iterative Mechanism to Improve Target Detection Based on Convolution Neural Network [J]. Control Engineering, 2021,28 (12): 2469-2477.

[13] Ayoun A, Smets P. Data association in multi‐target detection using the transferable belief model[J]. International journal of intelligent systems, 2001, 16(10): 1167-1182.
[14] Wei Zhifei, Song Quanhong, Li Fang, Yang Qingyu, Wang Aihua. Research on target detection algorithm based on neural network model compression technology [J]. Aerospace Defense, 2021,4 (04): 107-112
[15] Wang Yadong, Tian Yonglin, Li Guoqiang, Wang Kunfeng, Li Dazi. Overview of 3D target detection based on convolutional neural network [J]. Pattern Recognition and Artificial Intelligence, 2021,34 (12): 1103-1119.
[16] Zhu Q, Ma K, Wang Z, et al. YOLOv7-CSAW for maritime target detection[J]. Frontiers in Neurorobotics, 2023, 17.
[17] Huang L, He M, Tan C, et al. Retracted: Jointly network image processing: multi‐task image semantic segmentation of indoor scene based on CNN[J]. IET Image Processing, 2020, 14(15): 3689-3697.
[18] Zhang Shao Research on target detection algorithm based on convolutional neural network under laser active imaging [D]. University of Chinese Academy of Sciences (Changchun Institute of Optics, Precision Machinery and Physics, Chinese Academy of Sciences), 2021.
[19] Wang Chunlei Research on high-speed maneuvering target detection based on deep neural network [D]. Xi'an University of Electronic Science and Technology, 2021. DOI: 10.27389/d.cnki.gxadu.2021.003101
[20] Zeng N, Wu P, Wang Z, et al. A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-14.
[21] Li Guangshuai, Li Xiang, Tan Xiaofeng, Yan Jiwei SAR Image Target Detection Based on Convolutional Neural Network [C]//. National Security Geophysical Series (16) - Big Data and Geophysics, 2020:132-139.
[22] Ou Xianfeng, Yan Pengcheng, Wang Hanpu, Tu Bing, He Wei, Zhang Guoyun, Xu Zhi. Research on moving target detection method based on depth frame difference convolution neural network [J]. Journal of Electronics, 2020,48 (12): 2384-2393
[23] Cao Zhe, Cao Yucong, Li Jianfei, Xu Chaofan, Tan Qimeng. Research on space station cabin target detection based on convolutional neural network [J]. Journal of Northwest Polytechnical University, 2020, 38 (S1): 84-89
[24] Cong Longjian, Liu Yanxin, Jin Songzhi, Hao Mengxi, Liu Yanyangshuo, Zhou Bin, Zhang Hui. A convolutional neural network design for small-scale target detection [J]. Journal of Northwest Polytechnical University, 2020,38 (S1): 146-153
[25] Wang L, Zhu W, Zhang Y, et al. Multi-target detection and adaptive waveform design for cognitive MIMO radar[J]. IEEE Sensors Journal, 2018, 18(24): 9962-9970.
[26] Shi N, Li M, Deng Y, et al. Experimental demonstration of a multi-target detection technique using an X-band optically steered phased array radar[J]. Optics Express, 2016, 24(13): 14438-14450.
[27] Lin T H H, Lu T, Braun H, et al. Optimization of a multi-stage ATR system for small target identification[C]//Automatic Target Recognition XX; Acquisition, Tracking, Pointing, and Laser Systems Technologies XXIV; and Optical Pattern Recognition XXI. SPIE, 2010, 7696: 605-614.
[28] Fang Chen Research on lightweight method of target detection network model [D]. Tianjin University of Science and Technology, 2020. DOI: 10.27359/d.cnki.gtqgu.2020.000819
[29] Tian Zhuangzhuang Research on target detection method of remote sensing image based on convolution neural network [D]. University of National Defense Technology, 2020.
[30] Wang Yanan Research on salient target detection algorithm based on depth neural network [D]. University of Chinese Academy of Sciences (Xi'an Institute of Optics and Precision Machinery, Chinese Academy of Sciences), 2020.
[31] Song Jiaxing, Liu Qingwei, Luo Zhe, Yu Fan. Vehicle detection system based on deep neural network [J]. Transmission Technology, 2019, 33 (04): 3-9
[32] Li T, De la Prieta Pintado F, Corchado J M, et al. Multi-source homogeneous data clustering for multi-target detection from cluttered background with misdetection[J]. Applied Soft Computing, 2017, 60: 436-446.
[33] Jiao Jiali Several researches on target detection algorithms based on SVM and RPCA [D]. Nanjing University of Posts and Telecommunications, 2019.
[34] Lan T Y, Bendory T, Boumal N, et al. Multi-target detection with an arbitrary spacing distribution[J]. IEEE Transactions on Signal Processing, 2020, 68: 1589-1601.
[35] At dusk Research on abnormal target detection and recognition algorithm of coalbed methane well station [D]. Dalian University of Technology, 2019. DOI: 10.26991/d.cnki.gdllu.2019.004301
[36] Bai Jing, Moving target detection method based on depth convolution neural network Shaanxi Province, Xi'an University of Electronic Science and Technology, December 1, 2019
[37] Meng Fansheng Indoor object detection and attitude estimation based on enhanced self encoder [D]. Hangzhou Normal University, 2019.
[38] Swallow K M, Jiang Y V, Riley E B. Target detection increases pupil diameter and enhances memory for background scenes during multi-tasking[J]. Scientific reports, 2019, 9(1): 5255.
[39] Zhou Su, Zhi Xuelei, Liu Tong, Ning Hao, Jiang Lianxin, Shi Fanhuai. Small target traffic sign detection algorithm based on convolutional neural network [J]. Journal of Tongji University (Natural Science Edition), 2019, 47 (11): 1626-1632
[40] Li Z, Santi F, Pastina D, et al. Multi‐frame fractional Fourier transform technique for moving target detection with space‐based passive radar[J]. IET Radar, Sonar & Navigation, 2017, 11(5): 822-828.