

A One-stage Infrared Dim Small Target Tracking Method by Fusing Spatiotemporal Information

Da-Wei Li^{1*}, Xin-Yu Yang¹, Chen-Hui Cui², and Su-Zhen Lin²

¹ College of Electricity and Control Engineering, North University of China, 030051, P.R. China
lidawei@nuc.edu.cn, Yangxinyu210@163.com

² College of Data Science and Technology, North University of China, 030051, P.R. China
h2392636132@163.com, lsz@nuc.edu.cn

Received 1 January 2025; Revised 27 April 2025; Accepted 28 April 2025

Abstract. The susceptibility of infrared dim small targets to background clutter during motion causes existing Siamese network tracking methods to fail to track the target consistently and stably. In light of this, this paper puts forward a means that integrates the spatio-temporal information of targets, and reconstructs the Siamese network tracking framework for the task of tracking infrared dim and small targets. The deficiencies of the original two-stage Siamese network trackers are remedied by this framework. By using deep learning networks, it simultaneously accomplishes feature extraction and target matching, attaining multi-layer feature maps. As the various layers within the multi-layer feature maps concentrate on diverse aspects, a channel and spatial attention approach is adopted to fuse these feature maps with respect to semantics and particulars. Finally, to further improve tracking accuracy, the multi-head attention mechanism is combined with the Convolutional Neural Network. In an anchor-free form, it directly predicts target classification and regression scores to complete the tracking task. The realization results show that the proposed infrared dim small target tracking method has better performance on the public infrared dim small target tracking dataset, it achieved a success rate with an Area Under the Curve of 58.7% and an accuracy rate reaching 78.6%.

Keywords: machine vision, object tracking, infrared dim small target tracking, Siamese network, multi-head attention

1 Introduction

Infrared dim small target tracking technology is one of the important application technologies of Infrared Search and Track System (IRST) [1], and it is also widely used in the fields of missile detection, aerospace, remote sensing and space monitoring [2]. Due to the special characteristics of infrared dim small targets, for instance, the target dimensions typically range from 2×2 to 9×9 pixels [3], the target lacks edge and texture features, and the background environment is complex and changeable during the motion process, which makes it difficult for the tracker to obtain discriminative feature map, thus causing tracking offset problems [4]. Overcoming these challenges requires the design of dim small target tracking methods capable of extracting discriminative feature maps. Among the tracking methods for visible light targets, the Siamese network-based tracking method has gained popularity among researchers because it can satisfy the effective balance of accuracy and speed [5]. Existing infrared dim small target tracking methods also adopt this tracking framework, obtaining better tracking results relative to DCF tracking methods and meeting the requirements of engineering applications. Under the interference of similar distractors and background clutter, the existing single-target tracking methods are unable to effectively extract the real target, resulting in poor tracking performance. In situations such as target intersections, the disappearance and reappearance of targets, and target flickering, the existing multi-target tracking methods often match the wrong target ID numbers, causing errors in trajectory association and leading to tracking failures.

Qian [6] et al. utilized the style recalibration and multi-style feature module based on the SiamRPN [7] tracking method in a natural light scene, thus enhancing the ability of the network to conduct feature extraction on infrared dim and small targets. In addition, the utilization of the side-window filter has restrained the impact exerted by most background clutter on infrared dim and small targets. Shan [8] et al. relying on the single-stage

* Corresponding Author

Siamese network tracking framework OTrack [9], adopted multi-head attention [10] as the key module of the network. In the course of feature extraction, they established a bidirectional information exchange between the template frame image and the search frame image for the purpose of getting better tracking results. Li [11] et al. ameliorated the SiamFC [12] tracking method with the integration of the convolutional channel attention mechanism, the stacked channel attention mechanism [13] and the spatial attention mechanism [14], which led to an improvement in the tracking performance. Chen [15] et al. with the aim of coping with the challenges brought about by dim, small and rapidly moving targets, initially established a re-detection model predicated upon a two-stage Siamese network, so as to mitigate the interference of the background on infrared dim and small targets. Thereafter, in an effort to resolve the issue where the tracking method would lose the target on account of its rapid movement, they put forward a three-stage global re-detection mechanism for re-detecting the target, thereby realizing robust tracking. Lv [16] et al. with the intention of enhancing the tracking accuracy of infrared dim and small targets, put forward an efficacious Siamese network tracking approach. This approach encompasses three modules, specifically the feature extraction module, the feature fusion module, and the multi-head attention module. It is capable of extracting target features with high efficiency, establishing the contextual correlation of targets, and facilitating a more effective differentiation between targets and the background. To address the issues that traditional convolutional cross-correlation networks are prone to getting trapped in local optima and have relatively low tracking accuracy, Cui [17] et al. proposed an infrared dim and small target tracking method based on the Siamese network and Transformer. This method utilizes the Histogram of Oriented Gradients (HOG) feature map to expand the target information in the deep feature map and employs the multi-head attention mechanism in Transformer to search for the area where the target is located in the feature map of the search frame, ensuring robust tracking of the target. The existing target tracking methods mainly aim at the problem of relatively low target tracking accuracy under background interference, which leads to the inability to distinguish well between the real target and the similar target in the case of interference from similar objects. The tracking method is likely to deviate, and then the tracking fails. Therefore, how to extract discriminative feature maps and response maps under the interference of similar objects and fully correlate the effective target features is an urgent problem to be solved.

This paper puts forward a one-stage tracking methodology for infrared dim and small targets by integrating spatio-temporal information. This methodology efficaciously exploits the spatio-temporal information of targets within the ambit of the motion process and synchronously undertakes the target matching operation during the feature extraction stage, thus facilitating the procurement of multi-layer feature maps with augmented discriminative prowess. By availing itself of the channel and spatial attention mechanisms, it deftly amalgamates the multi-layer feature maps with assorted information, further intensifying the informational plenitude of the feature maps. Additionally, by leveraging the combination of multi-head attention and Convolutional Neural Network (CNN), the classification and regression confidence maps are yielded, which endows the predicted bounding boxes of the tracking method with greater precision. The experimental results evince that the proposed tracking method attains a superior level of tracking accuracy in tracking environments featuring complex and variable backgrounds. The primary contributions are expounded as follows:

- 1) Harness the one-stage Siamese network to consolidate the procedures of feature extraction and target matching, and capitalize on the spatio-temporal information of target displacement to augment the discriminative capacity of the feature maps.
- 2) By capitalizing on the channel and spatial attention mechanisms, an effective integration of the multi-scale feature maps is carried out, resulting in an enrichment of the information encompassed by the feature maps.
- 3) Through the integration of multi-head attention and CNN network for the classification and regression of trackers, the accuracy of the corresponding feature maps is enhanced.

The primary content of this paper is illustrated as follows. The first part elaborates on the problems that infrared dim and small targets face and the current research status. The second part comprises a comprehensive and detailed analysis of the overall proposed tracking method and each of its modules.

In the third section, the comparison of dim small target tracking public datasets, ablation experiment results and analysis are given. The fourth section makes the summary of this paper.

2 Method

2.1 Overall Framework

The basic ideology of the one-stage infrared dim and small target tracking approach (TSTrack) which is put forward in this paper and grounded on spatio-temporal feature correlation is as follows: 1) A one-stage target tracking framework is employed. The processes of feature extraction and cross-correlation computation during the tracking are unified, which can effectively diminish the quantity of network parameters, expedite the network inference velocity, and augment the discriminative power between the target and the background of the multi-layer feature maps generated by the network. 2) Link the spatio-temporal features of moving targets. Owing to the continuous alteration of the target's neighboring background during movement, multi-frame template images can effectively adapt to the background variation and raise tracking accuracy. 3) Multi-layer feature association: A multi-layer feature association method for infrared dim and small targets is proposed in this paper. The method associates and blends the feature maps of different hierarchies output by the one-stage network module. As a result, it can raise the information level of the feature maps, completely hold the semantic and detailed information of the targets, improve the correctness of the network's bounding box regression, and lower the possibility of tracking failure. 4) New classification and regression output module: The classification and regression output module combining the multi-head attention and CNN can get rid of the tedious operation of setting the anchor box, and output the classification and regression confidence map with long distance dependence in an anchor-free way.

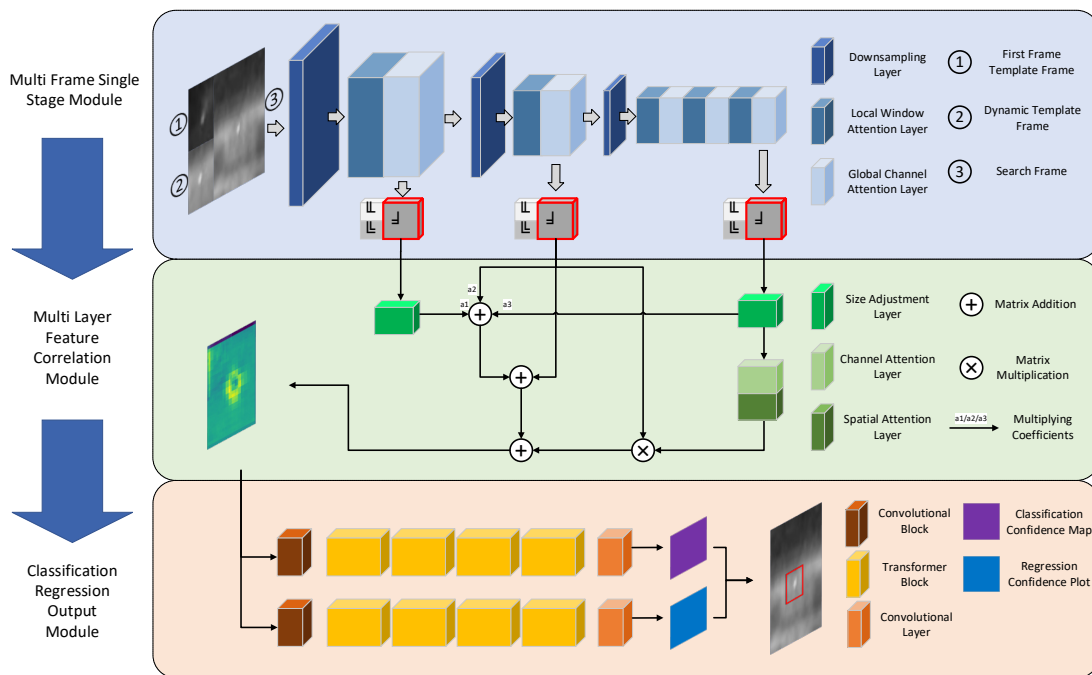


Fig. 1. Overall structure of the TSTrack tracking method

The overall framework of TSTrack tracker is shown in Fig. 1. The tracker comprises a multi-frame one-stage module, a multi-layer feature correlation module and a classification regression output module; wherein the multi-frame one-stage module and the multi-layer feature association module obtain a mutual correlation feature map, and the classification regression output module predicts the target classification probability and target location information in the feature map.

2.2 Multi-Frame One-stage Module

With the aim of strengthening the discriminative capacity of the foreground and background information within the feature map of the search frame image, a multi-frame one-stage module for unified feature extraction and cross-correlation computation is proposed, as shown in Fig. 2.

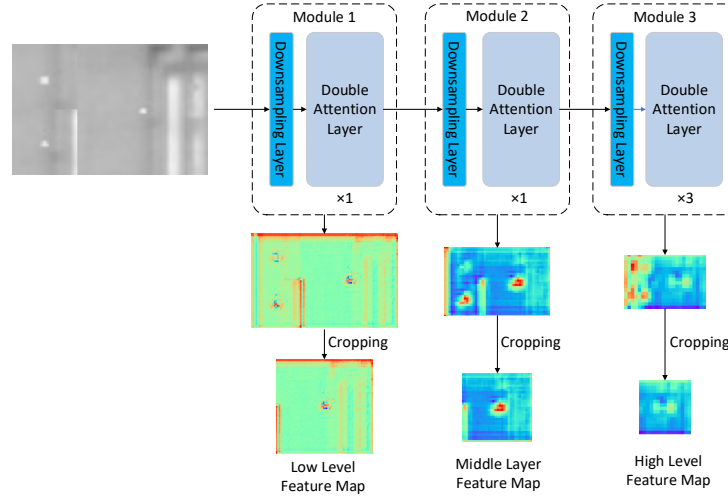


Fig. 2. Multi-frame one-stage module structure

The network has the advantages of small number of module parameters and fast inference speed. The network inputs are the first frame template frame (tracking the manually selected target and its neighborhood background image in the first frame of the sequence image), the dynamic template frame (the image that is closest to the current tracking frame image and the predicted classification confidence level meets certain requirements) and the search frame image (the local area of the current tracking image). In addition, as the network deepens, the three images can extract semantic and detailed feature information. The three-frame unified input permits simultaneous feature extraction and cross-correlation calculation for the search frame and two template frames. During network deepening, the correlation of feature details between the two template frames and the search frame features is established, strengthening the target information in the search frame's feature map, subduing the background area, and ceaselessly refining the accuracy of the network's subsequent feature extraction.

The multi-frame one-stage module is a refinement of the DaViT [18] network. It mainly boosts the window magnitude of the local window attention level in the dual attention layer to obtain local details of windows of different extents. The multi-frame one-stage module can be split into three sub-modules. Each sub-module comprises a down-sampling layer network and a dual attention layer. The dual attention layer includes a local window attention part and a global channel attention part. The network structure of the dual attention layer is shown in Fig. 4(a). Be aware that the first module contains a dual attention layer, the second module includes a dual attention layer, and the third module consists of three dual attention layers.

Multi-frame one-stage module process is: first, input the first frame template frame, dynamic template frame and search frame image and according to certain rules for local zoom cropping (two template frame image cropped to 128×128 size, the search frame image cropped to 256×256 size), cropped images in accordance with the splicing shown in Fig. 3 for the combination to obtain the combination of frame images $f_{joint} \in \mathbb{R}^{3 \times 384 \times 256}$.

Second, the combined frame image is passed through a downsampling layer network of the first module in Fig. 2. The downsampling layer network is composed of a convolutional layer and a normalization layer. The intention is to reduce the dimensions of the combined frame images, increase the channel count, further decrease the computational quantity, and improve the running speed of the tracker. Afterwards, have the obtained downsampling feature maps go through the double attention layer containing a local window attention layer and a global channel attention layer. Next, trim the output feature maps to obtain the low layer feature map $f_{low} \in \mathbb{R}^{96 \times 64 \times 64}$. Finally, the low layer feature map is processed through the second module depicted in Fig. 2 accompanied by

a cropping operation, thereby obtaining the middle layer feature map $f_{mid} \in R^{192 \times 32 \times 32}$. Subsequently, the intermediate-level feature map is fed into the third module and subjected to a cropping operation, resulting in the high layer feature map $f_{high} \in R^{384 \times 16 \times 16}$. Among them, the cropping operation mainly crops and deletes the feature maps mapped by the two template frames, and only retains the feature map region mapped by the search frame.

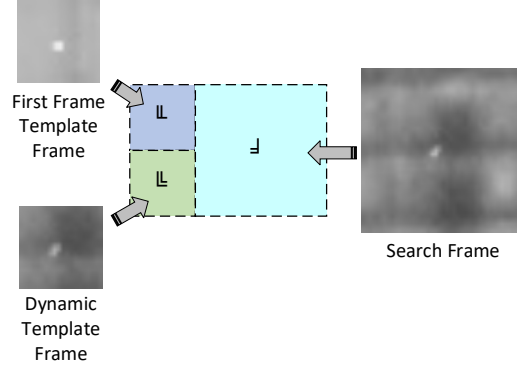


Fig. 3. Schematic diagram of splicing mode

As shown in Fig. 4, it's the network structure of the double attention layer. Fig. 4(a) showcases the detailed network structure of the dual-attention layer. Fig. 4(b) unfolds the detailed network structure of the local window attention layer, and Fig. 4(c) discloses the detailed network structure of the global channel attention layer. The process of double attention is as follows: after inputting the downsampling feature map f_{down}^i of the i module, it passes through the local window attention layer, the feedforward neural network layer, the global channel attention layer, and the feedforward neural network layer in series, and passes through each network layer by a residual connection operation to avoid the loss of information. The local window attention layer and the global channel attention layer will be analyzed in the following.

For the local window attention layer, the input downsampling feature map f_{down}^i is first flattened in wide and high dimensions and passed to the Layer Norm for normalization. Then, the normalized feature vector is restored to the size of the feature map. The restored feature map is divided into four equal parts according to the channel dimension, and input into four Reshape layers respectively. It facilitates the operation of multi-head attention within the local window (the window partition sizes of the first module are 4×4 , 8×8 , 16×16 and 32×32 , the window partition sizes of the second module are 2×2 , 4×4 , 8×8 and 16×16 , and the window partition sizes of the third module are 1×1 , 2×2 , 4×4 and 8×8 , respectively). The formula is as follows:

$$f_{reshape}^i = \text{Norm}(\text{flatten}(f_{down}^i)) \quad (1)$$

$$f_{n \times n}^i, f_{2n \times 2n}^i, f_{4n \times 4n}^i, f_{8n \times 8n}^i = \text{clip}(\text{reshape}(f_{reshape}^i)).$$

In the formula, the operation of flattening the width and height dimensions of the feature map is represented by *flatten*; *Norm* represents the Layer Norm normalization layer; *reshape* represents the operation of restoring the width and height dimensions; *clip* represents the window partitioning operation; $i=1,2,3$ represent different modules in a multi-frame one-stage module; $f_{n \times n}^i, f_{2n \times 2n}^i, f_{4n \times 4n}^i, f_{8n \times 8n}^i$ denotes four different scales of local feature maps generated after window partitioning operation, where $n=4$ for module 1, $n=2$ for module 2, and $n=1$ for module 3.

The principle of segmenting the local window attention layer based on window size is founded on the size of the targets in the two template frames and one search frame image. Consider the first module. After down-sampling, the target size is within 16×16 . Window sizes of 4×4 and 8×8 can extract the local characteristics of the target effectively, and a 16×16 window size can extract the global features of the target well. A 32×32 window size can extract the features of the target and the neighboring background effectively, thus enhancing the generalization power of the network model.

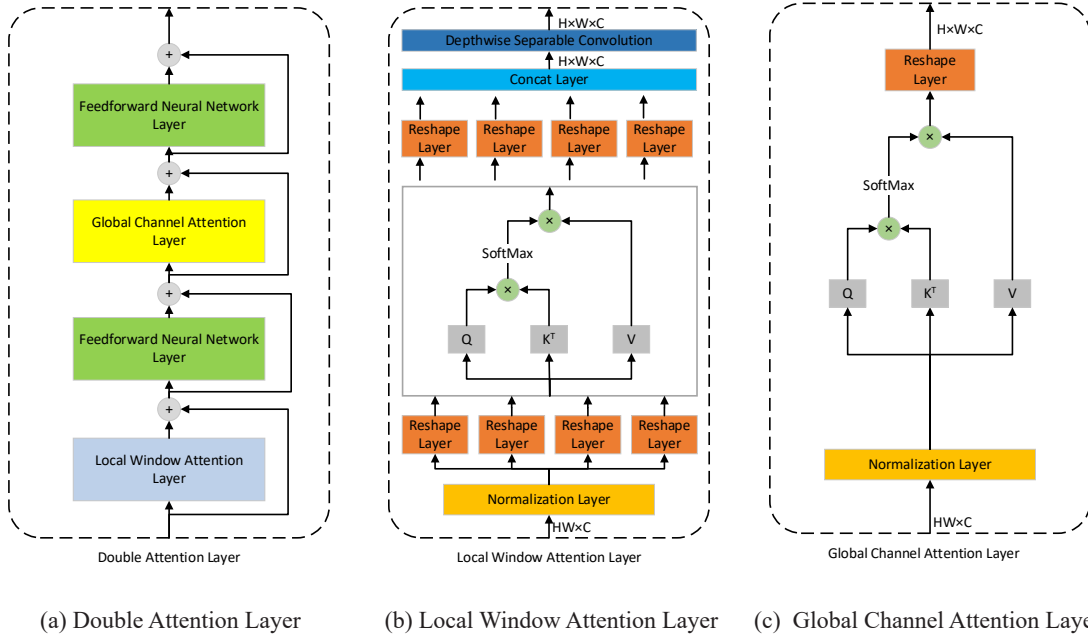


Fig. 4. Double Attention Layer, Local Window Attention Layer and Global Channel Attention Layer Network architecture

After dividing the window, the four local feature maps $f_{n \times n}^i, f_{2n \times 2n}^i, f_{4n \times 4n}^i, f_{8n \times 8n}^i$ are respectively flattened and inputted into the multi-head attention network in parallel, in particular, the head count of this multi-head attention mechanism network is 2, serving to extract the image features within the local window domain. The specific formula is shown hereafter:

$$f_k^i = \text{MultiHead}(\text{flatten}(f_{k \times k \times k \times n}^i)). \quad (2)$$

In the formula, *MultiHead* represents Multi-head attention; $k=1,2,3,4$.

Firstly, restore the four feature vectors, which have undergone the information interaction in the multi-head attention, to the size of the downsampling feature map through modifying the width and height in the reshape layer. Next, send them through the Concat layer and combine them in the dimension direction to acquire a feature map with the same size as the down-sampled feature map, and finally passed through a depth-separable convolution for the dimensionality direction of the information interaction, which generates the localized windowed attention feature map f_{local}^i .

The global channel attention layer is similar to the local window attention layer, the difference is mainly that there is only one Reshape layer, the global channel attention layer can carry out the information interaction of the feature map globally, and establish the long-distance information dependence. The specific process is as follows: after the feedforward neural network and residual connection, after a flattening operation, input to the multi-head attention mechanism network with 8 heads, and finally restored to the original size by the Reshape layer, to yield the global channel attention feature map f_{global}^i . After that, the feature map after dividing the window is subjected to a flattening operation and inputted into a Transformer network, which, in particular, has a head count of 2 for extracting image features within the local window; finally, the width and height are restored to the downsampling feature map size by a Reshape layer, and then passed through a Concat layer to obtain the feature map of the exact same size as the one obtained from the downsampling feature map by splicing it in the dimension direction. The size of the downsampling feature map matches that of the feature map completely. At last, by means of depthwise separable convolution for information interaction in the dimension direction, the local window attention feature map is created.

The global channel attention layer is analogous to the local window attention layer, and the key difference is that it has only one Reshape layer, which performs a flattening operation of the whole image width and height of the local window attention feature map, and after that inputs it into a Transformer network with a header number of 8, and then finally restores it to its original size through the Reshape layer. The global information interaction of the feature map is carried out to establish long distance dependency.

2.3 Multi-Layer Feature Association Module

In order to further boost the accuracy of the predicted bounding boxes in the tracking method, this paper presents a multi-layer feature association module based on the attention mechanism for associatively fusing the high, middle and low layer feature maps output from the multi-frame one-stage module, and the obtained associated feature maps have both the target's apparent detail information and deep semantic information.

As shown in Fig. 5, the high, middle and low layers of feature maps output from the multi-frame one-stage module are firstly obtained, where the high layer feature map has more semantic information about the target, the low layer feature map has more detailed information about the target, and the middle layer feature map maintains a balance between semantic and detailed information. In our paper, we consider the feature map of the middle layer to be the baseline feature map and obtain the crucial information of other layers through correlation and fusion operations. Since the scales and channel numbers of the three layers of feature maps are different, the high layer feature map first goes through a convolution block 1, and then nearest neighbor interpolation is performed to enlarge the feature map by two times, and the feature map with the same size as that of the middle layer feature map f'_{high} is output. The formula for convolution block 1 is shown below:

$$f'_{high} = \text{ReLU}(\text{BatchNorm}(\text{Conv2}(\text{Conv1}(f_{high})))) . \quad (3)$$

For this formula, *Conv1* indicates a 1×1 convolutional layer possessing 384 input channels and 1536 output channels. *Conv2* indicates a 1×1 convolutional layer with 1536 input channels and 192 output channels. *BatchNorm* symbolizes the batch normalization layer, and *ReLU* symbolizes the activation function layer. The goal of the convolutional block is to prevent information loss to the greatest extent during the dimension reduction of high-level feature maps. First, increasing the dimension before decreasing it can enhance the generalization performance of the feature map and retain the channel features possessing essential information.

The low layer feature map passes through a convolution block 2, which performs both dimensional scaling up and scaling down to output a feature map f'_{low} of the same size as the middle layer feature map. The formula for convolution block 2 is as follows:

$$f'_{low} = \text{ReLU}(\text{BatchNorm}(\text{Conv}(f_{low}))) . \quad (4)$$

In the formula, *Conv* represents a 2×2 convolutional layer with 96 input channels, 192 output channels and a step size of 2.

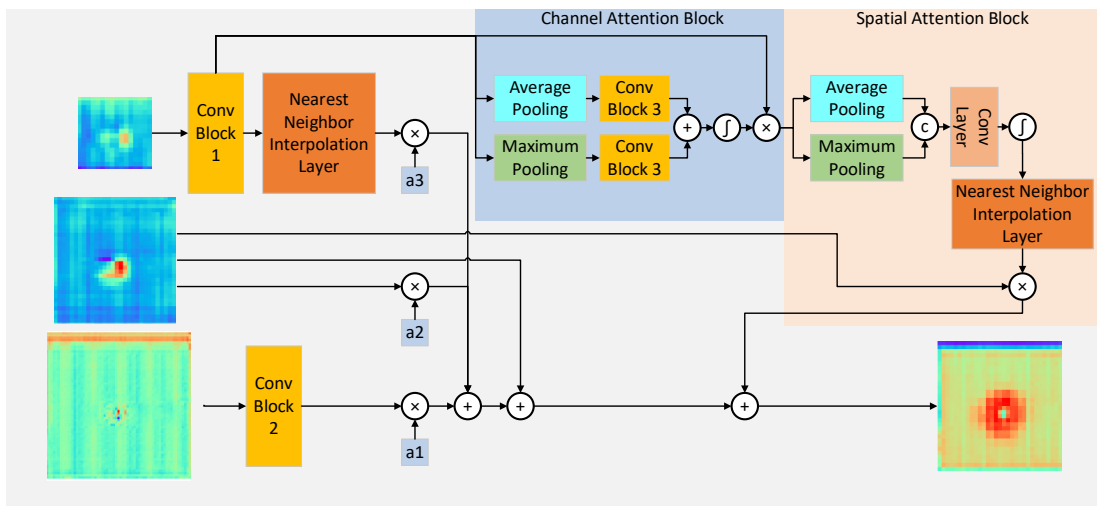


Fig. 5. Multi-layer feature association module network structure

After that, f'_{high} , f'_{low} and f'_{mid} are multiplied and summed with the weights of the learnable parameters, and then summed with to obtain the multi-information feature map, as shown in the following equation:

$$f'_{mid} = f_{mid} + (\alpha_1 f'_{high} + \alpha_2 f'_{mid} + \alpha_3 f'_{low}). \quad (5)$$

In the formula, α_1 , α_2 and α_3 are learnable parameters, participating in the overall training of the tracker.

From a theoretical perspective, high layer feature maps possess greater amounts of semantic information and can effectively tell targets apart from backgrounds. Consequently, mapping the high-scoring zones in high layer feature maps to middle layer feature maps can improve the discriminatory ability of the intermediate-level feature maps. Shown in Fig. 5, firstly, a channel attention block filled with effective information is added to the high layer feature map to derive a channel feature map. The formula goes as follows:

$$M_c(f'_{high}) = \sigma(\text{ConvB}(\text{AvgPool}(f'_{high})) + \text{ConvB}(\text{MaxPool}(f'_{high}))) \times f'_{high}. \quad (6)$$

In this formula: σ refers to the Sigmoid function, *AvgPool* refers to the average pooling layer, and *MaxPool* refers to the maximum pooling layer, after the above operation, the feature map changes from $192 \times 16 \times 16$ to $192 \times 1 \times 1$, *ConvB* represents the convolution block 3, and the operation flow of the convolution block 3 is as follows: the feature map following pooling first traverses a 1×1 convolutional layer, where the number of channels is reduced to 48 by this layer. and then passes through a ReLU activation function layer and a 1×1 convolution layer to raise the number of channels to 192.

Later on, the channel feature map will pass through the spatial attention module. In this module, the feature map after the Sigmoid operation will proceed through the nearest neighbor interpolation layer, and then the feature map will be interpolated and scaled up to twice its original dimensions. In addition, the interpolated feature map will also be multiplied with the middle layer feature map. The multiplication operation can enhance the target area in the middle layer feature map and improve the discrimination ability of the middle layer feature map. This is given by the formula as follows:

$$M_s(M_c(f'_{high}), f'_{mid}) = \text{Inter}(\sigma(\text{Conv}([\text{AvgPool}(M_c(f'_{high})), \text{MaxPool}(M_c(f'_{high}))]))) \times f'_{mid}. \quad (7)$$

In this formula, σ refers to the Sigmoid function, *Inter* refers to the nearest neighbor interpolation layer, *AvgPool* refers to the average pooling layer, and *MaxPool* refers to the maximum pooling layer, and the feature map f'_{high} is changed from $192 \times 16 \times 16$ size to $1 \times 16 \times 16$ size after passing through these two layers, $[\ast, \ast]$ represents the Concat operation, and the Conv has a convolution kernel sized 3×3 , and the padding for the convolution layer is set at 1. The dimension of the feature map after the Concat process is shrunk to 1.

Finally, the output feature map is summed with the multi-information feature map f'_{mid} to obtain the final association feature map f_{Merge} .

2.4 Classification Regression Output Module

The traditional CNN classification regression output module is limited by the sensory field and cannot establish long-distance dependencies in the feature map, while Transformer can extract key information in the global range of the feature map and establish dependencies within the target area, between the target and the background, and between the background and the background, thereby enhancing the prediction precision of the classification and bounding box regression network module. Based on this concept, this chapter puts forward a classification and regression output module that combines CNN with Transformer. The classification regression output module of this chapter is shown in Fig. 6.

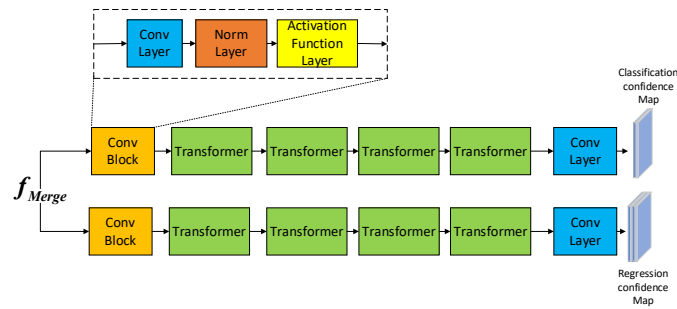


Fig. 6. Structure of the categorical regression output module

As an example, the classification network module in its entirety is composed of a convolutional block, four Transformer networks and a convolutional layer. The convolutional block consists of a convolutional layer, a BatchNorm normalization layer and a ReLU activation function layer. The Transformer network has a structure similar to what's shown in Fig. 4(c). The final convolutional layer applies a 1×1 convolutional kernel and a Conv2D convolution with an output dimension of 2. Differing from the classification network module, the output dimension of the final convolutional layer of the regression network module is 4.

3 Experimentation and Analysis

The experimental configuration is detailed in subsection 3.1, covering the chosen training and test datasets as well as experimental particulars. In subsections 3.2 and 3.3, qualitative and quantitative experiments are carried out with two contemporary infrared dim small target tracking techniques and three generalized target tracking methods to deliberate and dissect the merits and drawbacks of the methods proposed in this paper. Moreover, in subsection 3.4, ablation experiments and analysis are executed to exhibit the efficacy of each module within the overall methodology.

3.1 Experimental Setup

To objectively appraise the performance of the proposed tracking method, the publicly available infrared dim and small target datasets are adopted for training and testing. The training dataset originates from the “infrared dim small motion target detection dataset in complex background” [19], and the testing dataset originates from the “infrared image dim small aircraft target detection and tracking dataset in ground/air background” [20]. During the testing process, the same test data as SiamIST was maintained and the first four video sequences in the test dataset, which contained larger sized or multiple infrared targets, were deleted.

The comparison methods include SiamIST [6], SiamITO [8], Stark [21], TOMP [22], and TransT [23] tracking methods. Among them, the first two are infrared dim small target tracking methods and the last three are generalized target tracking methods.

The codes of all the methods presented in this paper were composed in Python 3.8 by using Pytorch 1.8.0. All codes were trained on a Linux server equipped with Inter Xeon (32GB, CPU), NVIDIA A5000 (24GB, GPU); and tested on a Windows 11 laptop equipped with Inter i7 12700H (16GB, CPU), NVIDIA RTX 3060 laptop (6GB, GPU). The request for codes can be made by email.

3.2 Qualitative Comparison

In this subsection, six video sequences are selected for demonstration, including videos 21, 20, 15, 14, 8, and 5, as shown in Fig. 7.

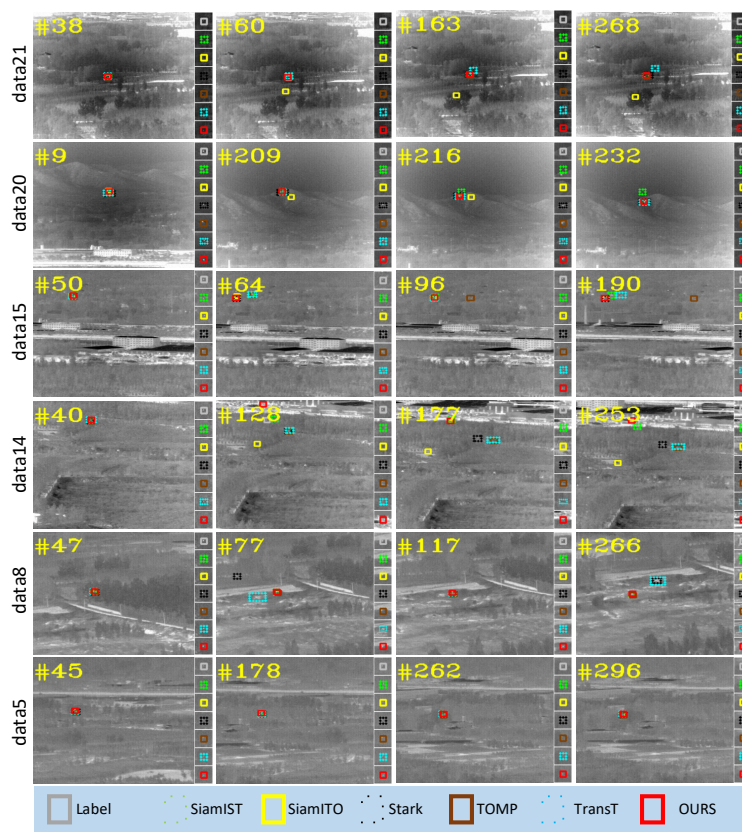


Fig. 7. Display of tracking results of different tracking methods

In the video sequence “data21” in Fig. 7, the overall gray value of the target keeps changing and showing flickering in the image, which leads to the loss of the real target by the SiamITO tracking method in the 60th frame image. In addition, as the target passes through the forest environment with low gray value, the target is close to submerged in the background. For instance, the real target was lost by the TransT tracking method in the 163rd frame image. In the 268th frame image, in addition to the tracking failures of the above two tracking methods, although the Stark tracking method tracked the real target, the estimated size of its bounding box was too big.

In the video sequence “data15” in Fig. 7, the target is displaced in the 64th frame due to infrared camera shake, which leads to the tracking failure of the SiamIST, TOMP and TransT tracking methods, but in the subsequent movements of the target, such as the 96th frame, the SiamIST and TransT tracking methods accurately track the real target again, and only the TOMP tracking method determines the background as the real target. However, in the subsequent motion of the target, such as the 96th frame, the SiamIST and TransT tracking methods accurately track the real target again, and only the TOMP tracking method determines the background as the real target tracking failure. At frame 190, multiple similar interfering objects appear around the target, causing the SiamIST and TransT tracking methods to shift to the false target, and the TOMP tracking method still fails to track it.

In the “data14” video sequence in Fig. 7, due to infrared camera shake, the target is shifted for a long distance, as in frame 128, and all tracking methods fail. However, in the subsequent tracking process, such as frame 177, the SiamIST tracking method and the proposed method track the real target again. The target crosses the background of urban buildings, and whenever there is a large discrepancy between the target and the background, only the method presented in this paper can track the actual target anew. This indicates that the proposed method can better obtain the feature map with discriminative power and reduce the influence of background clutter on the tracking method.

In the “data8” video sequence within Fig. 7, the target is extremely blurred as a result of the jitter of the infrared camera, although it does not produce a large displacement, resulting in the tracking failure of Stark and TransT tracking methods in the 77th frame. However, when the lens is stabilized and the target is imaged clearly, as in the 117th frame image, the two methods of tracking failure track the real target again. During the subsequent target motion, similar interfering objects appear around the target, resulting in the tracking failure of Stark and TransT again at frame 266.

In the “data5” video sequence in Fig. 7, the target does not have dramatic appearance or gray value changes, and the background does not have large changes, and all tracking methods track the real target.

The comparative experiments have shown that the method put forward in this paper can precisely track the real target in many complex background settings, except when there are extreme background interferences and camera jitters.

3.3 Quantitative Comparison

In this subsection, comparisons were made on the performance of different tracking methods by means of the success rate graph and the precision rate graph, as shown in Fig. 8. In addition, the area under the Area Under the Curve (AUC) of the success curve values and the accuracy rate values with a threshold of 10 are further compared as shown in Table 1.

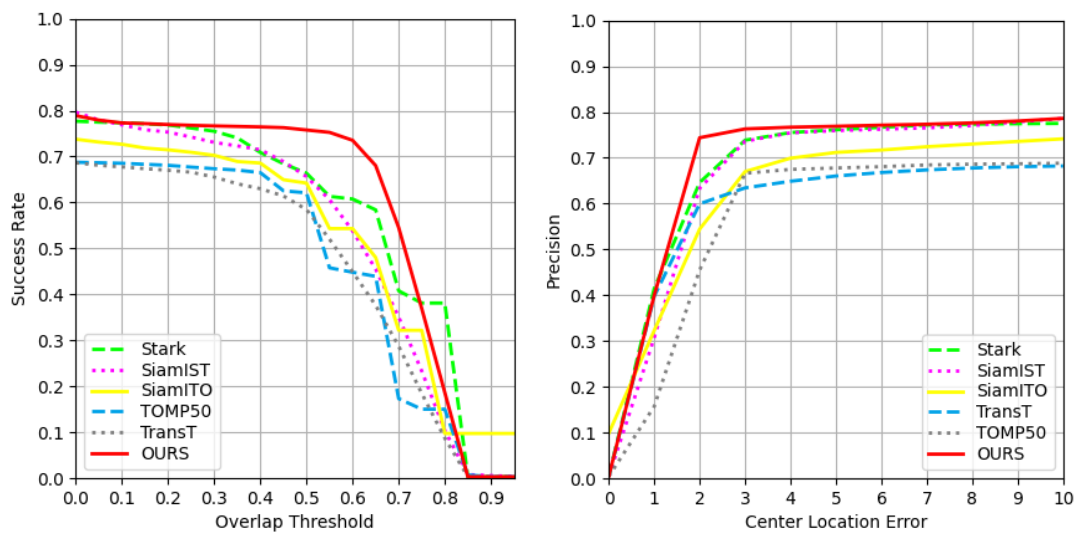


Fig. 8. Success rate and accuracy curves

Table 1. Comparative results of quantitative experiments

	Success ratio	Precision (threshold=10)
TransT	0.455	0.682
TOMP	0.460	0.688
Stark	0.558	0.775
SiamIST	0.521	0.786
SiamITO	0.515	0.741
Ours	0.587	0.786

Among these items, the success rate and the precision rate are expounded as follows. Success rate is figured out as the ratio of the quantity of images where there is an intersection and a union between the predicted bounding box of the tracking method and the target-labeled bounding box smaller than the specified threshold to the total frame number of the video sequence.

Precision rate represents the ratio of the number of pictures in which the Euclidean distance from the center point position of the predicted bounding box of the tracking method to that of the target-labeled bounding box is under the specified threshold to the total number of frames in the video.

From Table 1, it can be seen that the tracking method proposed in this paper can track infrared dim small targets robustly, and the AUC value is improved by 6.6 percentage points and the predicted target bounding box is more accurate relative to the SiamIST tracking method, although the accuracy rate remains the same; relative to the Stark tracking method, the AUC value has gone up by 2.9 percentage points, and the accuracy rate has risen by 1.1 percentage points.

Furthermore, with the aim of further exploring the performance of the tracking method put forward in this paper in respect of inference time, the inference velocity of this method on the test laptop is roughly 22 Frames Per Second (FPS), and it takes about 0.044 seconds to process one frame of image. On a workstation with an NVIDIA RTX 3090 GPU and an Intel Xeon Platinum 8362 CPU, the inference speed is approximately 25 FPS, and it takes about 0.04 seconds to process one frame of image. This demonstrates that on high-performance computing platforms, the method can basically meet the requirements for real-time tracking.

3.4 Ablation Study

In this subsection, for the purpose of verifying the efficacy of each component module of the proposed method, we executed ablation studies on various modules.

Among them, the ablation of the multi-frame one-stage module can be specifically divided into four parts: (1) Ablation of multi-frame: the dynamic template frames input to the tracking method are replaced with the first template frames, and the spatiotemporal characteristics of the target are not introduced into the tracking method. (2) Ablation of the improved DaViT network: the original DaViT network is used as the multi-frame one-stage module in this chapter. (3) Ablation of the one-stage framework: the multi-frame one-stage model is only used as a feature extraction network, and the inter-correlation computation stage adopts convolutional inter-correlation. (4) Overall ablation of the multi-frame one-stage module: the multi-frame, the improved DaViT network and the one-stage framework are all ablated.

The ablation method of the multi-layer feature association module is that after the high layer feature map has completed the convolution of block and interpolation operations and the low layer feature map has completed the convolution with the block, the two feature maps are directly added to the intermediate-level feature map.

The method for ablating the classification and regression output module is to use a convolutional neural network in place of the multi-head attention network.

In Table 2, upon comparing 7 with 8, with the application of the classification and regression output module proposed in this chapter, the AUC has gone up by 3.7 percentage points and the precision has increased by 4.0 percentage points. It can be verified that the classification and regression output module introduced in this chapter, integrating CNN and multi-head attention, can improve the precision of bounding box regression. This network module employs multi-head attention. This enables the establishment of the long-distance dependence property and facilitates the regression branch to concentrate more on the boundary outlines of the target, avoiding the loss of information in the dimension compression process of the convolution operation.

Comparing 6 with 8, after using the multi-layer feature association module proposed in this chapter, the AUC has been boosted by 4.6 percentage points, and the precision has increased by 6.6 percentage points. which proves that the multi-layer feature association module can, to some extent, effectively associate and fuse feature maps at different levels, so that the middle layer feature map has richer details and semantic information.

Comparing 5 with 8, after using the multi-frame one-stage module proposed in this chapter, the AUC has improved by 12.3 percentage points, while the precision has increased by 10.2 percentage points, which verifies the effectiveness of the multi-frame single-stage module proposed in this paper, and the operation of mutual correlation computation in the process of feature extraction can enhance the target information in the feature map of search frames during the deepening process of the network module, inhibit the background area, and continuously optimize the accuracy of the subsequent network feature extraction, in addition, the multi-frame one-stage module correlates the spatiotemporal characteristics of the target motion, which can further improve the robustness of the tracking method in this chapter in the face of complex background transformations.

Table 2. Comparative results of ablation studies

Index	Multi-frame single-stage module			Multilayer feature association module	Classification regression output module	Success ratio (%)	Precision rate (%)
	Multi-frame	Modified DaViT	Single stage framework				
1	√	—	√	√	√	54.7	74.1
2	—	√	√	√	√	56.8	77.3
3	—	—	√	√	√	54.3	73.7
4	√	√	—	√	√	49.1	70.9
5	—	—	—	√	√	46.4	68.4
6	√	√	√	—	√	54.1	72.2
7	√	√	√	√	—	55.0	74.6
8	√	√	√	√	√	58.7	78.6

To further explore the extent of the contribution of different components in the multi-frame one-stage module to the AUC and accuracy, the multi-frame, the improved DaViT network and the one-stage framework are sequentially ablated and analyzed. Comparing 1, 2, and 4 with 8, respectively, the AUC and accuracy rate increased by 4.0 and 4.5 percentage points after improving the original DaViT network; the success rate and accuracy rate increased by 1.9 and 1.3 percentage points after using the multi-frame template frame image; and the AUC and accuracy rate increased by 9.6 and 7.7 percentage points after utilizing the framework of one-stage tracking. This shows that the one-stage tracking framework has a great influence on enhancing the tracking accuracy of infrared dim and small targets under complex backgrounds. In addition, in order to further analyze the gap between one-stage and two-stage infrared dim small target tracking methods, 3 and 5 are compared, and the AUC and accuracy rate are improved by 7.9 and 5.1 percentage points after using the one-stage tracking framework, which indicates that the robustness of infrared dim small target tracking can be effectively improved by utilizing the one-stage tracking framework.

4 Conclusion

The paper proposes a one-stage tracking method for infrared dim and small targets that fuses spatio-temporal information, which is utilized to track such targets under the influence of background clutter. To heighten the discriminative ability of the target feature maps, it integrates the feature extraction and target matching processes and incorporates the spatio-temporal information in the target's motion to generate feature maps that can effectively distinguish the background from the real targets. To increase the accuracy of the predicted bounding boxes of the tracking method, the paper adopts a multi-layer feature fusion module and a classification and regression output module combining CNN and multi-head attention. This improves the detail and semantic information of the feature maps and outputs more accurate classification and regression results. The combination of these three modules further improves the performance of tracking infrared dim and small targets. The test results on the test dataset demonstrate that the method in this paper has better performance.

5 Acknowledgement

This work was supported by the Natural Science Foundation of Shanxi Province of China (No. 202303021211147).

References

- [1] S. Ailneni, S.K. Kashyap, V. Naidu, A. Kumar, Angle Only Tracking for Infrared Search and Track (IRST) System Mounted on an Aircraft, in: Proc. 2021 Seventh Indian Control Conference (ICC), 2021.
<https://doi.org/10.1109/ICC54714.2021.9703161>
- [2] S. Xiao, Y. Ma, F. Fan, J. Huang, M. Wu, Tracking small targets in infrared image sequences under complex environmental conditions, *Infrared Physics & Technology* 104 (2020) 103102.
<https://doi.org/10.1016/j.infrared.2019.103102>
- [3] S.S. Rawat, S.K. Verma, Y. Kumar, Review on recent development in infrared small target detection algorithms, *Procedia Computer Science* 167(2020) 2496-2505.
<https://doi.org/10.1016/j.procs.2020.03.302>
- [4] K. Qian, H.X. Zhou, S.H. Rong, B.J. Wang, K.H. Cheng, Infrared dim-small target tracking via singular value decomposition and improved Kernelized correlation filter, *Infrared Physics & Technology* 82(2017) 18-27.
<https://doi.org/10.1016/j.infrared.2017.02.002>
- [5] S. Javed, M. Danelljan, F.S. Khan, M.H. Khan, M. Felsberg, J. Matas, Visual object tracking with discriminative filters and Siamese networks: a survey and outlook, *IEEE transactions on pattern analysis and machine intelligence* 45(5) (2023) 6552-6574.
<https://doi.org/10.1109/TPAMI.2022.3212594>
- [6] K. Qian, S.J. Zhang, H.Y. Ma, W.J. Sun, SiamIST: Infrared small target tracking based on an improved SiamRPN, *Infrared Physics & Technology* 134(2023) 104920.
<https://doi.org/10.1016/j.infrared.2023.104920>
- [7] B. Li, J.J. Yan, W. Wu, Z. Zhu, X.L. Hu, High performance visual tracking with siamese region proposal network, in: Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
<https://doi.org/10.1109/CVPR.2018.00935>
- [8] J.H. Shan, Y. Yang, H. Liu, T. Liu, Infrared Small Target Tracking Based on OTrack Model, *IEEE Access* 11(2023) 123938-123946. <https://doi.org/10.1109/ACCESS.2023.3329063>
- [9] B.T. Ye, H. Chang, B.P. Ma, S.G. Shan, X.L. Chen, Joint feature learning and relation modeling for tracking: A one-stream framework, in: Proc. 17th European Conference on Computer Vision, ECCV 2022, 2022.
https://doi.org/10.1007/978-3-031-20047-2_20
- [10] K. Han, Y.H. Wang, H.T. Chen, X.H. Chen, J.Y. Guo, Z.H. Liu, Y.H. Tang, A. Xiao, C.J. Xu, Y.X. Xu, Z.H. Yang, Y.M. Zhang, D.C. Tao, A survey on vision transformer, *IEEE transactions on pattern analysis and machine intelligence* 45(1) (2023) 87-110.
<https://doi.org/10.1109/TPAMI.2022.3152247>
- [11] Y.C. Li, S. Yang, Infrared small object tracking based on Att-Siam network, *IEEE Access* 10(2022) 133766-133777.
<https://doi.org/10.1109/ACCESS.2022.3171037>
- [12] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P.H.S. Torr, Fully-convolutional siamese networks for object tracking, in: Proc. 14th European Conference on Computer Vision, ECCV 2016, 2016.
https://doi.org/10.1007/978-3-319-48881-3_56
- [13] J. Hu, L. Shen, S. Albanie, G. Sun, E.H. Wu, Squeeze-and-excitation networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42(8)(2020) 2011-2023.
<https://doi.org/10.1109/TPAMI.2019.2913372>
- [14] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in: Proc. the 15th European Conference on Computer Vision, ECCV 2018, 2018.
https://doi.org/10.1007/978-3-030-01234-2_1
- [15] J.J. Chen, B. Huang, J.N. Li, Y. Wang, M.X. Ren, T.F. Xu, Learning spatio-temporal attention based siamese network for tracking UAVs in the wild, *Remote Sensing* 14(8)(2022) 1797.
<https://doi.org/10.3390/rs14081797>
- [16] Z.C. Lv, Y.L. Li, J.B. Yuan, J. Wang, SiamITO: A Lightweight Siamese Network for Infrared Tiny Object Tracking, in: Proc. 2022 China Automation Congress (CAC), 2022.
<https://doi.org/10.1109/CAC57257.2022.10055103>
- [17] C.H. Cui, S.Z. Lin, D.W. Li, X.F. Lu, J. Wu, Infrared dim small target tracking method based on Siamese network and Transformer, *Journal of Computer Applications* 44(2)(2024) 563-571.
<https://doi.org/10.11772/j.issn.1001-9081.2023020167>
- [18] M.Y. Ding, B. Xiao, N. Codella, P. Luo, J.D. Wang, L. Yuan, Davit: Dual attention vision transformers, in: Proc. the 17th European Conference on Computer Vision, ECCV 2022, 2022.
https://doi.org/10.1007/978-3-031-20053-3_5
- [19] X.L. Sun, L.C. Guo, W.L. Zhang, Z. Wang, Y.J. Hou, Z. Li, X.C. Teng, A dataset of semi-synthetic detection for small infrared moving targets under complex backgrounds, *China Scientific Data* 9(3)(2024) 1-17.
<https://doi.org/10.11922/csdata.2021.0015.zh>

- [20] B.W. Hui, Z.Y. Song, H.Q. Fan, P. Zhong, W.D. Hu, X.F. Zhang, J.G. Ling, H.Y. Su, W. Jin, Y.J. Zhang, Y.Q. Bai, A dataset for infrared detection and tracking of dim-small aircraft targets under ground / air background, *China Scientific Data* 5(3)(2020) 1-12.
<https://doi.org/10.11922/csdata.2019.0074.zh>
- [21] B. Yan, H.W. Peng, J.L. Fu, D. Wang, H.C. Lu, Learning spatio-temporal transformer for visual tracking, in: *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
<https://doi.org/10.1109/ICCV48922.2021.01028>
- [22] C. Mayer, M. Danelljan, G. Bhat, M. Paul, D.P. Paudel, F. Yu, L.V. Gool, Transforming model prediction for tracking, in: *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
<https://doi.org/10.1109/CVPR52688.2022.00853>
- [23] X. Chen, B. Yan, J.W. Zhu, D. Wang, X.Y. Yang, H.C. Lu, Transformer tracking, in: *Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
<https://doi.org/10.1109/CVPR46437.2021.00803>