A Lightweight Single-target Tracking Model for Underwater Sonar Scenarios

Yu Li
Merchant Marine College
Shanghai Maritime University
Shanghai, China
1079995614@qq.com

Mingzhi Chen*
School of Mechanical Engineering
University of Shanghai for Science and Technology
Shanghai, China
mingzhichen2008@163.com

Daqi Zhu
School of Mechanical Engineering
University of Shanghai for Science and Technology
Shanghai, China
zdq367@aliyun.com

2024 9th International Conference on Automation, Control and Robotics Engineering (CACRE) | 979-8-3503-5030-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CACRE62362.2024.10635057

***Abstract*—Compared with optical images, underwater sonar images suffer from blurry edges, obscure features, and noisy underwater environments, while the small computing power of AUVs further limits the application of large models in underwater environments. To improve the tracking performance of AUV. In this paper, on the basis of siamese network, we designed a lightweight single-object tracking model applied in multibeam forward-looking sonar scenarios, which incorporates the attention mechanism in the feature extraction stage to make the model focus on the target candidate region with high weight, and next the feature fusion module is used to dilute the non-target region features and highlight the features of the region with high similarity, and finally, the output is detected to obtain the target position. The experimental results show that this model can achieve 63% accuracy while keeping the 0.75M parameter count on our sonar dataset.**

Keywords—Sonar scenarios; Single-target tracking; Lightweight model; Feature fusion; Attention mechanism

I. INTRODUCTION

Target tracking is one of the important research projects in intelligent robot vision, which can be widely used in public surveillance, automatic driving, intelligent identification tracking, and so on. In underwater environments, acoustic devices are commonly used for underwater target tracking, which are categorized into: traditional acoustic sensor arrays (TASA), underwater sensor networks (UWSN), and imaging sonar [1]. Typically, AUVs use active imaging sonar to detect targets at long range. Researchers have carried out many in-depth research works on visual target tracking, which are mainly categorized into: correlation filtering algorithms [2], and deep learning-based algorithms. Correlation filtering online target tracking algorithm started in MOSSE, using Fourier transform to achieve the effect of fast computation and high frame rate, which is favored by the industry. Deep learning has received a lot of attention in the classification and recognition of targets. With the proposal of Deep Learning Tracker (DLT) algorithms, target tracking has begun the era based on deep learning [3]. As the first deep learning-based algorithm SiamFC uses a twinned convolutional network to extract features, matches the template and the features of the searched image, and thus tracks the target [4]. The method extracts target features through a pre-trained convolutional neural network, however, twinned convolutional networks focus on the localization problem and are not able to accurately adjust to changes in the shape of the target. To address this problem, the SiamRPN algorithm combines the twin network with Fast-RCNN's RPN to deal with the target tracking problem in a classification and regression manner, which accurately tracks the target location and effectively regulates the shape [5]. Subsequent version of DaSiamRPN [6] increase the model discriminative properties from data, while improving the search area to adapt long time tracking，and SiamRPN++ [7] introduce more training datasets and incorporate location-balanced ResNet to obtain better generalization ability and solve the problem that the Siam network could not be deepened. To improve target tracking accuracy, the ATOM algorithm divides the target tracking into target estimation and classification [8]. Since the proposal of Transformer, it has been widely used in natural language processing. With help of Transformer [9], TransT[10] incorporates Transformer inside the target tracking, and establishes the correlation between templates and the search region through large-scale offline training.The MixFormer [11] algorithm proposes a hybrid attention module, which integrates the two tasks of feature extraction and feature fusion, and allows the model to abandon the traditional feature extraction network, thus realizing more targeted feature extraction, and incorporates online templates to realize the matching of multiple templates with the search region, which greatly solves the problems of target deformation, target occlusion, and fast movement.

However, most of the above studies are for ground or airborne target tracking, and there are few studies on underwater target tracking. Literature [12] created a representation of region information in light of the Gaussian particle filter, which proposed the weighted integration strategy combining the area and invariant moment. Some scholars study on underwater fish tracking based on visual images [13]. This paper [14] describes the development of the tracking filter that fuses USBL and processed sonar image measurements for tracking underwater targets for the purpose of obtaining reliable tracking estimates at steady rate, even in cases when either sonar or USBL measurements are not available or are faulty. Literature [15] carried out target

tracking based on forward-looking sonar images with the application of SiamFC. Underwater target tracking based on acoustic images is more difficult than visual target tracking, there are additional difficulties in low image resolution, few target texture features, almost no color features, and many noise clutter points.

Constrained by the limited arithmetic power of the AUV like Fig. 1, a lightweight underwater sonar target tracking model is designed in this paper. Firstly, a concise backbone is designed as the feature extraction part of the twin network, and an attention mechanism is incorporated to increase the weight on the region of interest and reduce the noise effect. The feature extractor trains offline to avoid interference in online low-resolution videos. Secondly, for the template and features of the region to be searched, multi-dimensional feature interactions are performed to further disperse the noise effects, the feature map regions fused with similar targets are weighted to highlight the targets to be searched, and finally, the design of the detection head is performed. To ensure the detection speed, the features of classification and regression are shared, and then two-branch prediction and regression are used. The experimental results show that compared with the traditional twin network, this model has a better tracking effect and lighter model size.

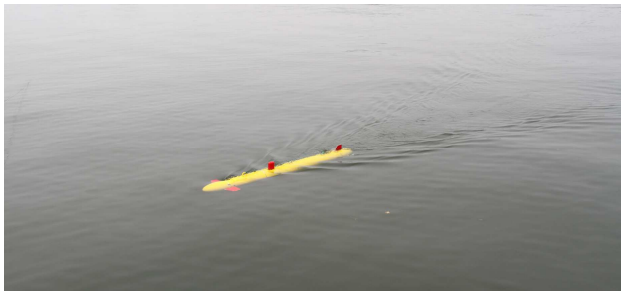

Fig.1. AUV named “Haishi Jinqiangyu”

II. DATASETS

The experimental dataset is several continuous moving videos of a moving remotely operating vehicle (ROV) captured by an AUV utilizing a multibeam sonar in a river. While the video was acquired, the AUV was kept immobile and the ROV was moved under human control, and in order to make the experimental dataset closer to the real situation, the ROV was controlled to move under the occlusion, overlap the original paths, move out of the field of view and return again, move between bubbles created by itself, and dive down to the depths and then float up again. In the meantime, the surroundings will suffer from interference from fish activity. Fig. 2 are the data samples in different cases. There are 12 video segments, totaling about 28 minutes in length, and due to the slow movement of the underwater target, sampling is taken at intervals of frames to produce the training and test sets.

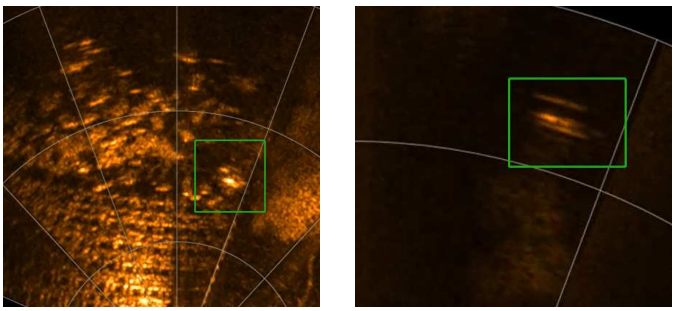

Fig.2. Obscured target and target moving to deeper water

III. ALGORITHM

Since AUVs have to ensure mobile dexterity, most of them are small in size and have limited terminal arithmetic. Therefore, the algorithms have to be as lightweight as possible while maintaining a good effect. The algorithms designed in this way are more composite to the operational scenarios of AUV.

The flowchart of the lightweight sonar target tracking algorithm is shown in Fig. 3. In order to keep the lightweight, the idea similar to SiamFC is still chosen to establish two branches to process the initial frame and the current frame respectively, and then the discriminative similarity is used to get the tracked target area.Through the previous position of the target, a proximity region can be first selected as the area to be searched. Thus, the inputs to the model are a 112*112 template and a 256*256 area to be searched. After the shared weight feature extraction backbone, this link uses the fused attention mechanism to make the model focus on the higher-weight region. The extracted templates with different specifications are fused with the search region features, and the smaller template features are used as a convolution kernel in searching the searched region features. While considering to maintain the lightweight characteristics of the model, the ordinary convolution is improved with depth separable convolution[16], which can reduce the number of parameters without significantly reducing the performance. In the specific implementation process, to avoid the influence of noise on the search background, a multilevel feature fusion operation is adopted, in which the high-level convolution is continuously carried out parallelly, and the features of the original search region are then fused. Finally, we get a feature map that generally suppresses noise and highlights the target. At last, the feature map is input to the detection head for target detection, and the maximum response position is the target's current position, thus realizing tracking.

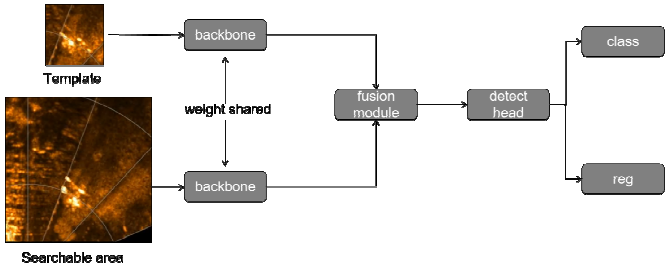

Fig.3. The overall structure of the algorithm

A. Backbone

Due to the low resolution of the sonar image and the small size of the target, the target is seen as an ill-defined bright yellow spot from a human viewpoint. Fish, obstacles, etc. in the environment are also shown as bright yellow spots. This leads to excessive noise in the image display. How to distinguish between the target bright spots and noise bright spots becomes the first problem to be solved to improve the correct rate of recognizing and tracking the target.

The first thing that comes to mind is that sonar images have a very distinct color differentiation, with the background appearing black and the range detectors generally appearing bright yellow. This limits the number of channels above the color dimension, so it naturally occurs to us that we could use a channel attention module to allow the model to learn to distinguish between background and target. It is trained to give high weight to a portion of the channels that show a strong correlation with the target color.

Once we have obtained the possible targets filtered through the channel domain, we would like to avoid jumps in tracking by further manipulation of the spatial domain to restrict the model's attention to regions where the likelihood of the target's occurrence is high, such as within a certain radius of the target. Since the ultimate goal is to restrict the weighting of the attention to a certain region, we design a Spatial channel Attention (SA) module that uses the local optimal weights instead of the all-all pixel attention weights for the region. For the selection of the optimal weights, we use a composite of maximum pooling and average pooling to finally get the attention matrix and then do the operation with the feature map.A detailed depiction of Attention module is shown in Figure. 4.

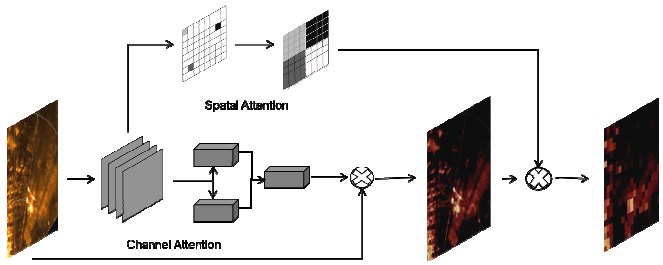

Fig.4. Attention module process

B. Fusion Module

Due to limitations in computing power and the impact of noise, the sonar target tracking model is not updated online. Since the sonar image has unclear details, directly following the template features to perform similarity matching in the search area, on the one hand, is hindered by the blurred and small details of the target, which will cause great interference from non-targets, and on the other hand, is affected by the air bubbles and obstacles during the target's moving process, all of which will make the accuracy of the tracking degrade. We refer to the idea of feature enhancement, using multi-dimensional features of the region to be searched superimposed on the template features, which has two advantages. First, in the process of feature superposition, when the template features are operated as a sliding window, the feature maps of the regions with large similarities will be more prominent, and the background and noise regions will be more evenly faded. The second is that the fused feature maps are then inspected by the detection head, which provides a secondary assurance of the similarity between the tracked target and the template.

We use the template features of the first frame as the convolution kernel and perform the convolution operation in the region to be searched. We call this step feature fusion because the results output from each step of the convolution are obtained by the template and the current region together. Considering the light weight of the model, we replace the traditional convolution with the deep separable convolution in this step as shown in Fig. 5, and finally output the fused feature map. In order to ensure that the deeper semantic information can also be fused and disclosed, we design a multi-layer fusion operation. Finally, to avoid the features of the region to be searched in the initial state being diluted in the process of continuous fusion, we perform a “concat” operation on the result at the end of each fusion, and then send it to the next layer of fusion.

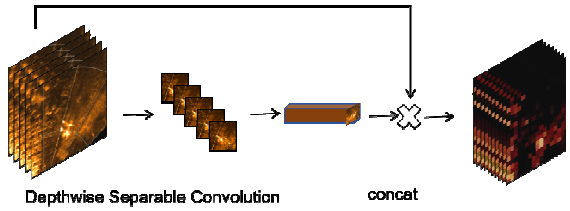

Fig.5. Fusion module

IV. Results of the Experiment

A. Experimental Environment

The algorithm is implemented using the Pytorch framework. The hardware and software environments are: Intel(R) Core (TM) i5-13400, 2.50 GHz; 16 GB of RAM; and NVIDIA GeForce GTX3060 with 12 GB of video memory. Development environment: Windows 11 64-bit operating system; CUDA11.8 and CUDNN8.7 for GPU acceleration; Pytorch deep learning framework, version 2.0.1 and torch-vision 0.15.2; Python, version 3.8; Visual Studio Code. The number of interval frames in the dataset was random in the interval [1, 3]. Some of the not-labelable and occluded images were manually eliminated after manual screening. The training batch size is set to 32, and the Adam optimizer is used with an initial learning rate of 1e-4 and weight decay of 1e-5 for a total of 300 rounds of iterations.

B. Evaluation indicators

Precision (P) indicates the number of successfully tracked frames as a percentage of the total number of frames successfully tracked. To judge whether the tracking is successful or not, the Euclidean distance from the center of the predicted frame to the center of the real frame is calculated to be less than 20, and the precision reflects the proximity between the center of the predicted frame and the center of the real frame. Since the boundary of the target is difficult to define when labeling itself, and there are errors in the boundary of the IoU, the accuracy in terms of distance is chosen as a fist indicator. And Expect Average Overlap rate

(EAO) is a more comprehensive single-target tracking performance evaluation index by combining tracker average overlap and robustness, and the larger the EAO value, the better the tracking performance.

C. Results

A comprehensive comparison of the effectiveness of current target tracking algorithms with smaller model sizes on homegrown datasets is presented below Table I:

TABLE I. ALGORITHM PERFORMANCE COMPARISON

Algorithm	Evaluation indicators		
	P	***EAO***	***Size(M)***
SiamFC	0.574	0.396	-
SiamFC++	0.530	0.414	13.9
SiamRPN	0.577	0.433	5.24
SiamRPN++	**0.673**	**0.495**	11.2
DaSiam	0.512	0.428	19.6
LightTrack	0.520	0.470	**1.97**
Ours	**0.631**	**0.452**	**0.75**

With the above experimental results, we can see that in terms of target tracking in sonar images, the accuracy of our model can reach 63.1%, which is only 4.2% less than that of SiamRPN++, which is better than other twin neural network series models.The EAO of this model is 4.3% less compared to the optimal one, but the number of parameters of the model is significantly reduced, by about 93% compared to the accuracy optimal model. Compared to LightTrack [17], which has the lowest number of parameters, the accuracy is improved by about 11% and the amount of parameters is reduced by about 61%， only 0.75 million parameters.The lightness and validity of this model is demonstrated.

In order to further validate the effect of the individual modules on the model, ablation experiments were conducted on the sonar dataset with SiamFC as the benchmark. The other group uses our original benchmark. The results of the comparison are presented in Table II.

TABLE II. ABLATION EXPERIMENT RESULTTS

Algorithm	***P***
SiamFC	0.574
SiamFC+Attention	0.582
SiamFC+Attention+Fusion	0.604
Our benchmark	0.363
Benchmark+Attention	0.519
Benchmark+Attention+Fusion	**0.631**

The experimental results show that in the SiamFC-based experiments, the addition of the Attention module and the Fusion module can further increase the model's tracking effectiveness in the sonar environment, with accuracy improvements of 0.8% and 2.2%, respectively. In the second set of experiments, due to our benchmark structure is too simple to recognize the target with various noise interference, the accuracy is only 36.3%, but after the feature processing of Attention module, it can separate the background from the high weight region to reduce the interference, and the accuracy is increased by 15.6%. the Fusion module can overcome the shallow feature hierarchy brought by the separation convolution problem, resulting in a further 11.2% increase in accuracy.

As shown in Fig. 6 and 7, the target generates a lot of bubbles during the motion process, which produces a certain amount of occlusion on the subsequent motion path, resulting in feature blurring. At this time, SiamFC and LightTrack incorrectly recognize the noise as the target due to the lack of deep feature extraction means. Our algorithm with the addition of attention will filter out the highlighted region as the region of interest and reduce the black region weight. In the subsequent feature fusion stage, weighting the template features with the region of interest, the non-target will be diluted and blurred out due to different contours and features, while the real target will be continuously deepened with features. Finally, the feature map that highlights the target region and dilutes the rest of the noise is obtained and finally recognized accurately. In Fig. 8, although the algorithms all accurately tracked the target, our model incorporates an attention region and integration of the target region, so that the recognition frame is more accurately targeted.

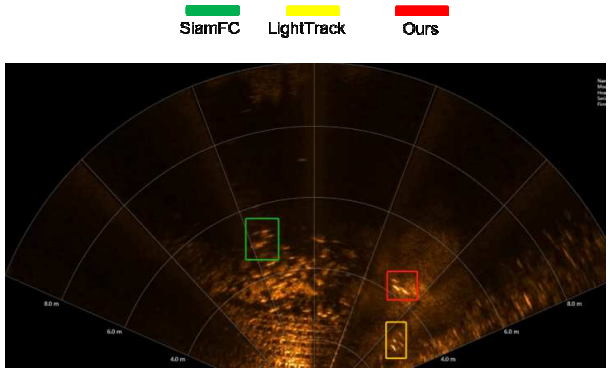

Fig.6. Bubble occlusion scene 1

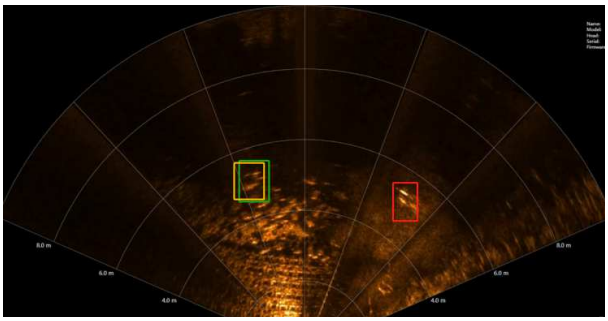

Fig.7. Bubble occlusion scene 2

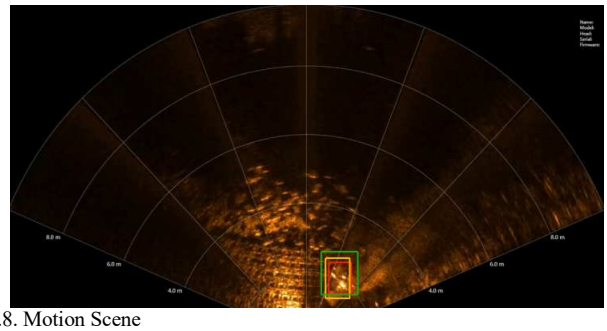

Fig.8. Motion Scene

V. CONCLUSION

In this work, we review the existing algorithms for target tracking and propose a lightweight tracking algorithm that can be applied to end devices for target tracking for applications in underwater sonar scenarios. The algorithm is based on twin neural networks, and in the feature extraction stage, for the characteristics of single color composition of sonar images and highlighting of object regions, a combined attention module is designed to increase the weights of regions of interest. Then for the characteristics of sonar imaging with low pixels, unclear edges, and much noise influence of template matching, the feature fusion module is designed as a means of composite template and searching region features, which can effectively lighten the noise influence and highlight the region with high similarity to the target. Finally, the resulting feature map is detected to get the current target position information. The experimental results show that the tracking of sonar images can reach 63% accuracy. In the future, we will further discuss the model improvement method that can update the template features online and take the template position into account, so as to realize more accurate single-target tracking of sonar.

REFERENCES

[1] S. Hare, A. Saffari, P. H. S. Torr, “Struck: Structured output tracking with kernels,” 2011 International Conference on Computer Vision(ICCV), 2011: 263-270.

[2] B. Liu, X. Tang, R. Tharmarasa, T. Kirubarajan, et al., “Underwater Target Tracking in Uncertain Multipath Ocean Environments,” IEEE Transactions on Aerospace and Electronic Systems, 2020, 56(6): 4899-4915.

[3] N. Wang, D. Y. Yeung, “Learning a deep compact image representation for visual tracking,” Advances in neural information processing systems, 2013, 20(3): 1326-1350.

[4] J. F. Henriques, R. Caseiro, P. Martins, J. Batista, “High-Speed Tracking with Kernelized Correlation Filters,” IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(3): 583-596.

[5] B. Li, W. Wu, Z. Zhu, J. Yan, X. Hu, “High Performance Visual Tracking With Siamese Region Proposal Network,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 8971-8980.

[6] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, W. Hu, “Distractor-aware Siamese Networks for Visual Object Tracking,” The European Conference on Computer Vision (ECCV), 2018: 101-117.

[7] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, J. Yan, “SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 4277-4286.

[8] M. Danelljan, G. Bhat, F. S. Khan and M. Felsberg, “ATOM: Accurate Tracking by Overlap Maximization,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 4655-4664.

[9] S. Gao, C. Zhou, J. Zhang, “Generalized Relation Modeling for Transformer Trackin,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 18686-18695.

[10] Chen X, Yan B, Zhu J, et al., “Transformer tracking,” IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2021: 8126- 8135.

[11] Cui Y, Jiang C, Wang L, et al., “MixFormer: End-to- end tracking with iterative mixed attention,” IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2022:13608-13618.

[12] Zhang Td, Wan L, Zeng Wj. et al. “Object detection and tracking method of AUV based on acoustic vision,” China Ocean Eng 26, 623–636 (2012)

[13] X. Li, Z. Wei, L. Huang, J. Nie, W. Zhang, et al., “Real-time underwater fish tracking based on adaptive multi-appearance model,” 2018 25th IEEE international conference on image processing (ICIP), 2018: 2710-2714.

[14] Filip Mandić, Ivor Rendulić, Nikola Mišković, Đula Nađ, "Underwater Object Tracking Using Sonar and USBL Measurements", Journal of Sensors, vol. 2016, Article ID 8070286, 10 pages, 2016.

[15] X. Ye, Y. Sun, C. Li, “FCN and Siamese Network for Small Target Tracking in Forward-looking Sonar Images,” OCEANS, 2018: 1-6.

[16] Howard A G, Zhu M, Chen B, et al., “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” arXiv preprint arXiv:1704.04861, 2017.

[17] Yan B, Peng H, Wu K, et al., “Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search,” IEEE Conference on Computer Vision and Pattern Recognition. 2021: 15180-15189.