Multi-Target Instance Segmentation and Tracking using Yolov8 and BoT-SORT for Video SAR

Shangqu Yan
College of Electronic Science and Technology
National University of Defense Technology
Changsha, China
shangqu_yan@163.com

Yaowen Fu*
College of Electronic Science and Technology
National University of Defense Technology
Changsha, China
fuyaowen@sina.com

Wenpeng Zhang
College of Electronic Science and Technology
National University of Defense Technology
Changsha, China
zhangwenpeng@hotmail.com

Wei Yang
College of Electronic Science and Technology
National University of Defense Technology
Changsha, China
yw850716@sina.com

Ruofeng Yu
College of Electronic Science and Technology
National University of Defense Technology
Changsha, China
yuruofeng0819@sina.com

Fatong Zhang
College of Electronic Science and Technology
National University of Defense Technology
Changsha, China
zft38768@163.com

2023 5th International Conference on Electronic Engineering and Informatics (EEI) | 979-8-3503-2707-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/EEI59236.2023.10212903

***Abstract*—Moving target detection and tracking is a crucial application field of video SAR. It can offer significant support for military intelligence reconnaissance and civil monitoring. Compared with the traditional moving target detection and tracking methods, the deep learning-based moving target detection and tracking methods can more fully extract the moving targets' shadow features in video SAR frame images to improve detection and tracking accuracy. This paper proposes a multi-target instance segmentation and tracking framework based on the YOLOv8 model and the BoT-SORT algorithm for video SAR. Firstly, the detection effects of YOLOv8-seg models with different sizes are analyzed, and the most suitable detection model for video SAR moving targets is selected. Secondly, the BoT-SORT algorithm is used to track the targets detected by the detector to complete the whole detection and tracking process. Based on the experimental results, it has been determined that the proposed framework is highly effective in detection and tracking accuracy. Additionally, it has the capability to achieve real-time tracking.**

Keywords—Video SAR, moving targets detection and tracking, YOLOv8, BoT-SORT

I. INTRODUCTION

Synthetic aperture radar (SAR) is an active radar imaging system [1], it breaks through the resolution limit of traditional radar antenna aperture and obtains high resolution in azimuth direction and range direction by synthetic aperture and pulse compression, respectively, which greatly improves the radar imaging ability of the target area [2]. As an important research direction in SAR, moving target detection and tracking can realize the tracking of moving targets based on highly accurate detection, which can improve the reconnaissance ability of military targets [3].

In 2003, Sandia National Laboratories (SNL) proposed the concept of video SAR [4] and achieved high frame rate imaging on an airborne platform. The advent of video SAR has provided a key technology for high-resolution video surveillance of a target area. Video SAR not only has the characteristics of traditional SAR but also has the advantage of continuously observing the target area. Meanwhile, in video SAR, moving target's shadow is an effective feature for detection and tracking, which can effectively improve detection and tracking accuracy [5]. With the rapid development of deep learning, deep learning-based feature extraction networks can adaptively extract the rich abstract features of the target, and achieve higher accuracy than traditional algorithms in the field of target detection and tracking. Therefore, numerous researchers have also applied deep learning to the video SAR moving targets detection task and achieved many research results. Target detection and tracking algorithms based on deep convolutional neural networks (DCNN) have been successfully applied to radar [6]-[9], and all have achieved good performance. DCNN can automatically extract the features of targets from data for detection, recognition, and tracking, and remove some limitations of traditional methods. However, DCNN is not an end-to-end detection network, which has the problem of slow detection speed. Therefore, Ding et al. [10] applied a faster region-based convolutional neural network (Faster R-CNN) [11] to detect the moving targets' shadows, and adopted an improved density-based clustering algorithm to suppress false alarms in the initial detection. However, the classical Faster R-CNN has poor generalization ability, and its detection performance will severely degrade for unknown scenes that are not used in the training dataset, resulting in many false alarms.

To solve the above problems, we proposed a multi-target instance segmentation and tracking framework. In the detector part of the framework, the YOLOv8 model [12] with strong generalization ability is used to segment the ground moving targets' shadows in video SAR frame images to realize the detection of moving targets. Then, the BoT-SORT algorithm [13] is used to track the detection results to achieve the tracking of moving targets. This new framework has been verified by the video SAR data released by the SNL.

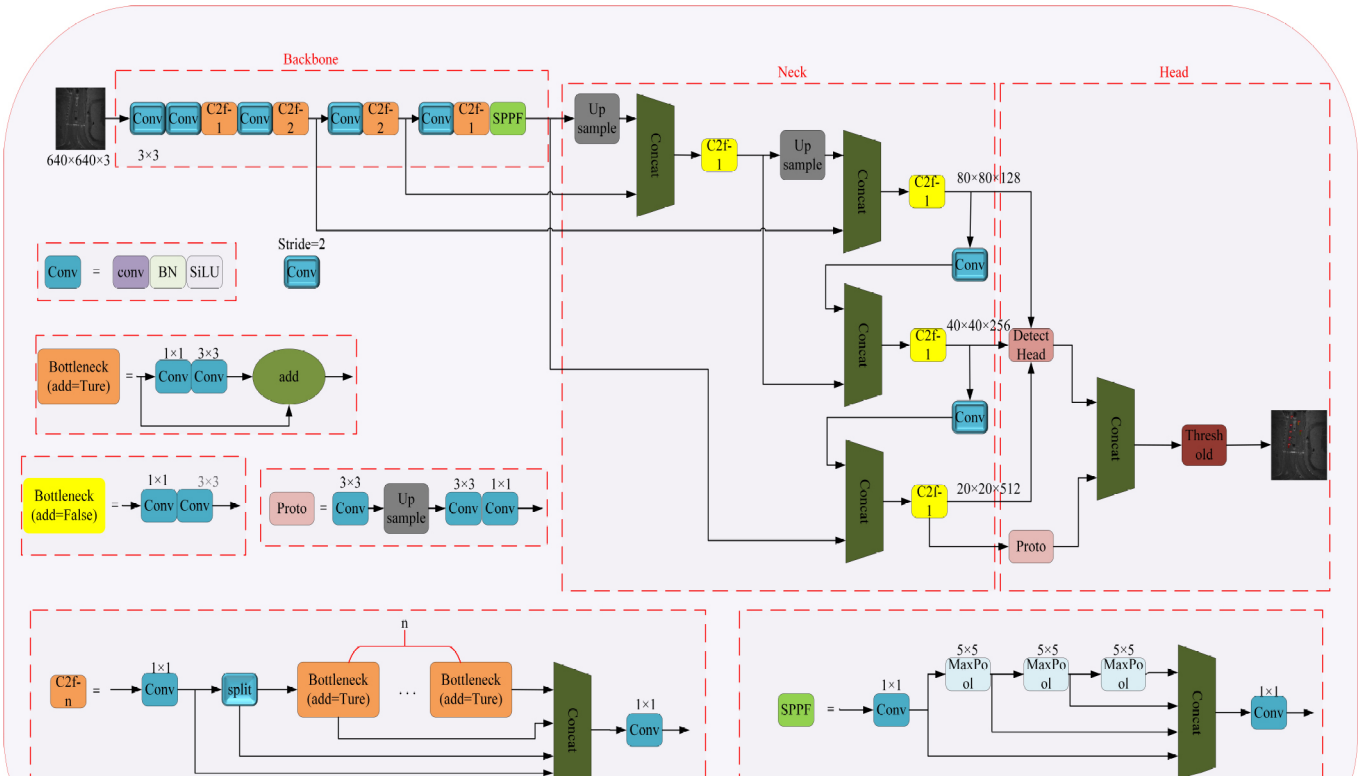

TABLE I. SCALING PARAMETERS FOR THE YOLOV8-SEG MODEL

	YOLOv8n-seg	*YOLOv8s-seg*	*YOLOv8m-seg*	*YOLOv8l-seg*	*YOLOv8x-seg*
Depth_multiple	0.33	0.33	0.67	1.0	1.0
Width_multiple	0.25	0.50	0.75	1.0	1.25
C2f-n (Backbone)	1, 2, 2, 1	1, 2, 2, 1	2, 4, 4, 2	3, 6, 6, 3	3, 6, 6, 3
C2f-n (Neck)	1, 1, 1, 1	1, 1, 1, 1	2, 2, 2, 2	3, 3, 3, 3	3, 3, 3, 3
Maximum number of channels	1024	1024	768	512	512

II. A FRAMEWORK FOR MULTI-TARGET INSTANCE SEGMENTATION AND TRACKING

A. Structure and Details of the YOLOv8-seg Model

The YOLOv8s-seg model's structure is depicted in Fig. 1. It comprises three modules: backbone, neck, and head. Compared with the YOLOv5 model [14], in the backbone of the YOLOv8 model, the convolution kernel in the first convolutional module is changed from 6×6 to 3×3, and the C3 module is replaced by the C2f module, where the C2f module uses three convolution modules and n bottlenecks. This enables the model to be further lightweight and obtain richer gradient flow information. At the same time, the idea of dual-stream feature pyramid networks (FPN) is used in the neck module, and the 1×1 convolution module before up-sampling is removed, which can better perform feature fusion compared with the YOLOv5 model. The head module of the YOLOv8-seg model adopts the idea of YOLACT [15], which is divided into two branches: detect head and proto. The detect head is a decoupled head and anchor-free structure, which is used to realize target detection and classification. The loss functions used in the detection function of the detect head are complete intersection over union (CIoU) and distribution focal loss (DFL), and the loss function used in the classification function is the binary cross-entropy (BCE) loss function. Their expressions are:

$$CIoU = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{1}$$

Where b and b^{gt} represent the center point of the predicted and true bounding boxes, ρ represents the Euclidean distance of the two center points, and c represents the diagonal distance of the minimum closure region that can contain both the two boxes. $\alpha = v/(1-Iou)+v$ is the weight function, and $v = \frac{4}{\pi^2}(\arctan {}^{w^{gt}}\!/_{h^{gt}} - \arctan {}^{w}\!/_{h})^2$ is used to calculate the similarity of the aspect ratio of two boxes.

$$DFL = -((y_{i+1} - y)\bullet\log(s_i) + (y - y_i)\log(s_{i+1}) \tag{2}$$

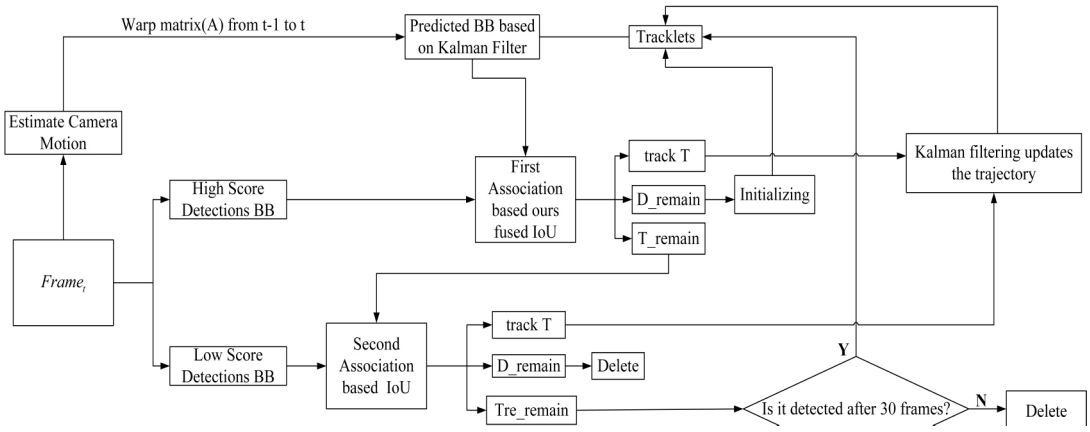

Fig. 2. Overall process of the BoT-SORT algorithm

where s is the sigmoid output of the network, y is the coordinates of the bounding box, and i represents the variation range of the box.

$$BCE = -\frac{1}{N}\sum_{i}^{N}[\text{target}[i]\bullet\log(\psi[i]) + (1-\text{target}[i])\bullet\log(1-\psi[i])] \tag{3}$$

where target$[i]$ is the label of the target, $\psi[i]$ is the label predicted by the network, and N is the total number of targets.

Another branch, proto, is used to obtain the prototype masks with the loss function BCEWithLogitsLoss, which is essentially the BCE loss plus a sigmoid function. The final detection results are obtained by weighting and thresholding the results of the two branches.

It should be noted that the YOLOv8-seg has five scaling models, as shown in Table 1. Therefore, the YOLOv8-seg model can adjust the depth and width of the network by two parameters depth_multiple and width_multiple to achieve the state-of-the-art (SOTA) effect. Where depth_multiple controls the size of n in the C2f-n module, and width_multiple controls the size of the number of channels in the convolutional module. The effect of the YOLOv8-seg models with different sizes on the detection results will also be explored in the following experiments.

B. The BoT-SORT Tracking Algorithm

At present, most multi-target tracking algorithms adopt the strategy of matching high-score bounding boxes to identify targets, but the low-score bounding boxes are directly discarded, which results in a large number of missing alarms. In the BoT-SORT tracking algorithm, the authors track multiple targets by associating each detection bounding box, including the low-score detection bounding boxes, which can greatly improve the overall performance in the tracking process. The process of this tracking algorithm is shown in Fig. 2.

The overall process of the BoT-SORT algorithm is similar to ByteTrack [16] but with two improvements:

- The state vector of the Kalman filter (KF) is improved.
- The camera motion compensation method is used to improve the prediction of bounding boxes.

Initially, the state vector of the KF in the SORT algorithm consists of seven elements: $x=[x,y,s,a,\hat{x},\hat{y},\hat{s}]^T$, where x and y are the center coordinates of the bounding box, s is the area of the bounding box, a is the aspect ratio of the bounding box, and $\hat{\bullet}$ represents the predicted value of the state. The state vector of the improved KF in the BoT-SORT algorithm is: $x=[x,y,w,h,\hat{x},\hat{y},\hat{w},\hat{h}]^T$. The new state vector changes the aspect ratio a to width w and height h. The predicted width and high can better match the target's bounding box and greatly improve the IoU in tracking matching.

In addition, the KF is unsuitable for tracking nonlinear moving targets because it is a uniform linear motion model. Therefore, the BoT-SORT algorithm uses OpenCV's global motion estimation (GME) technology to improve this shortcoming. Firstly, the key points of the image are extracted, and then sparse optical flow is used for feature tracking based on local outlier suppression. Random sample consensus is then used to compute the affine transformation matrix, which transforms the predicted bounding box from the k-1 frame coordinates to its next k frame coordinates. This improvement compensates for the problem of blurred bounding boxes caused by the irregular motion of the targets.

III. Experiments Result

A. Datasets, Metrics, and Implementation details

The video SAR dataset used in this paper is a 30-second video of Kirtland Air Force Base released by SNL. The FPS of the video is 30, and the resolution of each image frame is 720×660. After converting, it is possible to obtain 900 continuous SAR images. The "labelme" software is used to mark the moving targets' shadows for 900 continuous SAR images. For irregular shapes marked by "labelme", when the 'json' format is converted to 'yolo' format, the corresponding bounding boxes are generated according to the marked shapes. According to guidelines found in the literature [13], we split the dataset into three parts in order: 70% for training, 10% for validation, and 20% for testing in our experimental study.

We evaluated the effect of detection and tracking in terms of well-accepted explicit metrics, such as precision, recall, mean average precision at IoU = 0.5 (mAP50) and mAP50-95 in detection task [17], and multiple object tracking accuracy (MOTA) and identification switch (IDSw) [18] in the tracking task.

All the experiments were implemented using PyTorch and ran on a desktop with Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz and two NVIDIA GeForce RTX 3090 GPUs. In the training part of the detector, we trained 100 epochs with batch size 8, the detector model is trained with the optimizer SGD with the momentum of 0.937, the weight decay of 5e-4, and the initial learning rate of 0.01. We adopted the YOLOv8-seg model pre-trained on the coco128 dataset to make the detector fit fast. In addition, the gain of the bounding box loss function is set to 7.5, the gain of the category loss function is set to 0.5, and the gain of the DFL is set to 1.5. In the tracking process, we set the threshold of the high-score bounding box in the first association as 0.5, and in the second association, if the IoU is less than 0.1, we reject the match.

B. Detection Result

Table 2 shows the comparison of detection results under different size models. The metrics of bounding boxes and masks are calculated separately, and the best results under each metric are marked in bold.

From the comparison of bounding boxes' metrics in this table, it can be seen that the detection performance of the YOLOV8m-seg model is the best value in recall rate, mAP50, and mAP50-90, and the recall rate is 5% higher than that of the YOLOv8x-seg model. In the metrics comparison of masks, the YOLOv8x-seg model has better results on mAP50 and mAP50-90. However, the recall rate is still not as high as the YOLOv8m-seg model, and the mAP50 is only 2.2% higher than the YOLOv8m-seg model. This indicates that the YOLOv8m-seg model can detect the real targets well. In addition, since the YOLOv8m-seg model has a smaller depth and width, it will be faster to train. Therefore, considering all factors, the YOLOv8m-seg model is selected as the detector in the subsequent tracking experiment.

TABLE II. COMPARISON OF DETECTION PERFORMANCE UNDER DIFFERENT SIZE MODELS

Models	*Box (Precision*	*Recall*	*mAP50*	*mAP50-95)*	*Mask (Precision*	*Recall*	*mAP50*	*mAP50-95)*
YOLOv8n-seg	**0.979**	0.488	0.735	0.44	**0.918**	0.457	0.692	0.32
YOLOv8s-seg	0.969	0.569	0.773	0.494	0.911	0.535	0.722	0.334
YOLOv8m-seg	0.956	**0.763**	**0.866**	**0.563**	0.874	**0.697**	0.777	0.346
YOLOv8l-seg	0.943	0.709	0.834	0.539	0.887	0.667	0.770	0.349
YOLOv8x-seg	0.952	0.713	0.839	0.535	0.914	0.684	**0.797**	**0.418**

C. Tracking Result

According to the above, after the video sequence used for testing is input into the trained detector, the detection bounding boxes and masks of the ground targets' shadows can be obtained. Then the results are input into the BoT-SORT tracker, and the Kalman filter and Hungarian algorithm are used to complete the association and matching.

Fig. 3 shows the results of the BoT-SORT tracking algorithm. The test video has 180 consecutive frame images for detection and tracking. In this result, to better show the effect of targets movement and tracking, the tracking results of every five frames starting from frame 868 are placed in Fig.3. It is evident from Fig.3 that after the target with id 55 passes through the occlusion of the gate, it can still be detected in frame 888, and the corresponding id is successfully locked for tracking. This can reflect that the BoT-SORT algorithm can track the target well in the presence of short-term occlusion. However, the target with id 61 was not successfully tracked in frame 868, which is determined by the detector performance. The YOLOv8 model has too large down-sampling rate and receptive field, and lacks a detection head for tiny targets. Therefore, when it is applied to the moving targets detection task of video SAR, there will be more missing alarms, which is also the focus of our improvement in the future.

The metrics results of the tracking test based on the YOLOv8m-seg model and the BoT-SORT algorithm are shown in Table 3. Among the metrics, a larger MOTA value and a smaller IDSw value indicate a better tracking effect, and FPS represents how many frames per second the framework can reason. It can be seen from Table 3 that the MOTA of the proposed framework is 49.7%, which shows that the framework has good tracking precision in the video SAR moving targets tracking task. The IDSw is 33, indicating that the IDs of targets have changed 33 times during the tracking process. Since the video SAR frame images contain more speckle noise, there is a conversion of the targets' ID of adjacent frames in the matching process. How to reduce IDSw and improve MOTA is the focus of our subsequent study. The FPS of the whole framework is 24.8, which has the ability for real-time detection and tracking.

TABLE III. MULTI-TARGET TRACKING RESULTS

Method	*MOTA*	*IDSw*	*FPS*
YOLOv8m-seg + BoT-SORT	49.7%	33	24.8

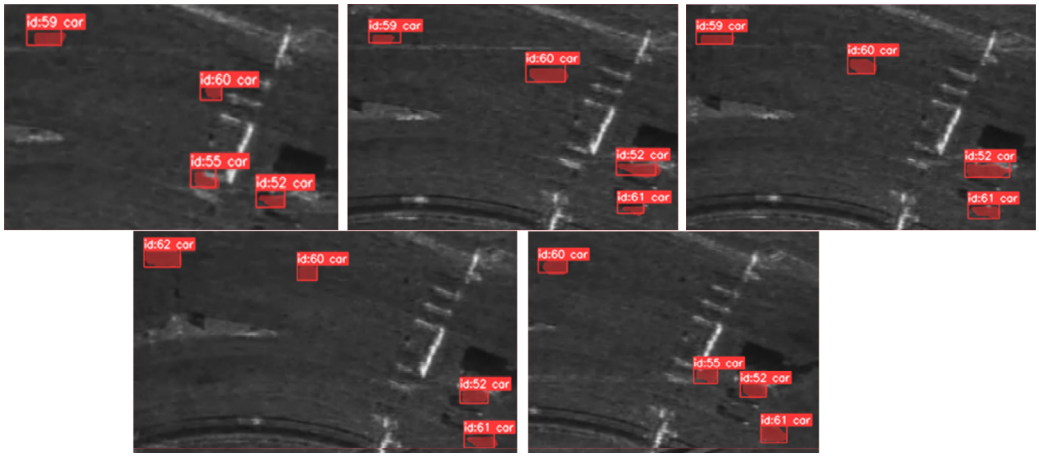

Fig. 3. Examples of multi-target tracking results based on the BoT-SORT Algorithm

IV. Conclusion

In this paper, we used the YOLOv8 model proposed in early 2023 and the BoT-SORT algorithm proposed in late 2022 for real-time video SAR moving targets detection and tracking. At the same time, the structure of the YOLOv8 model is analyzed in detail, the detection ability of YOLOv8-seg models with different sizes in video SAR moving targets detection task is discussed, and the model suitable for video SAR moving targets detection is determined. Finally, the detector combined with the BoT-SORT algorithm was used for video SAR moving targets tracking experiment. The experimental results show that our proposed framework has good tracking precision and real-time performance in video SAR moving targets detection and tracking task. However, since the detector and tracking algorithm are not perfect for tiny targets, it will cause a certain phenomenon of missing alarm and ID switching. In future work, we will combine the characteristics of moving targets' shadows in video SAR to improve the detection accuracy and tracking precision for tiny targets.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61871384 and Grant 61401486, and the Science and Technology Innovation Program of Hunan Province under Grant 2022RC1092.

References

[1] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek and K. P. Papathanassiou, "A tutorial on synthetic aperture radar," IEEE Geoscience and Remote Sensing Magazine, vol. 1, no. 1, pp. 6-43, March 2013.

[2] T. Tanaka, I. N. S. Parwata and P. E. Yastika, "Synthetic Aperture Radar Interferometry: Utilizing Radar Principles," IEEE Geoscience and Remote Sensing Magazine, vol. 8, no. 4, pp. 111-116, Dec. 2020.

[3] K. Suwa, K. Yamamoto, M. Tsuchida, S. Nakamura, T. Wakayama and T. Hara, "Image-Based Target Detection and Radial Velocity Estimation Methods for Multichannel SAR-GMTI," IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 3, pp. 1325-1338, March 2017.

[4] L. Wells, K. Sorensen, A. Doerry and B. Remund, "Developments in sar and ifsar systems and technologies at sandia national laboratories," 2003 IEEE Aerospace Conference Proceedings (Cat. No.03TH8652), Big Sky, MT, USA, 2003, pp. 2_1085-2_1095.

[5] H. Yan, X. Mao, J. Zhang and D. Zhu, "Frame rate analysis of video synthetic aperture radar (ViSAR)," 2016 International Symposium on Antennas and Propagation (ISAP), Okinawa, Japan, 2016, pp. 446-447.

[6] Y. Zhang, S. Yang, H. Li and Z. Xu, "Shadow Tracking of Moving Target Based on CNN for Video SAR System," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 2018, pp. 4399-4402.

[7] H. Fang, Y. Liu, G. Liao, H. Li, C. Chen and X. Liu, "Moving Target Tracking Based on Shadow with Unsupervised Deep Track in Video-SAR," 2021 CIE International Conference on Radar (Radar), Haikou, Hainan, China, 2021, pp. 1192-1195.

[8] Z. Liu, D. K. C. Ho, X. Xu and J. Yang, "Moving Target Indication Using Deep Convolutional Neural Network," IEEE Access, vol. 6, pp. 65651-65660, 2018.

[9] Y. Chung, P. Chou, M. Yang and H. Chen, "Multiple-Target Tracking with Competitive Hopfield Neural Network Based Data Association," IEEE Transactions on Aerospace and Electronic Systems, vol. 43, no. 3, pp. 1180-1188, July 2007.

[10] J. Ding, L. Wen, C. Zhong and O. Loffeld, "Video SAR Moving Target Indication Using Deep Neural Network," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 10, pp. 7194-7204, Oct. 2020.

[11] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[12] R. Hamadi, H. Ghazzai and Y. Massoud, "Image-based Automated Framework for Detecting and Classifying Unmanned Aerial Vehicles," 2023 IEEE International Conference on Smart Mobility (SM), Thuwal, Saudi Arabia, 2023, pp. 149-153.

[13] N. Aharon, R. Orfaig, B. Bobrovsky, BoT-SORT: Robust Associations Multi-Pedestrian Tracking. arXiv 2022, arXiv:2206.14651

[14] X. Xiang, Z. Wang and Y. Qiao, "An Improved YOLOv5 Crack Detection Method Combined With Transformer," IEEE Sensors Journal, vol. 22, no. 14, pp. 14328-14335, 15 July15, 2022.

[15] T. Meng and W. Zhang, "Fast Video Object Segmentation via Dynamic YOLACT," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 2400-2404.

[16] L. Shen, M. Liu, C. Weng, J. Zhang, F. Dong and F. Zheng, "ColorByte: A real time MOT method using fast appearance feature based on ByteTrack," 2022 Tenth International Conference on Advanced Cloud and Big Data (CBD), Guilin, China, 2022, pp. 1-6.

[17] X. Sun, Y. Fu, W. Zhang, W. Yang, R. Yu and F. Zhang, "Multi-Channel SAR Moving Target Detection by Integrating STAP and Faster R-CNN," 2022 7th International Conference on Signal and Image Processing (ICSIP), Suzhou, China, 2022, pp. 327-332.

[18] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. LealTaix´e, and B. Leibe, "Hota: A higher order metric for evaluating multi-object tracking," in International journal of computer vision, vol. 129, no. 2, pp. 548–578, Oct. 2021.