STN-Track: Multiobject Tracking of Unmanned Aerial Vehicles by Swin Transformer Neck and New Data Association Method

Xiangkai Xu, Zhejun Feng, Changqing Cao, Chaoran Yu, Mengyuan Li, Zengyan Wu, Shubing Ye, and Yajie Shang

***Abstract*—The gradual development of remote sensing object tracking technology based on unmanned aerial vehicles (UAV) videos has become one of the main research directions in the field of visual tracking. However, due to characteristics of the UAV platform, typical visual tracking algorithms currently applied to natural scenes cannot be used directly. Small-scale objects in UAV remote sensing videos are difficult to detect and have the problem of tracking identity switching. In order to solve these problems, we designed the Swin transformer neck-YOLOX (STN-YOLOX) object detection algorithm as the detection module, and the G-Byte data association method as the tracking module. We then combined the two into a new multiobject tracking algorithm named STN-Track. We used STN-Track to conduct experiments on the UAVDT and VisDrone MOT datasets. The experimental results show that compared with the current state-of-the-art (SOTA) methods, our STN-Track has improved detection and tracking accuracy of small-scale objects and greatly improved identification capabilities for object tracking. Compared with the SOTA ByteTrack algorithm, MOTA of STN-Track can be improved by up to 3.2%, AP_S can be improved by up to 4.4%, MT can be improved by up to 6.8%, and IDSW can be reduced by up to 28.0%.**

***Index Terms*—Data association, multiobject tracking (MOT), object detection, Swin transformer, unmanned aerial vehicles (UAV).**

I. Introduction

VIDEO object tracking, one of the basic tasks of computer vision, is a process of predicting changes of objects of interest in videos based on scene understanding and video analysis. In recent years, unmanned aerial vehicles (UAV) technology has been widely used in military and civilian fields, such as road monitoring and target search in harsh environments [1]. The gradual development of remote sensing object tracking technology based on UAV remote sensing videos has become one of the main research directions in the field of visual tracking [2]. However, due to the characteristics of the UAV platform, UAV videos have difficult problems such as small-scale objects, lack of texture, low resolution, and complex backgrounds. With these problems, typical visual tracking algorithms currently applied to natural scenes cannot be used directly, and it is difficult to guarantee robustness and adaptability. At the same time, due to the limited computing resources of the UAV itself, it cannot withstand operations with too high complexity. It is also extremely challenging to develop a tracking algorithm with low complexity while ensuring accuracy.

According to different tracking tasks and application scenarios, object tracking technology can be divided into single-object tracking (SOT) and multiobject tracking (MOT) [3]. MOT can track multiple objects, or track the identity (ID) of different individuals of the same type of objects. At the application level, MOT is more in line with the needs of UAV remote sensing object tracking. The realization of an accurate and robust UAV remote sensing MOT algorithm has important research significance for further development and application of the object tracking field. MOT is used to track multiple objects simultaneously in a video or image sequence, while keeping the ID of each tracked object unchanged. According to the object initialization method, the current MOT algorithms can be divided into two categories: detection-based tracking (DBT) and detection-free tracking (DFT). Because the DFT algorithm relies too much on manual annotation and the process is cumbersome and complicated, it has gradually been replaced by the DBT algorithm. The current mainstream MOT algorithms are mainly online tracking algorithms based on object detection, which have a wide range of application scenarios. DBT mainly includes four parts: detection model, feature model, similarity metrics, and data association.

Object detection models are a key part of MOT, and the results determine the performance to a certain extent. At present, object detection based on deep learning has become the mainstream detection model of MOT, which promotes the rapid development of MOT technology. Therefore, an accurate and efficient detector is very important for MOT algorithms. In recent years, a lot of excellent object detection algorithms have been developed based on convolutional neural network (CNN), such as the path aggregation network (PANet) [4] and cascade regional CNN (R-CNN) [5]. However, they still suffer from typical problems such as inaccurate segmentation edges and a weak ability to

Manuscript received 17 June 2022; revised 18 August 2022 and 21 September 2022; accepted 6 October 2022. Date of publication 10 October 2022; date of current version 19 October 2022. This work was supported in part by the National Natural Science Foundation of Shaanxi Province under Grant 2020 JM-206, in part by the State Key Laboratory of Laser Interaction with Matter under Grant SKLLIM2103, and in part by the 111 project under Grant B17035. *(Corresponding author: Changqing Cao.)*

The authors are with the School of Optoelectronic Engineering, Xidian University, Xi'an 710071, China (e-mail: xkxu@stu.xidian.edu.cn; zhjfeng@mail.xidian.edu.cn; chqcao@mail.xidian.edu.cn; 21051212260@stu.xidian.edu.cn; myli151024@stu.xidian.edu.cn; zywu_21@stu.xidian.edu.cn; sbye@stu.xidian.edu.cn; 20051212174@stu.xidian.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3213438

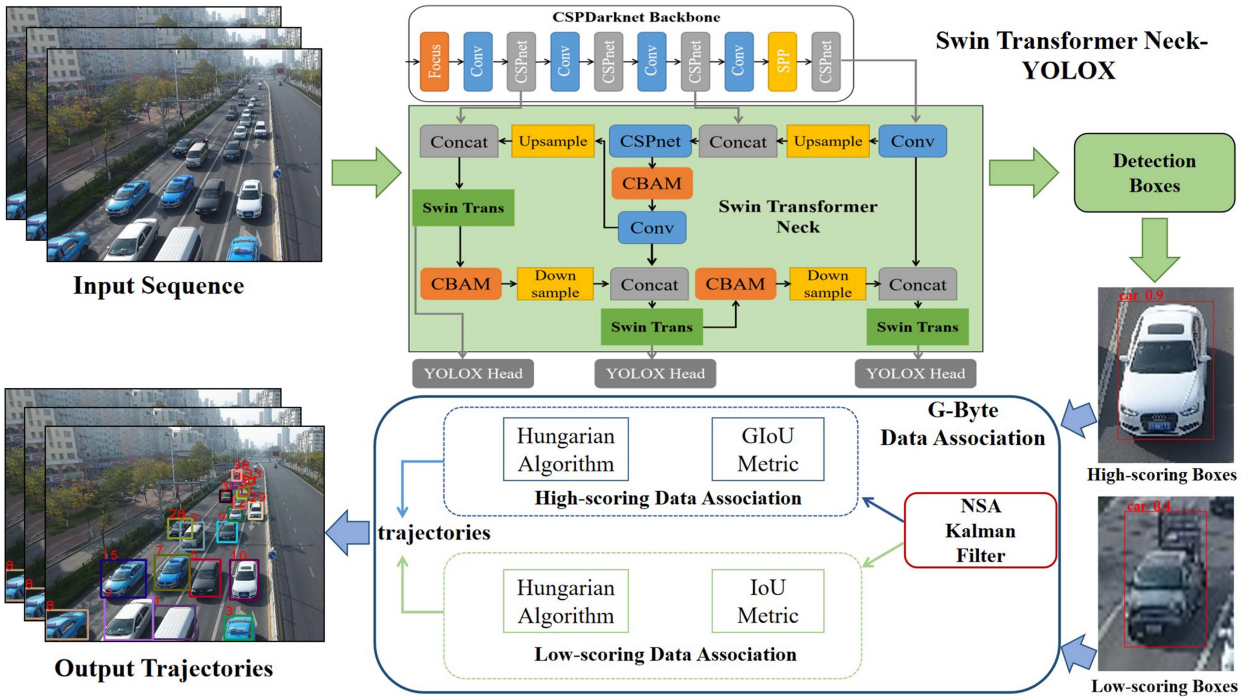


Fig. 1. Overall block diagram of proposed MOT algorithm, STN-Track. Input video sequence is first passed through designed STN-YOLOX to generate detection boxes. After dividing all detection boxes into low-scoring and high-scoring boxes, they are, respectively, sent to the G-Byte data association algorithm after adding GIoU matching and NSA Kalman filter to generate tracking results.

establish global relationships. CNNs are useful in extracting local information, but they lack the ability to extract long-range features from global information. This has a great impact on object detection in remote sensing images with high resolution, complex backgrounds, and small objects. Inspired by the use of self-attention in transformer models [6] to mine long-term correlation dependencies in text, many computer vision tasks involve the use of self-attention mechanisms to effectively overcome the limitations of CNNs, such as vision transformer (ViT) [7]. Self-attention mechanisms can more quickly acquire relationships between distant elements, pay attention to different regions of an image, and integrate information across the entire image.

Data association is the core of MOT. It first identifies objects through the motion and appearance models, then calculates the position distance and feature distance to measure the similarities between detection boxes and tracking boxes, and finally matches according to the similarities. At present, most of the classic MOT algorithms, such as SORT [8] and DeepSORT [9], will select a detection threshold, and only keep the detection results with a confidence score higher than this threshold for data correlation to obtain tracking results, and detection results below this threshold are directly discarded. However, due to the problems of small-scale objects, mutual occlusion of objects, and complex and changeable backgrounds in UAV remote sensing videos, simply discarding these low-score detection boxes will cause missed detection and trajectory interruption for the MOT algorithm, reducing the overall tracking performance.

In this article, we designed a novel MOT algorithm to solve these problems. The overall block diagram of the proposed MOT algorithm is shown in Fig. 1. The main contributions of this article can be summarized as follows.

1) We first propose a new object detection network named Swin transformer neck-YOLOX (STN-YOLOX). This network combines the advantages of CNN and the transformer network algorithms to improve the global information extraction ability.
2) Second, we designed a new data association method, named G-Byte. G-Byte retains all detection boxes and divides them into high-scoring and low-scoring boxes by confidence, using the noise scale adaptive (NSA) Kalman filter [10] and the generalized intersection over union (GIoU) [36] metric to predict the position of the trajectory in the new frame.
3) Finally, we used the designed STN-YOLOX object detection algorithm as the detection module, and the G-Byte association method as the tracking module. We combined the two into a new MOT algorithm, named STN-Track. We used STN-Track to conduct experiments on the UAVDT [11] and VisDrone [12] MOT datasets.

II. Related Work

A. Object Detection Model

In recent years, CNN-based object detection models have been favored by both academia and industry due to their high robustness and efficient performance [40], [41], [42]. Object detection algorithms are divided into one-stage and two-stage object detectors according to whether an R-CNN is required. Among them, the two-stage object detector needs a certain area

generation network part, so the running speed of the algorithm is limited: Fast R-CNN [13], based on R-CNN, involves the concept of a region of interest pooling layer, which can map the feature maps of candidate regions of different sizes to fixed-size feature maps; Faster RCNN [14] uses the CNN-based region proposal network to take an image feature map as input, and then output a series of candidate regions.

The one-stage object detector does not need the area generation network part, so the algorithm generally runs faster but the accuracy is slightly lower than that of the two-stage detector. On the basis of a feature pyramid network (FPN) [15], Lin et al. [16] proposed RetinaNet, which further improved the performance of the single-stage object detection algorithm. The YOLO series algorithm is an example of a single-stage algorithm. YOLOv3 [17] is the third version of the YOLO series, which improves the speed and accuracy of object detection in three ways: multiscale feature detection, multilabel task, and anchor box clustering. YOLOv4 [18] is based on YOLOv3 and uses the cross-stage partially connected Darknet (CSPDarknet) [19] and the PANet [20] to improve model performance. YOLOX [31] is a new YOLO network proposed in 2021. It adds advanced detection techniques such as anchor-free method, decoupled head, data enhancement, and the SimOTA label assignment strategy [21] based on the original, thus realizing a better trade-off between accuracy and speed. However, due to the relatively weak ability of CNN to capture distant features, the problem of establishing global relationships in images has not been solved, so the effect is not satisfactory when applied to remote sensing images.

The transformer model is a new type of deep neural network that has emerged in recent years. It was initially applied in the field of natural language processing and later extended to computer vision tasks. The transformer's network structure is composed of only the self-attention mechanism and the feedforward neural network, completely avoiding the CNN network structure. Compared with a CNN network, the advantage of the transformer is to use self-attention to capture global contextual information. ViT [7] is a representative state-of-the-art (SOTA) model in the field of image recognition. It only uses a self-attention mechanism, which makes the image recognition rate far higher compared with models based on CNNs. In 2020, Nicolas et al. [22] combined CNN and transformer to propose a complete end-to-end DETR object detection framework, applying transformer architecture to object detection for the first time. Zhou et al. [23] proposed the deformable DETR model, which draws on the variable CNN. Zheng et al. [24] proposed end-to-end object detection with an adaptive clustering transformer to reduce the computational complexity of the self-attention module. Due to the large amount of computation of transformer models, Liu et al. [34] proposed the Swin transformer to solve the problems of traditional transformer models with large amounts of computation and poor detection of dense objects.

Due to the high resolution of remote sensing images, the Swin transformer is very suitable for remote sensing object detection tasks, but its ability to collect local information is still weak. Therefore, we need to design a new object detection framework that combines the advantages of CNNs for processing the underlying vision and transformers for processing the relationship between visual elements and objects.

B. Data Association Tracking Algorithm

Data association first calculates the similarity between trajectories and detection boxes, and then matches according to the similarity. Feature model and similarity metrics are both important parts of data association. Among them, the motion model is used to predict the position of objects in video frames. It predicts tracking boxes of the current frame from the previous frame and matches the detection boxes in the current frame to achieve continuous tracking of objects. The appearance model aims to learn the discriminative features of objects, so that the same object features in different frames are more similar than different object features. The similarity metrics measure the similarity between the detection and the tracking boxes by calculating the feature distance and position distance. Frequently used metrics include IoU metric, Mahalanobis distance, and cosine distance [25].

SORT [8] uses the Faster R-CNN object detector, Kalman filter prediction to predict and update motion trajectories to track boxes, and the IoU metric as the matching criterion. Based on the SORT algorithm, the DeepSORT [9] algorithm does an additional cascade matching before IoU matching, adds deep appearance features, and extracts them as an embedded layer through the reidentification (Re-ID) network. This method can alleviate the occlusion problem to a certain extent and reduce the amount of ID switching. To improve the DeepSORT algorithm, the innovation of the MOTDT algorithm [26] is that it introduces a trajectory scoring mechanism. The longer the trajectory, the higher the reliability. The JDE algorithm [27] is improved based on MOTDT. The main advantage is that the detection and embedding networks are combined in the feature extraction stage to achieve a balance of speed and accuracy. Based on the JDE algorithm, FairMOT [28] improves the anchor-free method of object detection and introduces multilayer feature aggregation to deal with the problem of insensitivity to scale changes. ByteTrack [29] proposes a simple and efficient data association method, which separates high-scoring and low-scoring boxes and mines more real objects from the latter. ByteTrack has achieved SOTA results so far on the MOT20 [39] dataset. It can be seen from the above-mentioned tracking algorithms that the current research direction is focused on how to design better data association methods. Given that the DBT algorithm is more dependent on the object detection module, the detection ability and speed of detection algorithms are equally important.

III. Methodology

A. Swin Transformer Neck-YOLOX

The network structure of the STN-YOLOX object detection algorithm is shown in Fig. 2. The STN-YOLOX network consists of three parts: the basic YOLOX network framework, the network neck that integrates the Swin transformer block and the convolutional block attention module (CBAM) [30], and the network prediction head. The back end of the model performs feature map classification and bounding box regression tasks. In our model, each bounding box is divided into object and nonobject regions. The details of each module are described in the following.

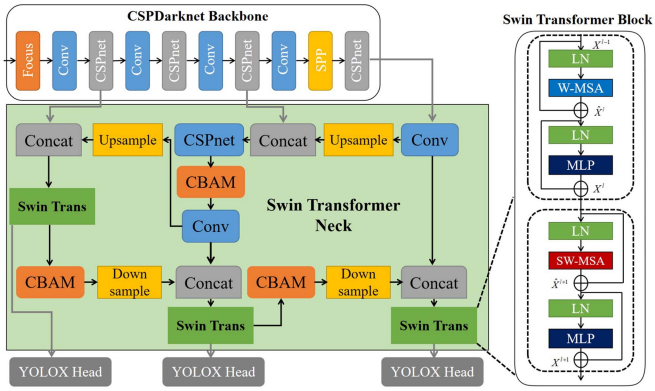


Fig. 2. Architecture of the STN-YOLOX. Left side shows detailed structure of the Swin transformer neck (STN); right side shows detailed structure of the Swin transformer block.

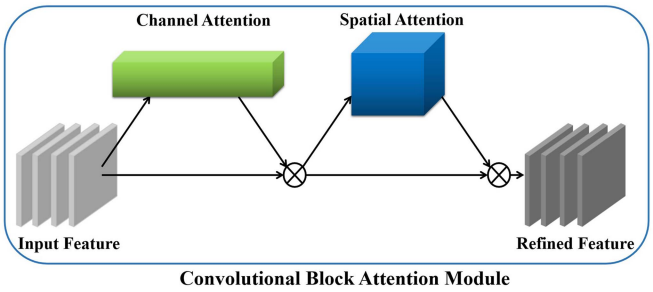


Fig. 3. Structure of the CBAM. Input feature maps pass through channel attention and SAMs in turn to obtain refined feature maps.

YOLOX [31] converts the YOLO family detectors to an anchor-free approach and performs other advanced detection methods. There are four models of YOLOX — YOLOX-S, YOLOX-M, YOLOX-L, and YOLOX-X and the number of model parameters increases sequentially. In view of the importance of tracking speed in UAV MOT tasks, we chose YOLOX-S as the basic network. The backbone feature extraction network of YOLOX is CSPDarknet [19], and the specific structure is shown in Fig. 2. CSPDarknet has four important features: it uses residual network, CSPNet network structure, focus network structure [32] to increase the number of feature layer channels, and SPP structure [33]. The backbone part of CSPNet still uses the original residual blocks, while the other part, such as the residual edge, is directly connected to the end after some processing. The SPP structure performs feature extraction through maximum pooling of different pooling kernel sizes to improve the receptive field of the network.

CBAM performs attention mapping in the channel and space dimensions, which can help the network pay more attention to identifying objects. CBAM includes two submodules, a channel attention module and a spatial attention module (SAM), as shown in Fig. 3. CBAM sequentially infers the attention map along the two submodules from the input feature map, and then multiplies the attention map with the feature map. This design can not only save parameters and computing power, but also ensure that the module can be integrated into other existing network architectures. Therefore, we added the CBAM to the STN to enhance the network's ability to find regions of interest in large-area images. In order to expand the receptive field of the network without increasing the amount of parameters, we replaced some convolutional layers of the CBAM module with dilated convolutions. We put the CBAM module behind the Swin transformer module to ensure that the Swin transformer module can learn the original feature maps.

B. Swin Transformer Neck

The network neck of YOLOX uses a structure similar to PANet [4], combining PAN and FPN. The network neck aggregates the parameters of the three feature layers generated by the backbone network from different backbone layers to different detection layers. However, for remote sensing images, ordinary CNN networks have poor ability to solve problems such as small-scale objects and low resolution. Inspired by ViT [13], we add the Swin transformer block to the network neck of YOLOX to replace some CSPNet modules of the original network. The Swin transformer [34] improves the network's ability to capture global information and reduce the amount of computation as much as possible. As a general vision backbone network, the Swin transformer achieves SOTA performance in tasks such as object detection and semantic segmentation. In recent years, the application of the Swin transformer in remote sensing object detection and instance segmentation has performed outstandingly [35].

Since the resolution of remote sensing images is usually large, traditional transformer models will have large amounts of computation. Considering the model detection speed, we introduced the Swin transformer to solve the problem that the traditional transformer has a large amount of calculation and a poor detection effect on dense objects. The structure of the Swin transformer block is shown in Fig. 2. The block consists of window multihead self-attention (W-MSA), shifted-window multihead self-attention (SW-MSA), and a multilayer perceptron. LayerNorm layers are inserted in the middle to make training more stable, and a residual connection is applied after each module. This part can be expressed as follows:

$$\begin{aligned}\hat{X}^{l} &= W-MSA(LN(X^{l-1}))+X^{l-1}\\ X^{l} &= MLP(LN(\hat{X}^{l}))+\hat{X}^{l}\\ \hat{X}^{l+1} &= SW-MSA(LN(X^{l}))+X^{l}\\ X^{l+1} &= MLP(LN(\hat{X}^{l+1}))+\hat{X}^{l+1}.\end{aligned} \tag{1}$$

As shown in Fig. 4, W-MSA and SW-MSA use a special calculation method. W-MSA controls the calculation area of the MSA [6] of the traditional transformer model within the range of windows (window size is set to 7 by default). This calculation method greatly reduces the computational complexity of the network and reduces the complexity to a linear scale of the image size. Since W-MSA lacks connections across windows, SW-MSA needs to provide different window segmentation after W-MSA to realize cross-window communication. The implementation process is shown in Fig. 4. The result of window segmentation of the input image through W-MSA is shown in Fig. 4(b). Then, the image is moved up and left circularly by half the size of the window, and the blue and red areas are moved

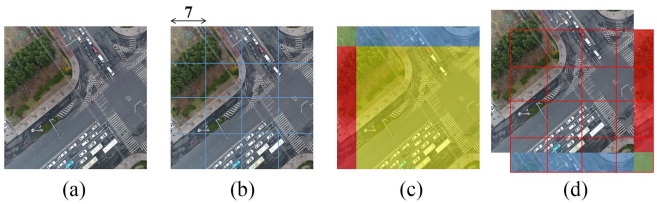


Fig. 4. Mechanism of action of shifted windows (a) input image, (b) window segmentation (window size is set to 7) of input image through the W-MSA), (c) action of the shifted windows, and (d) different window segmentation method through the SW-MSA.

Algorithm 1: Pseudo-Code of **G-Byte**.

Input: UAV video sequence V, detection score threshold τ_{high} τ_{low}, object detector Det, tracking score threshold ε, NSA Kalman Filter $NSA-KF$

Output: Tracks T of UAV remote sensing video

1. Initialization: $T \leftarrow \emptyset$; $d_{high} \leftarrow \emptyset$; $d_{low} \leftarrow \emptyset$
2. **For** *frame* f_k in V **do**
3. $d_k \leftarrow Det(f_k)$
4. **For** d in D_k **do**
5. **If** d.scoce $> \tau_{high}$ **Then**
6. $d_{high} \leftarrow d_{high} \cup \{d\}$
7. **else if** d.scoce $> \tau_{low}$ **Then**
8. $d_{low} \leftarrow d_{low} \cup \{d\}$
10. **end**
// *predict tracks* //
11. **For** t in T **do**
12. t $= NSA-KF$(t)
13. **end**
// *first association* //
14. Associate T and d_{high} using GIoU metric
15. $d_{unmatched} \leftarrow$ unmatched object boxes from D_{high}
16. $T_{unmatched} \leftarrow$ unmatched tracks from T
// *second association* //
17. Associate $T_{unmatched}$ and d_{low} using IoU metric
18. $T_{(re\text{-}unmatched)} \leftarrow$ unmatched tracks from $T_{unmatched}$
19. Delete tracks $T_{re-unmatched}$
// *initialize new tracks* //
20. **For** d in $d_{unmatched}$ **do**
21. **If** d.scoce $> \varepsilon$ **Then**
22. $T = T \cup \{d\}$
23. **end**
24. **Return** T;

to the lower and right sides of the image. The implementation process is shown in Fig. 4(c) and (d). Finally, the window is divided again based on the shifted image, and a window segmentation method different from that of W-MSA will be obtained.

C. G-Byte Data Association Method

Most of the current MOT algorithms focus on optimizing data association, while ignoring that the results of the object detection module can have a large impact on MOT tasks. Due to problems of complex environmental interference, small and dense objects, and motion blur in UAV videos, there are many low-scoring detection boxes. If detection boxes with a low confidence score are directly ignored, problems such as missed detection and trajectory departure can occur. Therefore, inspired by the ByteTrack [29] algorithm, we designed a new UAV remote sensing data association algorithm, G-Byte. G-Byte used the GIoU metric method to associate data between high-scoring detection boxes and used NSA Kalman filter for trajectory prediction.

The algorithm keeps all detection boxes and divides them into high-scoring and low-scoring boxes. The pseudocode of the algorithm is shown in Algorithm 1. The level of confidence is divided into high-scoring boxes (detection boxes with scores greater than τ_{high}) and low-scoring boxes (those with scores between τ_{high} and τ_{low} are low-scoring boxes). In the first data association, we used high-scoring boxes to match new trajectories generated by NSA Kalman filter. Similarity was calculated using the GIoU metric between detected and predicted boxes. Since IoU only considers the overlap ratio between two boxes, it cannot reflect the distance and intersection between them, so on this basis, we used GIoU to solve this problem (lines 14–16 in Algorithm 1). In the second data association, we matched low-scoring boxes with unmatched trajectories after the first step data association (lines 17–18 in Algorithm 1). We deleted the unmatched detection boxes the second time and treated them as backgrounds. We kept tracks that were not matched and removed them from tracks when they remained more than 30 frames. Finally, we filtered the unmatched high-scoring boxes from the first data association. When the score of the detection box is greater than tracking score threshold ε and it appeared in more than two frames, we treated it as a new object and generate a new trajectory (lines 20–23 in Algorithm 1).

Throughout the algorithm, we used the NSA Kalman filter for trajectory prediction. As shown in (2), the Kalman filter generates a Kalman gain K_K when making predictions. K_K consists of the predicted estimated covariance P_K, the measurement model H_K, and the prefitted covariance C_K. C_K consists of measurement noise covariance R_K, P_K, and H_K. In the traditional Kalman filter algorithm, the noise scale is a constant matrix. However, in essence, the measurement noise level should vary with the detection confidence. Therefore, detection confidence S_K was added to the NSA Kalman filter to generate adaptive measurement noise covariance $\tilde{R}_K$. $\tilde{R}_K$ can reduce the influence of noise, thereby improving tracking accuracy.

$$\begin{aligned} K_K &= P_K H_K^T C_K^{-1} \\ C_K &= H_K P_K H_K^T + R_K(\widetilde{R}_K) \\ \widetilde{R}_K &= (1 - S_K) R_K. \end{aligned} \tag{2}$$

At the same time, we combined the GIoU and IoU metrics in our algorithm. IoU only considers the overlap ratio between two boxes, and cannot reflect the distance and intersection. The GIoU metric is proposed to solve this problem. The difference between these two measurement methods is described in (3). In the formula, the area of the detection box is A, the area of the prediction box is B, and C is the minimum rectangular frame

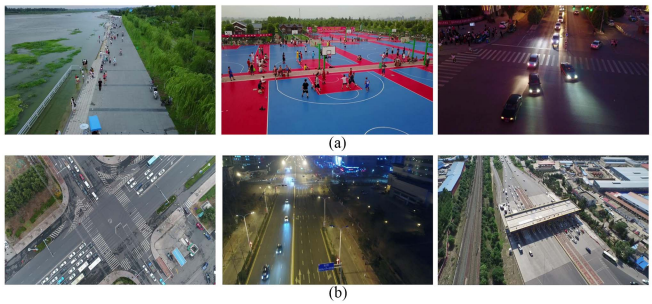


Fig. 5. UAVDT and VisDrone MOT datasets displays. (a) Portion of VisDrone dataset. (b) Portion of UAVDT dataset. From these examples, we can see difficulties for UAV MOT, such as small-scale objects, low resolution, and complex backgrounds. (a) VisDrone-MOT dataset. (b) UAVDT dataset.

area that can contain A and B. GIoU adds a measure of how the two boxes intersect based on IoU, and considers the area other than A and B in C.

$$S_{IoU} = \frac{A \cap B}{A \cup B}$$
$$L_{IoU} = 1 - S_{IoU}$$
$$S_{GIoU} = S_{IoU} - \frac{C - A \cap B}{C}$$
$$L_{GIoU} = 1 - S_{GIoU}. \quad (3)$$

In the whole algorithm, we do not use the feature models commonly used by other algorithms for matching. There are two reasons for this: first, UAV MOT has high requirements regarding tracking speed. For feature matching, the Re-ID network needs to be added to the algorithm. This operation will slow down the tracking algorithm. Second, in UAV videos, when the shooting angle is changed, the appearance characteristics of objects will change greatly based on their relative orientation to the camera. Adding Re-ID has little effect on the tracking accuracy of the network, which was proved in subsequent ablation experiments.

IV. Experiment

A. Dataset and Metrics

As shown in Fig. 5, we selected two large-scale, publicly available UAV remote sensing MOT datasets for training and testing our designed algorithm. The UAVDT dataset [11], proposed by ICCV2018, contains a total of 80 000 frames of pictures, which can be used for object tracking as well as object detection. The UAVDT dataset focuses on complex scenes and contains 50 video sequences with more than 80 000 frames. The dataset contains three object categories, cars, buses, and trucks. We used 40 video sequences in the dataset as a training set and 10 video sequences as a test set for experiments. The VisDrone MOT dataset [12] was collected by the Machine Learning and Data Mining Laboratory of Tianjin University. The dataset provides 96 video sequences, including training video sequences (56 in total, 24 201 frames), validation video sequences (7 in total, 2819 frames), and test video sequences (33 in total, 12 968 frames). The video sequences captured by different drone cameras covering various aspects, including location (14 cities in China), environment (city and country), objects (pedestrians, cars, buses, etc.), and density (sparse and crowded scenes). We chose five object classes to track experiments: pedestrians, cars, vans, buses, and trucks.

TABLE I
Initial Learning Rate Ablation Study

Method	Initial learning rate	AP (%)	AP_{50} (%)
STN-Track	0.001	21.4	36.8
	0.01	23.2	38.9
	0.02	26.0	41.3
	0.04	**26.3**	**42.9**
	0.1	24.0	40.6

The bold entities indicate the best result of the comparison methods.

We quantitatively analyzed the proposed MOT algorithm by various metrics [37]: average precision (AP), AP_{50} (AP value when IoU threshold is 0.5), AP_S (the average value of recall measurement of object frames smaller than 32 × 32 pixels), false positive (FP), false negative (FN), identification precision (IDP), identification recall (IDR), IDF1 score, MOT accuracy (MOTA), mostly tracked objects (MT), mostly lost objects (ML), ID switching (IDSW), and frames per second (FPS). Among them, AP, AP_S, and AP_{50} are used for object detection tasks, and MOTA, IDF1, IDSW, MT, ML, FP, FN, and FPS are used for MOT tasks. MOTA and IDF1 are two more important evaluation metrics. MOTA is a comprehensive evaluation of FP, FN, and mismatch rate, while IDF1 combines IDP and IDR.

B. Implementation Details

Our experimental hardware platform is a computer equipped with GEFORCE RTX 3060 GPU (12 G), and the compilation environment used by the computer is python 3.8 and PyTorch 1.8.1. The experimental training parameters were as follows: training schedule of 48 epochs, optimizer is SGD with weight decay of 5×10^{-4} and momentum of 0.9, batch size of 8, the initial learning rate is 0.04 with cosine annealing schedule. For G-Byte, the high detection score threshold τ_{high} is 0.6, the low threshold τ_{low} is 0.1, and the tracking score threshold ε is 0.7.

C. Study for Initial Learning Rate and Detection Score Threshold

We conducted ablation experiments with MOT algorithm parameter settings. The initial training learning rate and the detection score threshold τ_{high} are crucial hyperparameters in the tracking algorithm. Their selection is directly related to the final effect of the algorithm.

We used the STN-Track algorithm we designed for training, and set the initial learning rate to 0.001, 0.01, 0.02, 0.04, and 0.1 to conduct comparative experiments. After the optimal initial learning rate was determined, we set the detection score threshold τ_{high} to 0.5 to 0.8, and observed the impact of different parameters.

TABLE II
DETECTION SCORE THRESHOLD ABLATION STUDY

Method	Detection score threshold τ_{high}	MOTA (%)	IDF1 (%)
STN-Track	0.5	59.3	72.6
	0.6	**60.4**	**73.7**
	0.7	57.9	71.3
	0.8	52.5	69.9

The bold entities indicate the best result of the comparison methods.

TABLE III
DETECTION PERFORMANCE OF DIFFERENT METHODS

Method	Dataset	Parameter	AP (%)	AP_{50} (%)	AP_S (%)	Epochs
YOLOX-S	UAVDT	**8.94M**	23.3	38.0	18.9	48
	VisDrone		37.6	56.9	15.5	
YOLOX-M	UAVDT	25.28M	25.5	42.8	21.0	50
	VisDrone		38.8	58.6	16.7	
YOLOX-L	UAVDT	56.30M	26.0	**43.3**	21.4	55
	VisDrone		**41.9**	**61.1**	17.2	
YOLOX-S + Transformer	UAVDT	9.05M	24.0	39.3	20.8	48
	VisDrone		38.9	58.6	16.9	
YOLOX-S + STN	UAVDT	9.11M	24.8	40.7	22.1	48
	VisDrone		39.3	59.3	17.9	
STN-YOLOX	UAVDT	9.13M	**26.3**	42.9	**23.3**	48
	VisDrone		40.9	60.7	**19.8**	

The bold entities indicate the best result of the comparison methods.

It can be seen from Table I that when the initial learning rate is 0.04, the AP and AP_{50} values obtained by object detection are the highest. Therefore, setting the initial learning rate to 0.04 can make the object detection module of the network achieve the best results on the UAVDT dataset. From Table II, we can see that when the initial learning rate is constant, setting the detection score threshold τ_{high} to 0.6 can result in the highest MOTA and IDF1 scores. Therefore, when the detection score threshold is set to 0.6, the MOT algorithm can achieve the best tracking effect on the UAVDT dataset.

D. Experiment With STN-YOLOX

We conducted ablation experiments on the UAVDT test set and VisDrone MOT test-dev dataset to verify the importance of each component in the designed STN-YOLOX object detection network. The specific content of the experiment is shown in Table III. In the experiment, we gradually added the STN and CBAM to the YOLOX-S network, and compared with YOLOX-M, YOLOX-L, and YOLOX-S with traditional transformer modules. It can be seen from the table that after adding the STN, the detection effect of the network is improved, and the detection effect of small objects is better than the traditional transformer. After adding CBAM to the network, the detection effect increases considerably. For YOLOX-L, although it has good results in detection, the number of network parameters is too large. Having excessive network parameters is not conducive to the real-time object detection and tracking of UAVs.

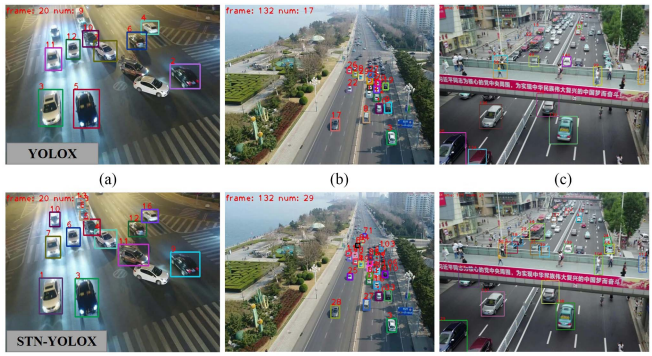


Fig. 6. Comparison of detection results of YOLOX and STN-YOLOX. (a) Night detection scene. (b) Small object detection scene in daytime environment. (c) Blurred scene due to shaking of camera on UAVs. (a) Night scene. (b) Small object scene. (c) Blur scene (camera shake).

Our proposed STN-YOLOX object detection network achieves detection performance similar to YOLOX-L, with almost no increase in the number of network parameters. Compared with YOLOX-S, on the UAVDT dataset, AP of the STN-YOLOX detection network increased by 3.0%, AP_{50} increased by 4.9%, and AP_S increased by 4.4%. On the VisDrone MOT dataset, AP of STN-YOLOX increased by 3.3%, AP_{50} increased by 3.8%, and AP_S increased by 4.3%.

We selected some representative examples from the two datasets for demonstration, as shown in Fig. 6. We used them to compare YOLOX with the STN-YOLOX network. The self-attention mechanism of the transformer can bring the global receptive field to the network, which can improve the network's ability to detect the edge of images. At the same time, the transformer has the ability to associate with the global context, which is very helpful for the localization of small objects. Fig. 6(a) shows a detection scene at night. It can be seen from the image that the STN-YOLOX network has a better edge detection effect for the night scene. Fig. 6(b) shows a small object detection scene in the daytime. We can clearly see that the detection effect of STN-YOLOX on small objects is significantly improved, and there are more detected small objects. Fig. 6(c) is a blurred scene due to the severe shaking of cameras on UAVs. As can be seen from the image, when a scene captured by the camera is blurred, the detection effect of the network is poor, and STN-YOLOX can improve the detection effect somewhat.

E. Comparison of STN-Track With SOTA MOT Algorithm

We combined STN-YOLOX with the proposed data association algorithm, G-Byte, to form a complete MOT algorithm, STN-Track. We compared the proposed tracking network with previous SOTA MOT methods. We selected some representative SOTA MOT methods: SORT [8], DeepSORT [9], MOTDT [26], JDE [38], and ByteTrack [29]. These have all previously achieved SOTA results on the MOT20 dataset. It can be seen from Table IV that the tracking accuracy and speed of the STN-Track algorithm surpassed most algorithms and achieved the best performance on many metrics, including MOTA, IDF1, MT, and IDSW. Except for STN-Track, detection modules of

TABLE IV
COMPARISON OF SPEED AND ACCURACY OF SOTA METHODS

Method	Dataset	MOTA (%) ↑	IDF1 (%) ↑	IDSW ↓	MT (%) ↑	ML (%) ↓	FP ↓	FN ↓	FPS ↑
SORT	UAVDT	55.5	68.1	196	46.0	22.4	9323	72 492	12.02
DeepSORT		56.5	69.1	175	48.8	21.3	10 193	71 194	8.82
MOTDT		56.3	68.8	172	48.3	21.6	**8343**	74 968	7.92
JDE		55.9	69.0	155	49.6	23.0	9105	64 990	8.18
ByteTrack		57.4	70.2	102	50.2	20.9	10 084	69 770	**12.07**
ByteTrack + Re-ID		57.4	71.2	124	48.7	20.1	10 143	71 298	9.08
STN-Track (ours)		**60.6**	**73.7**	**76**	**57.0**	**17.0**	12 825	**61 760**	11.39
SORT	VisDrone	35.5	47.0	1373	27.4	52.7	**7044**	79 980	**6.66**
DeepSORT		36.0	48.6	934	28.4	50.3	10 268	76 511	4.58
MOTDT		35.7	47.4	1129	27.2	51.9	9329	77 672	4.26
JDE		36.2	48.5	998	29.7	**50.1**	8776	78 980	5.37
ByteTrack		36.6	49.5	928	29.0	52.9	6495	79 375	6.42
ByteTrack + Re-ID		36.5	48.0	1104	28.5	51.0	7276	80 797	4.87
STN-Track (ours)		**38.6**	**52.6**	**668**	**31.4**	51.2	7385	**76 006**	6.18

The bold entities indicate the best result of the comparison methods.

all algorithms used YOLOX-S. Compared with the latest MOT algorithm, ByteTrack, on the VisDrone MOT dataset, MOTA of STN-Track increased by 2.0%, IDF1 increased by 3.1%, MT increased by 2.4%, and IDSW decreased by 28.0%. On the UAVDT dataset, MOTA of STN-Track increased by 3.2%, IDF1 increased by 3.5%, MT increased by 6.8%, and IDSW decreased by 25.4%. At the same time, the calculation speed of STN-Track also achieved good results, which was slightly lower on the FPS metric compared with ByteTrack (11.39 versus 12.07 FPS). This indicates that the G-Byte algorithm we designed greatly improves the object identification ability of the algorithm, which is very effective in the scenario where the drone is tracking objects. Compared with the classic algorithm, DeepSORT, MOTA of STN-Track can be improved by about 2.6%–4.1%, IDF1 can be improved by 4.0%–4.6%, MT can be improved by 3.0%–8.2%, and IDSW can be reduced by 28.5%–56.6%. We added the Re-ID network to the ByteTrack, and found that this not only failed to improve the tracking accuracy, but also reduced the tracking speed of the algorithms. From the above-mentioned experimental results, it can be concluded that STN-Track has superior tracking accuracy and speed when dealing with MOT tasks in UAV videos.

In order to show the advantages of the STN-Track MOT algorithm more clearly, we used that algorithm and the ByteTrack algorithm to conduct a more detailed comparison experiment. We selected representative UAV videos that contain some classic challenges of MOT tasks. The images in Fig. 7 are from the UAVDT dataset. We extracted three frames from the video for analysis. It can be seen from the images that the objects to be tracked in the video are relatively small and easily occluded by other objects. At frame 213, both algorithms detected a black car at the same time. At frame 256, the black car was not detected by the ByteTrack algorithm because it was blocked by a street light. At frame 293, ByteTrack detected the black car again and determined it as a new object, resulting in the loss of the original tracked object (object ID changed from 39 to 43). In contrast,

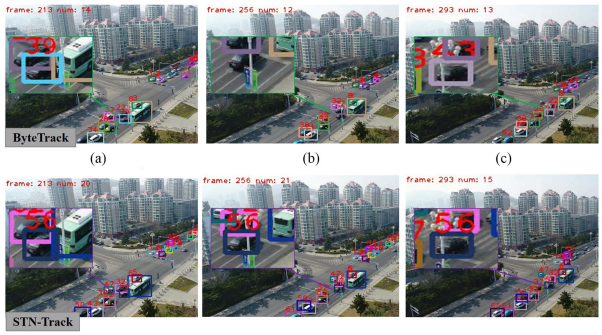


Fig. 7. Comparison of tracking results of ByteTrack and STN-Track. Video frames are from the UACDT dataset, including problems of small objects and object occlusion. (a) Frame 213. (b) Frame 256. (c) Frame 293.

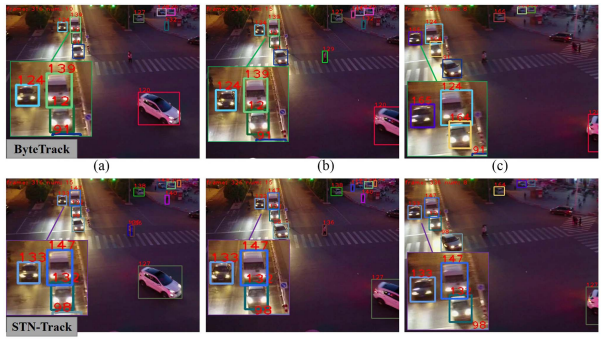


Fig. 8. Comparison of tracking results of ByteTrack and STN-Track. Video frames are from the VisDrone MOT dataset, including problems of shooting angle changes and image blurring. (a) Frame 316. (b) Frame 326. (c) Frame 336.

the STN-Track algorithm never lost the object (object ID 56) from start to finish, even with obstacles.

The images in Fig. 8 are from the VisDrone MOT dataset, and we also sampled three frames from the video for analysis. It can

be seen from the figure that rotation of the UAV leads to a change in the camera's shooting angle and a blurred image. At frames 316 and 326, both the ByteTrack and STN-Track algorithms detected the same four cars in the upper left corner of the video. At frame 336, due to the change in shooting angle and the blurred image, ByteTrack lost track of the three cars in the upper left corner (object IDs changed from 124, 139, 12 to 165, 124, 164). At the same time, STN-Track kept track of these three vehicles, and could track more other objects than ByteTrack. From these two examples, the superiority of STN-Track can be proved. For UAV videos with complex and changeable backgrounds, the improvement of the algorithm greatly improves the ability to identify object, and is very effective in situations such as occlusion by many objects, changes in shooting perspective, and blurred images.

V. Conclusion

In this article, we designed a novel algorithm, STN-Track, for MOT in UAV remote sensing videos. STN-Track consists of two parts: We first made improvements based on the YOLOX algorithm, replaced the original network neck with the designed STN, and proposed a new object detection network named STN-YOLOX. This network combines the advantages of CNN and the transformer network algorithm to improve the global information extraction ability. Second, we designed a new data association method named G-Byte. G-Byte retains all detection boxes and divides them into high-scoring and low-scoring boxes by confidence, and uses the NSA Kalman filter and GIoU metric to improve the tracking accuracy. The following experiments show that the STN-Track can improve the detection and tracking accuracy of small-scale objects and greatly improved identification capabilities for object tracking. In addition, the tracking speed of the algorithm is not inferior to that of most SOTA algorithms.

References

[1] J. Wu et al., "Multiple ship tracking in remote sensing images using deep learning," *Remote Sens.*, vol. 13, no. 18, Sep. 2021, Art. no. 3601.

[2] H. Menouar, I. Guvenc, K. Akkaya, A. Uluagac, A. Kadri, and A. Tuncer, "UAV-enabled intelligent transportation systems for the smart city: Applications and challenges," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 22–28, Mar. 2017.

[3] G. Ciaparrone, F. Luque Sánchez, S. Tabik, L. Troiano, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61–88, Nov. 2019.

[4] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.

[5] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3150–3158.

[6] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[8] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 3464–3468.

[9] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3645–3649.

[10] Y. Du, J. Wan, Y. Zhao, B. Zhang, Z. Tong, and J. Dong, "A comprehensive framework for MCMOT with global information and optimizing strategies in VisDrone 2021," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 2809–2819.

[11] D. Du et al., "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 370–386.

[12] G. Chen et al., "VisDrone-MOT2021: The vision meets drone multiple object tracking challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 2839–2846.

[13] R. Girshick, "Fast R-CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[15] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[16] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[17] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1–6.

[18] D. Wu, S. Lv, M. Jiang, and H. Song, "Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments," *Comput. Electron. Agriculture*, vol. 178, 2020, Art. no. 105742.

[19] C. Wang, H. Liao, Y. Wu, P. Chen, J. Hsieh, and I. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1571–1580.

[20] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.

[21] Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun, "OTA: Optimal transport assignment for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 303–312.

[22] C. Nicolas, M. Francisco, S. Gabriel, U. Nicolas, K. Alexander, and Z. Sergey, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[23] X. Zhou, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[24] M. Zheng, P. Gao, X. Wang, H. Li, and H. Dong, "End-to-end object detection with adaptive clustering transformer," in *Proc. 32nd Brit. Mach. Vis. Conf.*, Nov. 2021.

[25] H. Seyed, M. Anton, Z. Zhang, Q. Shi, D. Anthony, and R. Ian, "Joint probabilistic data association revisited," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3047–3055.

[26] L. Chen, H. Ai, Z. Zhuang, and C. Shang, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2018, pp. 1–6.

[27] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 107–122.

[28] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3069–3087, Nov. 2021.

[29] Y. Zhang et al., "ByteTrack: Multi-object tracking by associating every detection box," 2021, *arXiv: 2110.06864.*

[30] S. Woo, J. Park, J. Lee, and I. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[31] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv: 2107.08430.*

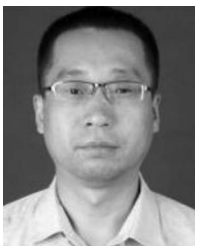
[32] W. Wu et al., "Application of local fully convolutional neural network combined with YOLO v5 algorithm in small target detection of remote sensing image," *Plos One*, vol. 16, no. 10, Oct. 2021, Art. no. e0259283. [Online]. Available: https://doi.org/10.1371/journal.pone.0259283

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[34] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.
[35] X. Xu et al., "An improved Swin transformer-based model for remote sensing object detection and instance segmentation," *Remote Sens.*, vol. 13, no. 23, Nov. 2021, Art. no. 4779.
[36] H. Rezatofighi, N. Tsoi, J. Gwak, Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 658–666.
[37] K. Bernardin and R. Stiefelhagen, "Research article evaluating multiple object tracking performance: The CLEAR MOT metrics," *J. Image Video Process.*, vol. 2008, Feb. 2008, Art. no. 1, doi: 10.1155/2008/246309.
[38] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 107–122.
[39] P. Dendorfer et al., "Mot20: A benchmark for multi object tracking in crowded scenes," 2020, *arXiv: 2003.09003.*
[40] X. Yao, H. Shen, X. Feng, G. Chen, and J. Han, "R^2IPoints: Pursuing rotation-insensitive point representation for aerial object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5623512, doi: 10.1109/TGRS.2022.3173373.
[41] X. Feng, J. Han, X. Yao, and G. Chen, "TCANet: Triple context-aware network for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6946–6955, Aug. 2021.
[42] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

Xiangkai Xu received the bachelor's degree in electronic science and technology in 2020 from Xidian University, Xi'an, China, where he is currently working toward the master's degree.

His main research interests include object detection and object tracking.



Zhejun Feng received the graduate degree in 2008 from Xidian University, Xi'an, China, where he is currently working toward the Ph.D. degree.

He is an Associate Professor. His research interests include photoelectric detection and signal processing.

Changqing Cao received the Dr.Eng. degree from Xidian University, Xi'an, China.

In 2011, he was an Associate Professor with Xidian University. His research interests include laser technology and its applications.

Chaoran Yu received the bachelor's degree in photoelectric information science and engineering from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2021. He is currently working toward the master's degree with Xidian University, Xi'an, China.

His main research interests are target detection and laser detection.

Mengyuan Li received the bachelor's degree in photoelectric information science and engineering from East China Jiaotong University, Nanchang, China, in 2020. She is currently working toward the master's degree with Xidian University, Xi'an, China.

Her main research interests include deep learning and object detection.

Zengyan Wu received the bachelor's degree in photoelectric information science and engineering from the North University of China, Taiyuan, China, in 2018. She is currently working toward the Ph.D. degree with Xidian University, Xi'an, China.

Her main research direction is optical heterodyne target detection.

Shubing Ye received the bachelor's degree in photoelectric information science and engineering from the Changsha University of Science and Technology, Changsha, China, in 2020. She is currently working toward the master's degree with Xidian University, Xi'an, China.

Her main research interest includes photoelectric detection.

Yajie Shang received the bachelor's degree in measurement and control technology and instruments from the Changchun University of Science and Technology, Changchun, China, in 2020. She is currently working toward the master's degree with Xidian University, Xi'an, China.

Her main research interests include generation of high-frequency millimeter waves by remodulation.