ArbiTrack: A Novel Multi-Object Tracking Framework for a moving UAV to Detect and Track Arbitrarily Oriented Targets

Yuqing Chen, Jiayu Wang, Qianchen Zhou, Huosheng Hu, Senior Member, IEEE

Abstract—**The operation of traditional multi-object trackers on a moving unmanned aerial vehicle (UAV) faces many difficulties due to the irregular motion of UAV, the occlusion problem, and in particular arbitrarily oriented targets that are densely distributed with complex backgrounds. To solve these difficulties, this paper proposes a novel multi-object tracking framework, namely ArbiTrack, for a moving UAV to effectively detect and track arbitrarily oriented targets on the grounds. The proposed framework consists of an oriented object detection module to capture ground objects, a multi-scale context aggregation (MCA) module to improve the detection accuracy of small objects, and an adaptive motion switching (AMS) module to deal with the nonlinear complexity among UAV and ground objects. Historical information from multiple moments is used in this framework to learn the spatio-temporal characteristics so that the occlusion problem can be solved effectively. Experiments are conducted by using our OriDrone dataset and the public dataset UAVDT dataset. Results demonstrate that the proposed method achieves state-of-the-art tracking performance.**

Index Terms—**Multi-object tracking, UAV, oriented object detection, multi-scale context aggregation, spatio-temporal evolutionary memory.**

I. INTRODUCTION

WITH the rapid development of computer vision and sensor technologies, multi-object tracking (MOT) technology shows great potential in many application scenarios, such as intelligent surveillance [1], autonomous driving [2], and advanced video analytics [3]. Tracking by detection (TBD) [4], [5], is currently the mainstream tracking paradigm for MOT, aiming to accurately detect objects of interest from video sequences and obtain the corresponding motion trajectories based on the temporal and spatial information or the visual features of the video data. In recent years, it has attracted the attention of many researchers [6], [7], [8], [9] due to the high maneuverability and flexibility of Unmanned Aerial Vehicles (UAV) [10], [11].

Traditional MOT algorithms can effectively detect and track multiple objects in the video captured by fixed or static cameras and show impressive performance. However, they have many challenges to manage multi-object tracking for a moving UAV.

UAVMOT [12] addresses that in the detection phase, a new gradient balanced focal loss to supervise the learning of object heat map (GBF loss) is proposed to address the impact on target detection accuracy due to the high flight altitude of the UAV and the fact that most of the targets in its viewpoint are small. In the tracking phase, a local relational filter is designed to cope with the frequent ID switching due to the irregular motion of the target triggered by the superposition of the camera motion due to the movement of the UAV camera. However, the limitation of the local relational filter is that the number of targets and related positions in the adjacent two frames do not change, but in the actual UAV flight situation, special circumstances such as occlusion will often occur resulting in the change of the number of targets and related positions in the adjacent two frames, which leads to the failure of the local relational filter.

In the detection stage, the object image captured from a moving UAV is usually non-axial, arbitrary direction, densely distributed, and complex background. The traditional object detection model focuses on upright or axially symmetrical objects. The rectangular bounding box predicted by the traditional object detection model cannot accurately describe the shape of the object and may lead to inaccurate positioning or misjudgment. Furthermore, UAVs fly at high altitudes, and the object in their view is small. In the tracking stage, the object motion state observed by an UAV is complex and nonlinear, which consists of the motion states of UAV and ground objects. The traditional Kalman filter cannot be used here for motion modeling.

However, the methods mentioned above have some limitations. In this paper, we propose a novel multi-object tracking framework for a moving UAV to effectively detect and track arbitrarily oriented targets, which is called ArbiTrack. To accurately characterize the shape and position of an arbitrarily oriented object, Oriented RepPoints [13] is used as a detector in our tracking network, which can capture geometrical information of an arbitrarily oriented object from

This work was supported by the National Natural Science Foundation of China under Grant 61203082. *(Corresponding author: Yuqing Chen).*

Yuqing Chen is with the Department of Automation, College of Marine Electrical Engineering, Dalian Maritime University, Dalian 116026, China (e-mail: chen@ dlmu.edu.cn).

Jiayu Wang is with Department of Automation, College of Marine Electrical Engineering, Dalian Maritime University, Dalian 116026, China (e-mail: jiayu_wang2804@163.com).

Qianchen Zhou is with Department of Automation, College of Marine Electrical Engineering, Dalian Maritime University, Dalian 116026, China (e-mail: zqc18768006620@163.com).

Huosheng Hu is with School of Computer Science and Electronic Engineering, University of Essex, Colchester CO43SQ, United Kingdom. (e-mail: hhu@essex.ac.uk).

a moving UAV. It can provide orientation information in addition to the traditional appearance and motion cues for the subsequent data correlation stage. The oriented object detection algorithm is to replace the traditional detection algorithm in MOT tasks. To improve the detection of small-scale objects, we construct a multi-scale context aggregation module (MCA) to fully utilize the contextual information and expand the sensory field, which helps us to capture fine-grained features effectively.

To solve the problem of inaccurate motion estimation due to the irregular motion of UAVs, we propose an Adaptive Motion Switching (AMS) module based on motion state feedback. It can adaptively select different motion estimation filters according to the different motion modes of the UAV to provide accurate motion information for the tracker. Moreover, it can significantly improve tracking in complex variable-speed scenarios, while maintaining high tracking accuracy relative to static scenarios. In addition, to solve the problem of prolonged complete occlusion that occurs during UAV tracking, a Spatio-temporal Evolutionary Memory (SEM) tracking algorithm is created based on Convolutional Gated Recurrent Unit (ConvGRU) [14], which combines the object's historical positional and motion information to learn its spatiotemporal characteristics, and models its spatio-temporal evolution to guide the localization and correlation of the objects in the input video sequences.

Since the bounding box obtained by the oriented object detector carries orientation information, its labeling differs from that of the traditional bounding box and the existing dataset cannot meet the requirements. Therefore, we built our oriented object dataset, OriDrone. Then, our algorithm is evaluated by conducting experiments on OriDrone and the public dataset UAVDT [15]. The main contributions of this article are summarized as follows:

- An MCA module is constructed to aggregate local and global contextual information, enhancing the feature representation of small-scale objects. An oriented object detector, unlike the traditional non-oriented rectangular bounding boxes, is employed to capture objects from the UAV perspective. This detector is not constrained to only upright or axially symmetrical objects but is adaptable to targets of any orientation and shape.
- An AMS module is developed to switch motion estimation filters according to different motion modes of UAV, which can provide accurate motion information for the tracker;
- A SEM tracking algorithm is designed to model the spatio-temporal evolution of the object and to guide the localization and association of objects in the input video sequences, addressing the issue of occlusion from the UAV perspective. Thus, the number of objects and their relative positions can vary between adjacent frames.
- A dataset OriDrone is constructed for oriented object tracking of a moving UAV in many different scenarios. Different from the existed UAV-based datasets, its bounding boxes are labeled with an orientation, which can significantly contribute to developing oriented object detection algorithms in MOT.

The rest of the paper is organized as follows. Section II reviews some related research work on the detection and tracking framework and then discuss the approach taken in the data association stage. In section III, the main framework of ArbiTrack is first described and then three key modules are described: the multi-scale context aggregation network, the adaptive velocity estimation module, and the data association method for learning spatio-temporal features of the object by combining the entire history information of the object. Experiments are conducted in Section IV to verify the feasibility and performance of the proposed framework. Finally, a brief conclusion and future work are given in Section V.

II. Related work

A. Tracking-by-Detection

The tracking by detection (TBD) paradigm divides the MOT task into three stages: object detection, feature extraction, and data association [16]. Specifically, the tracker in the TBD paradigm associates objects detected in different video frames with existing trajectories according to their similarity or initializes a new trajectory to deal with the list of emerging objects. The early MOT algorithm SORT [17] used Faster RCNN [18] as the object detection model, Kalman filter (KF) [19] for object trajectory prediction and Hungarian Algorithm [20] for object and trajectory matching. It has real-time performance, but no be able to solve practical problems due to the frequent switching of object and trajectory during tracking.

Subsequently, Wojke et al. proposed the Deep Sort algorithm [21] based on SORT, which used the Re-identification Network (Re-ID) to obtain the appearance information for data matching. They proposed the idea of cascade matching further to improve the accuracy of the algorithm tracking and matching. Lee et al. [22] designed a ReID network combined with Feature Pyramid Network (FPN) [23] to enrich the obtained information by fusing different levels of features. Sun et al. [24] proposed a deep affinity network for estimating object embeddings and predicting affinities between objects. These trackers performed well in crowded scenes and have good tracking accuracy.

To balance the accuracy and speed requirements of MOT, researchers began to propose a single-stage MOT algorithm. The main framework of the single-stage MOT algorithm was to add an embedded vector to the head of the detector for ReID learning and later data association. For example, JDE [25] first extracted feature vectors from the feature mapping of YOLOv3 [26] and applied automatic balance loss [27] to balance the importance of classification, regression, and Re-ID in the network. JDE was the first real-time single-stage MOT tracker, but its tracking accuracy of JDE does not show a significant advantage in comparison with the previous two-stage method.

Based on the anchor-free detector CenterNet [28], Zhang et al. [29] added the learning of embedded vectors to form the FairMOT Multi-object tracker, which balances the detection and Re-ID tasks by learning low-dimensional Re-ID features. At the same time, it effectively reduced the risk of over-fitting and achieves good accuracy and speed. CenterTrack [30] directly predicted the displacement of the center point of the object. It can simultaneously predict the position of the object and ReID learning in a unified network, avoiding many repeated calculations, thus improving the tracking speed. These methods were based on traditional object detectors and focused on vertical or axial objects from fixed cameras to achieve good tracking results. However, non-axial objects are densely distributed in aerial images with the complex background, which will bring difficulties in object detection and thus affect the tracking effect.

The thermal infrared target tracking is also investigated in recent years and it is not affected by illumination changes to track targets in some extreme weather [31]. There are many thermal infrared tracking methods are widely used, such as machine learning, Siamese Network, discriminative prediction model, multimodal tracks. Although some progress has been made, thermal infrared target tracking still faces many unresolved challenges, e.g., similarity interference, intensity change, deformation, occlusion.

Tracking in the UAV scenarios is one of the main components of target-tracking tasks. Apart from the aforementioned methods, Yuan et al. recently proposed ASTCA [32]. This approach used spatial-temporal context to improve the tracking robustness in UAV scenarios. While it is effective for object detection in small-scale and occlusion-heavy scenes, it focuses primarily on correlation filtering and lacks a mechanism to handle highly nonlinear motion patterns resulting from UAV maneuvers.

Therefore, this paper introduces an oriented object detector to obtain an accurate object box, which can provide a better object description and a more accurate initial tracking result, thereby improving the overall tracking performance.

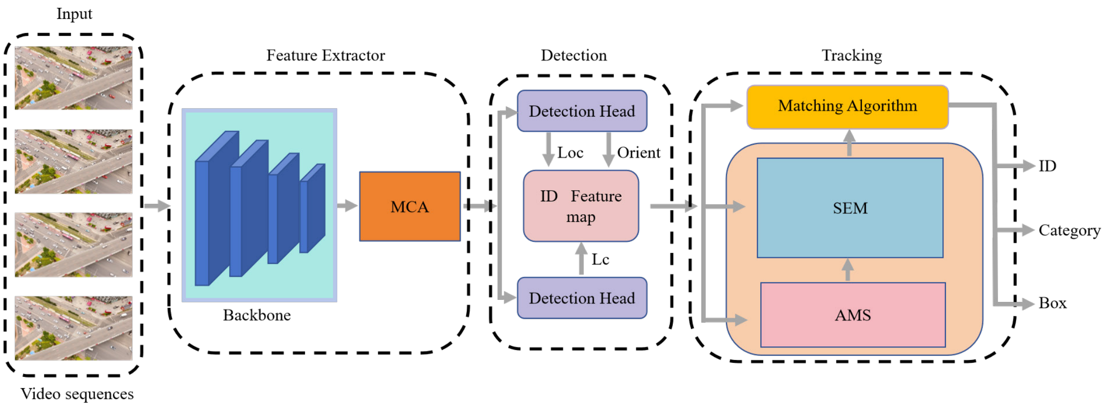


Fig. 1. Architecture of ArbiTrack Network

B. Data association

Data association [33] is a crucial step in MOT, especially in the TBD paradigm. It associates objects detected in different frames and assigns a unique ID number to each object based on the similarity between the trajectory and the detection frame. Data association follows two key cues: motion laws and object features. The object motion laws are commonly used in close-range matching and are characterized using position and motion similarity. Traditional methods [17], [21], [34] mostly used conventional Kalman filters to estimate object position and motion information. Some recent approaches [35], [36], [37] obtained better results with camera motion or low frame rates by designing neural networks to learn object motion.

Nonetheless, the methods discussed previously exhibit certain constraints when viewed from the context of UAV As the object motion law is the composition of the motion states of UAV and ground objects, the traditional Kalman filter cannot accurately estimate such a complex nonlinear motion. At the same time, learning object motion through a neural network could bring a non-negligible computational burden to the tracking network. The AMS module designed in this paper can accurately estimate the motion of the object without introducing a new neural network.

For object appearance, similar to the ReID task [37], most methods extracted the Re-ID features of each object and measured the appearance similarity by the cosine similarity of the Re-ID features to distinguish different objects. For example, MOTDT [39] first used appearance similarity to match and then uses IoU similarity to match unmatched trajectories. QDTrack [40] converts appearance similarity into probability by bidirectional softmax operation and uses nearest neighbor search to complete matching. Appearance similarity is helpful for remote matching and is often used to deal with occlusion problems. However, appearance features are often unreliable, such as motion blur, occlusion, and objects with high appearance similarity.

Recently, as the transformer has demonstrated its power in many computer vision tasks, several researchers [41], [42] have attempted to utilize it for MOT tasks. In the transformer-based MOT methodology each tracked object's ID features and geometric features were treated as a query propagated between frames. For example, Sun et al. proposed TransTrack

[35] based on DETR [43]. This was the first time that the transformer technique was applied to MOT. TransTrack used IOU to match detection and track boxes from two decoders. Meinhardt et al. [41] proposed a tracking query that can follow changes in the object's position, but the information came only from the previous frame of the image.

A significant limitation of these methods is that they are entirely independent of the previous history and only refer to the previous frame for data correlation. Our proposed SEM tracking module uses ConvGRU [14] to combine the entire history of the object with learning the spatio-temporal characteristics of the object and modeling its spatio-temporal evolution, thus predicting both partially and fully occluded object locations and obtaining a more reliable similarity metric.

III. Methodology

The main framework of ArbiTrack is first described in Section A. The proposed ArbiTrack consists of three key modules: the multi-scale context aggregation network constructed in Section B, the adaptive velocity estimation module designed in Section C, and the data association method for learning spatio-temporal features of the object by combining the entire history information of the object as described in Section D.

A. Overall Framework

In this paper, deformable convolution is used to obtain the set of adaptive points as a fine-grained representation in Oriented RepPoints. The detector can capture the critical geometric features of arbitrary direction, clutter, and non-axis aligned objects and proposes an effective method to learn high-quality adaptive points. In the architecture of ArbiTrack Network, shown in Fig. 1, the video sequences from the UAV perspective are used as inputs, and the features are extracted through the backbone network. To obtain richer context information to enhance the ability of small object detection, an MCA module is constructed to replace the feature pyramid structure (FPN) in the original detector.

After passing the detection head, the ID feature map is obtained, and the position of each object in the current frame and the angle of the bounding box are encoded. Subsequently, an adaptive motion switching module (AMS) is designed to accurately estimate the complex motion of the object from the perspective of the UAV. Finally, three components (the ID feature maps corresponding to each time step, the motion information obtained by the AMS module, and the history state of the previous frame) are input into the Spatio-Temporal Evolution Memory (SEM) module to model its spatio-temporal evolution to cope with the long-time occlusion problem.

B. Multi-scale Context Aggregation

The object image captured by a moving UAV is small due to its wide field of view and high altitude. The existing Feature Pyramid Network (FPN) is a valid solution to the problem of small object detection. It manages multi-scale object detection tasks by assigning different scale instances to various levels of feature layers. However, this network may have the problem of misassigning instances of similar sizes to various levels. Furthermore, the importance of features may not be strongly correlated with the level to which they belong, and the detecting loses the opportunity to gather more diverse information at a single feature level. Meanwhile, the multiple branches in FPN increase the computational cost and leads to substantial runtime delays.

In this paper, a multi-scale context aggregation method is designed to deal with the problem of object-scale variation and small object detection from the perspective of UAV. The proposed MCA uses cross-attention to capture local contextual information from neighboring feature layers and global average pooling to obtain global contextual feature vectors. Therefore, it aggregates local and global contextual information from multiple scales to generate features more conducive to identifying small objects. Inspired by [44], [45], fusion factors are used to weigh different input features according to their importance in this method, and the differentiated fusion of these features can deal with the multi-scale variation. The network structure of MCA is shown in Fig. 2.

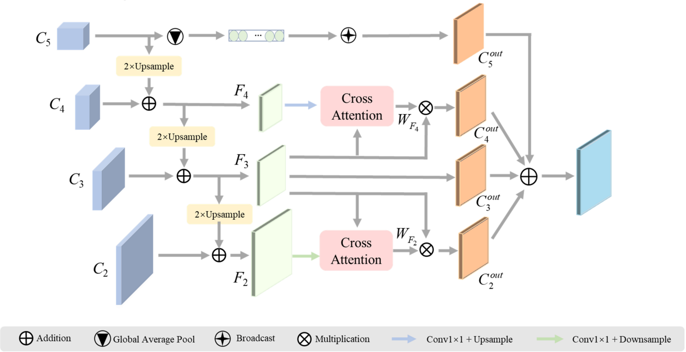


Fig. 2. Architecture of MCA module.

In MCA, the multiscale feature maps C_2, C_3, C_4, and C_5 are the module’s inputs which are outputs from the backbone network stage2, stage3, stage4, and stage5. After top-down feature aggregation, Cross Attention modules are designed to obtained local context information. F_2 and F_4 are reshaped to the same size as F_3, and then the transpose of F_2 is multiplied by F_3, while multiplying F_3 with the transpose of F_4 in another branch. Then the cross-relation weight maps W_{F_2} and W_{F_4} are calculated through the softmax as follows:

$$W_{F_n}^{xy} = \frac{\exp(F_n^x \cdot F_3^y)}{\sum\limits_{x=1} \exp(F_n^x \cdot F_3^y)}, \quad n = 2,4 \tag{1}$$

where F_n^x and F_3^y denote the x-th and y-th channels of F_n or F_3, respectively. $W_{F_n}^{xy}$ represents the element of W_{F_n} at the position (x, y), which expresses the correlation between the x-th channel of F_n and the y-th channel of F_3.

Finally, $C_{\text{out}}^{i}(i = 2,3,4,5)$ are outputs separately in traditional FPN. Unlike this, fusion factors are introduced to adapt and fuse features from different layers in our MCA, as shown in the following equation:

$$C_{\text{out}} = Conv\left(\frac{w_1 \cdot C_2^{out} + w_2 \cdot C_3^{out} + w_3 \cdot C_4^{out} + w_4 \cdot C_5^{out}}{w_1 + w_2 + w_3 + w_4 + \epsilon}\right) \quad (2)$$

where C_{out} is the final output feature, w_i (i =1,2,3,4) are the learnable weights and satisfies $0 \le w_i \le 1$. ϵ is a small value to avoid numerical instability and is usually taken to be 0.0001, and Conv is usually a convolution operation for feature processing.

As the network deepens, the features of small objects will be lost. We believe that localized information around the object is beneficial for small object detection, like a small object on the surface of the water with a high probability of being a boat. Therefore, cross-attention is used to capture the local context information between neighboring layers. After that, the global context feature vector by applying a global average pooling on the C_5 feature map. Finally, the four generated feature maps are aggregated. By using local and global contextual information, MCA effectively improves the representation of feature maps, which is more conducive to coping with scale variations and small object detection problems. Unlike traditional FPN, MCA only output the fused features that combine all the levels for subsequent prediction and no longer need to detect the different feature layers separately.

C. Adaptive Motion Switching Module

When the camera is fixed on a moving UAV, its motion brings new challenges to the multi-object tracking algorithms as the motion of ground objects is not a simple linear motion but a nonlinear motion against the motion states of UAV. The traditional Kalman filter gives the minimum error covariance estimation based on all available observations at the current time step under linear motion. Therefore, the traditional Kalman filter is difficult to deal with this complex superposition motion caused by UAV motion. A novel AMS module is proposed, which is shown in Fig. 3.

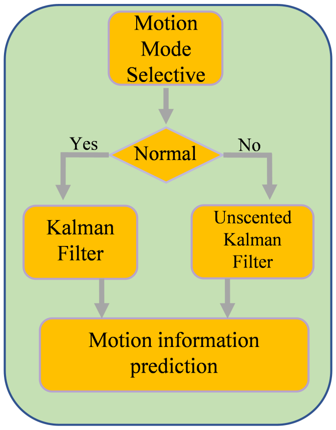


Fig. 3. Architecture of AMS module.

Motion Mode Selection. According to the linear-nonlinear characters of the motion, the motion state can be categorized into two types: normal mode and abnormal mode.

- Normal motion mode: If the UAV flies smoothly in the air at a fixed speed, or the object's speed does not vary suddenly, the motion of the object can be regarded as approximately linear motion. This case can be seen as normal motion mode.
- Abnormal motion mode: if the UAV is droved at a varied speed, such as sudden acceleration, the motion of ground objects shows a complex nonlinear motion from the perspective of the UAV. This case can be seen as abnormal motion mode.

To deal with these two modes, an acceleration threshold τa is presented. When the acceleration of the tracked object is less than τa, it is considered that the motion state does not change drastically and is in the normal mode, and vice versa. Then, Kalman filtering (KF) and Unscented Kalman Filtering (UKF [46]) are employed in the normal mode and abnormal mode to estimate the motion states of the tracked objects, respectively, as shown in Algorithm1.

```
Algorithm 1 Adaptive motion switching algorithm
Input:  Kalman Filter KF; Unscented Kalman Filter
        UKF; Acceleration threshold τa; The set of
        tracked objects T in each frame of UAV video
Output: the velocity and acceleration of the tracked
          object{v, a}
  /*Motion mode judgment, using different Kalman
  filters to estimate the motion information of tracked
  objects*/
for t int T do            //For each tracked object t int set T
    if t.a < τa then      //normal mode
        t←KF(t)
    else                  //abnormal mode
        t←UKF(t)
    end if
end for
return{v, a}              //Output motion information{v, a}
  of tracked objects
```

D. Spatio-temporal Evolutionary Memory

In aerial imagery, the trajectories of the tracked objects are usually complex and irregular and may sometime be obscured by the other objects. For this situation, most traditional MOT algorithms are difficult to deal with these complex trajectories and occlusion problems, resulting a non-continuous and unstable tracking behavior. Currently, more spatio-temporal feature is helpful. Therefore, the spatio-temporal feature extraction is the key to solving the problem for the tracked objects in our ArbiTrack.

GRU [47] is a gated recurrent neural network and exhibits good memory performance but with lower memory requirements compared to LSTM [48]. It has the advantage of being adaptive and capable of modeling non-fixed-length time series. In the basic structure of GRU, a reset gate and an update gate are designed, and the reset gate can control the degree of forgetting of historical information, while the update gate can control the retention of current information.

In SEM, ConvGRU [14] was used in the tracker part. It is an extension of classical GRU, which replaces 1D state vectors with 2D feature maps and uses convolution to compute a fully connected layer of state updates. At each time step t, the corresponding detector outputs ID feature map IDF_t, which was then passed to the ConvGRU with the previous state S_{t-1}. Different from the method in [49], the output motion information (v, a) of AMS was inputted into the ConvGRU as a matrix Mt. It guides the object position prediction of the ConvGRU. Consequently, the state update can be expressed as $S_t = GRU(S_{t-1}, IDF_t, M_t)$. The state matrix S_t includes the positions of the multiple objects $\{O_1, O_2, ...\}$ observed in time series $\{1, 2, ..., t\}$.

Specifically, the detector transmits the position, orientation, and their confidence value of each detected object to the tracker. The position is represented by an Oriented RepPoints point set, the orientation is the rotation angle of the object's bounding box, and the confidence value is the probability that the object is a natural object. Motion information can be used in ConvGRU, and the tracking accuracy and robustness are improved. Thus, in SEM tracker network, the previous historical information is used to learn the spatio-temporal characteristics of the object, model its spatio-temporal evolution, and guide the localization and association of objects in the input video sequence. In addition, it can predict the location of objects seen in the past but currently occluded. When they reappear, they can still be matched into the original tracking trajectory again.

To learn the spatio-temporal features of objects, L1 loss is used as the location prediction loss L_t in the tracker network, as follows:

$$L_t = \frac{1}{N}\sum_{i=1}^{N} | p_i^t - y| \tag{3}$$

where p_i^t is the center point coordinate of object O_i at moment t as predicted by the tracker based on historical information, and y is the valid center point coordinate of that object.

Matching algorithms. With a sufficiently accurate prediction of the object location, inspired by [30], a simple greedy matching algorithm is constructed. The algorithm idea is shown below:

$$\begin{cases} \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \le k & \text{Trajectory association} \\ \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} > k & \text{Trajectory creation} \end{cases} \tag{4}$$

$$k = \frac{\sqrt{w^2 + h^2}}{2} \tag{5}$$

where (x_1, y_1) is the center position of the object predicted by the tracker, and (x_2, y_2) is that detected by the detector in the current frame. Threshold k is computed from the width and the height of the object's bounding box.

If there is corresponding prediction within a radius k for the tracker, the predicted position is greedily associated with the closest detected position. Otherwise, a new trajectory is generated. The simplicity of this greedy matching algorithm once again highlights the advantages of our tracker. Simple position prediction is enough to link objects across time. No complex feature distance measurement or feature similarity matching is required.

When the object is occluded, the ConvGRU prediction is directly stored in the state matrix S, which is used as the tracking result to predict the object position of the next frame. The object's position and orientation are saved if it is occluded in the previous frame. When the object is recaptured by the detector near the center position of the object predicted by the tracker, the current detection result is checked for match with the ConvGRU prediction. Firstly, the distance to the center coordinate position is judged. Then, the probability of bounding-box matching is reflected using similarity Q_{cor} :

$$Q_{cor} = \mathcal{F}_t\left(b_j, b_t\right) \tag{6}$$

where $\mathcal{F}_t$ denotes the CIoU between the detection box b_j and the tracker's prediction box b_t . The CIoU improves the accuracy and robustness of the matching task by considering the bounding box's aspect ratio instead of judging its similarity based on the overlap area alone in IoU function.

IV. Experiments

A. Datasets and Metrics

Datasets. To validate the effectiveness of ArbiTrack, we conducted a series of experiments on our self-built OriDrone dataset and the public UAVDT dataset [15].

The annotation method for our OriDrone dataset was different from the existing UAV tracking datasets. Since ArbiTrack introduced the oriented object detection algorithm as a detector, a novel UAV-tracking dataset was required with the oriented object box annotation. The object tracking dataset was collected based on the DJI Spirit UAV Phantom 4 Pro, equipped with a 20-million-pixel CMOS camera. Many tracking experiments were conducted under different road conditions, different heights, and different UAV-speed and abundant object tracking data was obtained from the perspective of the UAV. The video resolution of the camera was 1280 × 720 and specialized for vehicle objects. In MOT task, it was divided into a training set (30 sequences) and a test set (10 sequences) consisting of various common scenarios, including parking lots, intersections, and viaducts. Each object was annotated by an oriented bounding box, category, and tracking ID in each frame.

The public UAVDT dataset is dedicated to vehicle detection and tracking, which consists of three categories: cars, trucks, and buses. It is divided into a training set (30 sequences) and a test set (20 sequences). The resolution of the video is 1080 × 540, including various familiar scenes such as squares, main streets, and toll stations.

Metrics. To comprehensively evaluate ArbiTrack with other state-of-the-art methods, several metrics were used to evaluate the performance of tracking [50], such as multi-object tracking accuracy (MOTA), Multi-object tracking precision (MOTP), ID switching (IDs). MOTA was the most critical metric. MOTA considered various errors in the tracking process as defined below:

$$\mathrm{MOTA} = 1 - \frac{\mathrm{FN} + \mathrm{FP} + \mathrm{IDs}}{\mathrm{GT}} \quad (7)$$

B. Implementation Details

Training. In the detector part, the original settings of Oriented RepPoints are maintained. Here, we provided hyperparameter values about the tracker network referring to [48], ConvGRU had a feature dimension of 256 and used 7 × 7 convolutions. Sequences of captured video frames were used to train the network. Specifically, they were divided into segments at a fixed time step and labeled to obtain the information sequences of the bounding boxes, which was used for training. The Adam optimizer was used to train the model for 30 epochs with a batch size 6 on information sequences. The learning rate was set to 1.25e-4, which was decayed by a factor of 10 at 10 epochs and 20 at 20 epochs, respectively. The ArbiTrack was trained and evaluated on a single GeForce RTX 3090 GPU.

C. Comparison with State-of-the-arts

OriDrone dataset. Due to the specificity of our dataset, the representation of its object boxes is oriented and different from the representation of regular rectangular object boxes. To obtain more unbiased results, our dataset was labeled with normal object boxes to facilitate the training of other methods. Then, we compared it with the existing methods in the test set of our dataset. The results of our method evaluated on the dataset are shown in Table I, where our method achieved 44.6% in MOTA and 56.9% in IDF1. ArbiTrack surpasses the previously representative tracker FairMOT [29] by 2.4% on MOTA and 2.1% on IDF1. Compared to TrackFormer [41], a tracker using transform, our method outperforms it by 6.7% on MOTA and 11.7% on IDF1. Meanwhile, ArbiTrack outperformed other trackers on IDs, showing that our method achieved better tracking performance in the face of the complexity of the UAV view. In addition, our tracker also achieved the best result on MT, demonstrating the high completeness of the tracking trajectory of our method and proving its strong robustness.

UAVDT dataset. To further validate the effectiveness of our ArbiTrack, we compared ArbiTrack to other methods on the UAVDT test set for MOT task. We relabeled the training set of the UAVDT so that it could be used to train our network. As can be seen, our method achieved 47.6% on MOTA and 67.4% on IDF1. ArbiTrack outperformed the previous most representative tracker FairMOT by 2.7% on MOTA and 6.5%

TABLE I

QUANTITATIVE COMPARISON OF ARBITRACK WITH OTHER METHODS ON THE ORIDRONE TEST SET

Method	MOTA↑(%)	MOTP↑(%)	IDF1↑(%)	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓
MOTDT[39]	-1.2	66.3	20.3	85	1298	63080	266317	1562	3725
IOUT[53]	26.7	71.2	38.4	453	780	39251	197869	2597	3746
GOG[54]	27.2	**75.4**	37.0	381	865	28429	207584	2068	**2694**
CMOT[55]	28.4	72.4	42.4	497	824	43446	189547	1221	3456
SORT[17]	38.6	71.4	43.6	482	568	**18624**	180831	1344	5874
DSORT[21]	39.8	70.8	48.3	527	524	27120	168629	1126	6580
JDE[25]	38.3	72.1	53.4	541	**462**	26896	173257	1628	7598
FairMOT[29]	42.2	72.5	54.8	563	427	24875	162587	1564	5862
TrackFormer[41]	37.9	73.7	45.2	425	536	25698	172298	942	6235
ours	**44.6**	74.3	**56.9**	**593**	520	24366	**156025**	**786**	5632

TABLE II

QUANTITATIVE COMPARISON OF ARBITRACK WITH OTHER METHODS ON THE UAVDT TEST SET

Method	MOTA↑(%)	MOTP↑(%)	IDF1↑(%)	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓
CEM[51]	-6.8	70.4	10.1	94	1062	64373	298090	1530	**2835**
SMOT[52]	33.9	72.2	45.0	524	367	57112	166528	1752	9577
GOG[54]	35.7	72	0.3	627	374	62929	153336	3104	5130
IOUT[53]	36.6	72.1	23.7	534	357	42245	163881	9938	10463
CMOT[55]	36.9	**74.7**	57.5	**664**	351	69109	144760	1111	3656
SORT[17]	39.0	74.3	43.7	484	400	**33037**	172628	2350	5787
DSORT[21]	40.7	73.2	58.2	595	358	44868	155290	2061	6432
JDE[25]	39.5	73.5	55.3	624	442	-	-	3124	8536
FairMOT[29]	44.9	72.7	60.9	672	365	-	-	2279	7163
MDP[56]	43.0	73.5	61.5	647	324	46151	147735	541	4299
ours	**47.6**	73.4	**67.4**	657	**265**	60435	**117760**	**494**	5687

higher on IDF1. It also outperformed other trackers on IDs, showing that our method can effectively reduce ID switching and has better robustness in complex scenarios.

D. Ablation Study

A series of ablation experiments were conducted in the OriDrone test set to validate the effectiveness of each module in our approach. In the ablation experiments, Oriented RepPoints and Sort were used as baseline models, and Resnet50 was used as the backbone network.

TABLE III
ABLATION EXPERIMENTS ON THE ORIDRONE TEST SET

Baseline	KF	UKF	AMS	MOTA↑(%)	IDs↓	FPS↑
√	√			43.2	1126	12
√		√		44.8	563	5
√			√	44.6	786	10

As shown in Table III, AMS employs KF and UKF to estimate the motion state of the tracked target in different motion modes, respectively. We report three key metrics for each module in the test set. Using our AMS module on MOTA is 44.6%, IDs are 786, and FPS is 10. When the motion state estimator uses the KF module, MOTA decreases to 43.2%, IDs become 1340, and FPS improves to 12. When the motion state estimator uses the UKF module, MOTA improves to 44.8%, IDs become 563, and FPS decreases to 5. Although the FPS (frames per second) using the KF module as the motion estimation module is the highest at 12 FPS, indicating fast processing speed; however, it has the lowest MOTA (multi-target tracking accuracy) at 43.2% and the highest number of ID switches at 1,340. This indicates that the KF module, although fast, performs poorly in terms of tracking accuracy and stability. The UKF module, when used as a motion estimation module, has the best performance in MOTA and the least number of ID switches, but it has the lowest FPS of 5FPS, which is slow in processing speed. The AMS module ensures high tracking accuracy while taking into account the processing speed and tracking stability, thus providing a better balance between performance and speed. This makes it ideal for use in processing complex motion estimation, especially in UAV application scenarios where both real-time and accuracy need to be considered.

TABLE IV
ABLATION EXPERIMENTS ON THE ORIDRONE TEST SET

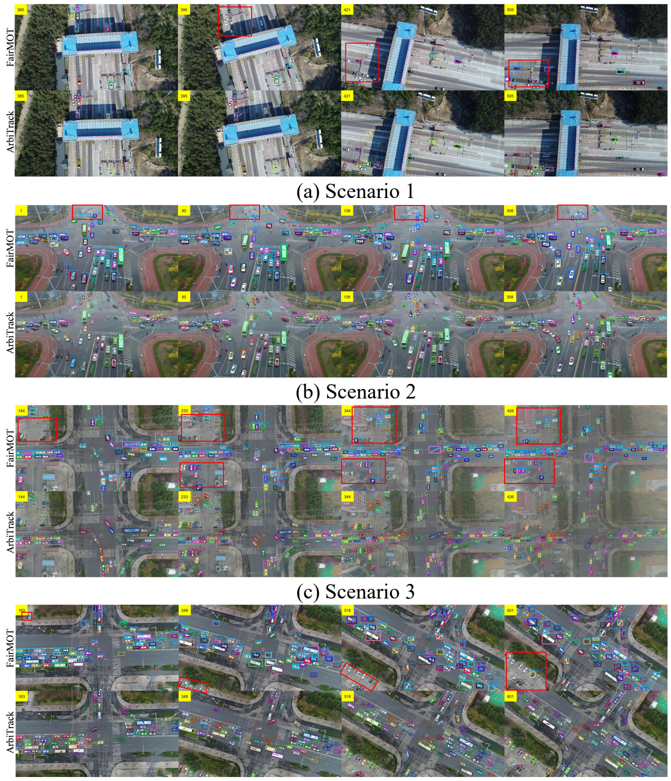
Baseline	MCA	AMS	SEM	MOTA↑(%)	IDF1↑(%)	IDs↓
√				38.4	44.6	2075
√	√			42.8	46.3	1340
√	√	√		43.2	53.2	985
√	√	√	√	44.6	56.9	593

As shown in Table IV, ArbiTrack contains three core components: the MCA module, the AMS module, and the SEM tracking algorithm. Three critical metrics are shown for each module in the test set. The baseline model is 38.4% on MOTA, 44.6% on IDF1, and 2075 on IDs. Adding the MCA module to the baseline model improves MOTA to 42.8%, decreases IDs to 1340, and reaches 46.3% on IDF1. Adding the MCA module and the AMS module to the baseline model, the MOTA improves to 43.2%, the IDF1 improves to 53.2%, and the IDs decrease from 1340 to 985. Adding all three modules, our MOT model reaches 44.6% on the MOTA and 56.9% on the IDF1.

The advantages of our ArbiTrack in terms of MOTA and IDF1 come from the three proposed modules. Specifically, the oriented object detector with the MCA module enables our tracker to accurately detect the object under complex UAV viewpoints, while the AMS module enhances the tracking effectiveness of ArbiTrack under complex UAV motions. The synergy of AMS and SEM improves the tracking stability of ArbiTrack under occlusion or complex UAV motions, which reduces the ID switching occurrences, and ensures the integrity of the output tracking trajectory.

E. Visualization

The tracking results are visualized, and the qualitative results were analyzed. We selected FairMOT [29], the next best tracking method in UAVDT and OriDrone. The yellow corner marker in the upper left corner of the image denotes the frame number of the tracking sequence where the image is located. To facilitate the observation of the object motion trajectory, the tracking trajectory is output while tracking the object in the resulting image of our method. Fig. 4 presents the tracking results of FairMOT and our method ArbiTrack for four different sets of scenarios in the UAVDT dataset.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

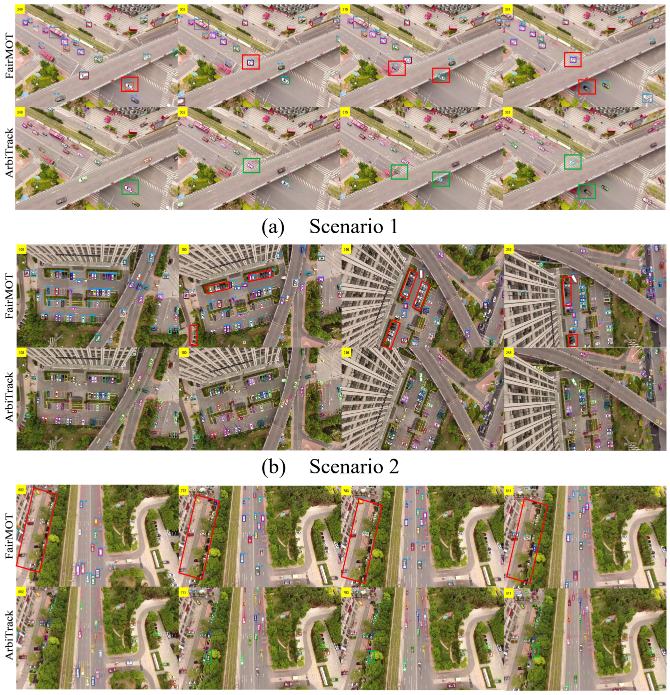
(d) Scenario 4

Fig. 4. Visualization results of UAVDT dataset. Red boxes indicate areas where many objects are lost, or ID switches occur.

In Fig. 4(a), the object position in its perspective rotates while the UAV suddenly rotates rapidly to the left during flight, and the original motion pattern will be disrupted. In this special scenario, FairMOT experienced object loss during the tracking process (as shown in the red box). Furthermore, ID switching occurred (as shown in the red box) after the perspective stabilized. However, ArbiTrack can accurately track objects without being affected by UAV rotations. In Fig. 4 (b), a dense vehicle intersection is monitored while the UAV is hovering in the air. In this case, vehicles that are farther away from the UAV appear to be objects with small size in the image captured from the UAV's viewpoint (marked by the red rectangular box at the top of the image), which cannot be tracked in FairMOT, but can be accurately tracked in ArbiTrack.

In Fig. 4(c), the object size in the UAV view becomes smaller when the UAV altitude keeps increasing. At the same time, there are more complex background information in the image. In this scenario, our ArbiTrack can accurately track the object against the complex background, while FairMOT cannot track all the objects due to the interference of the complex background. Fig. 4 (d) shows that when the UAV turns right during flight, the object motion trend under its perspective is a complex nonlinear motion that superimposes the UAV's motion and its own motion. At the same time, when turning right, the scene in the image undergoes rapid changes. In this scenario, our ArbiTrack can quickly adapt to scene changes and accurately track all moving objects in the image, while FairMOT can potentially lose objects.

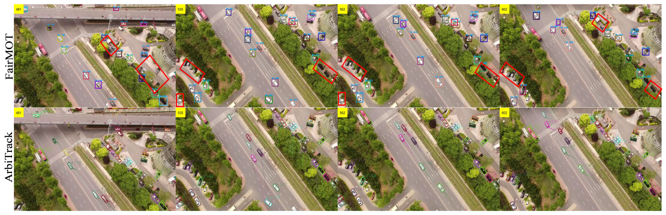
To further demonstrate the superiority of our ArbiTrack, four more challenging scenarios are selected for UAV tracking in comparison with the FairMOT method. The visualization results are shown in Fig. 5.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3



(d) Scenario 4

Fig. 5. Visualization results in the OriDrone dataset. The red boxes represent areas where many objects are missing, while the green boxes represent the objects that need attention.

- **Scenario 1 -- the object undergoes occlusion.** As shown in Fig. 5(a), three objects experienced occlusion (as shown in the green box) during the tracking. The vehicle with ID 58 undergone occlusion at frame 249 and was recaptured at frame 302 when it appeared once again (the occlusion time was around 2.5s). The two vehicles with IDs 56 and 63 undergone occlusions at frame 315 and were recaptured at frame 361 (the occlusion time was around 1.8s). For these occlusions at various times, our ArbiTrack has shown excellent tracking results benefited from the SEM module. However, FairMOT cannot effectively deal with the occlusion scene in this scenario in which three vehicles have switched their IDs after they are occluded.
- **Scenario 2 -- UAV rotates rapidly to the right.** Fig. 5 (b) shows the visualization results obtained in the scene when the UAV rotated rapidly to the right. It can be seen from the output trajectory points, our ArbiTrack can accurately track the object and ensure the integrity of the trajectory when the position and motion trend of the objects change drastically in response to the rapid rotation of the UAV. On the contrary, FairMOT lost many objects during rotation, resulting in many IDs switching.
- **Scenario 3 -- UAV performs acceleration motion.** Fig. 5 (c) shows the visualization results obtained in the UAV accelerated motion scene. The UAV is rapidly moving in the opposite direction along the road. In this scene, the scene changed rapidly from the UAV viewpoint while the objects moved at different speeds and movement law of the objects is complex. For example, the vehicle with ID 183 (shown in the green box) was captured in frame 775. Its motion direction and speed were different from most objects in the image and obscured by trees in frames 775 to 783. However, it was still accurately captured when it appeared in frame 783. Our method achieved good robustness in complex scenes, while FairMOT lost many objects during object tracking.
- **Scenario 4 -- The UAV camera rotates up and down.** Fig. 5 (d) shows the visualization results obtained in a scene where the camera rotated up and down rapidly when the UAV was hovering. In this scene, the object position in the UAV's viewpoint and the surrounding scene changed rapidly. It can be seen from the output trajectory points; our method tracked the whole trajectory for the objects under this scenario. In contrast, FairMOT lost many objects with a lot of IDs switching in this

scenario.

From the visualization results in Figure 4 and Figure 5, our ArbiTrack achieved good detection results for objects in complex backgrounds due to the proposed MCA module and the oriented object detector. Furthermore, based on the AMS and the SEM modules, our ArbiTrack can ensure the integrity and accuracy of the tracking trajectory in many scenarios such as occlusion, complex UAV motions. It significantly reduced the times of ID switching. Therefore, our ArbiTrack is valid in coping with tracking scenarios from the UAV viewpoint.

V. Conclusion

The primary motivation of this work is to design a multi-object tracker for a moving UAV. Our research is focused on the arbitrarily oriented object detection and tracking under UAV scenarios. A novel ArbiTrack framework was proposed for multi-object tracking of a moving UAV in this paper. An oriented object detector was used, and a multi-scale-context-aggregation module MCA was designed to aggregate local and global context information and enhance the feature representation of small-scale objects. To adapt to the complex UAV motion, an AMS module was proposed based on motion state feedback. According to the different motion modes of the UAV, two motion filters were adaptively selected and switched to provide accurate motion information for the tracker. Besides, a SEM tracking algorithm was presented to solve the problem of long-term complete occlusion during UAV tracking. In SEM, the historical position and motion state of the object were combined to learn the spatio-temporal characteristics of the objects. Its spatio-temporal evolution was modeled to guide the positioning and association of objects for the input video sequences. At last, a series of experiments were conducted using our OriDrone dataset and the public UAVDT dataset. The experimental results demonstrated that the proposed ArbiTrack framework can accurately track multiple objects by a moving UAV.

Our future work will further explore multiple-object tracking of UAV at night or in dark scenes by fusing visible and infrared images.

References

[1] Y. Jiao, X. Zhai, L. Peng et al., “A digital twin-based motion forecasting framework for preemptive risk monitoring,” *Adv. Eng. Informatics.*, vol. 59, pp. 1 02250, 2024, doi: 10.1016/j.aei.2023.102250.

[2] L. Wang et al., “CAMO-MOT: Combined appearance-motion optimzation for 3D Multi-Object Tracking with Camera-LiDAR Fusion,” *IEEE trans Intell Transp Syst*., vol. 24, no. 11, pp. 11981-11996, Nov. 2023, doi: 10.1109/TITS.2023.3285651.

[3] T. T. Nguyen, H. H. Nguyen, M. Sartipi and M. Fisichella, “Multi-vehicle multi-camera tracking with graph-based tracklet features,” *IEEE Trans Multimedia*., vol. 26, pp. 972-983, 2024, doi:10.1109/TMM.2023. 3274369.

[4] G. Brasó, L. Leal-Taixé. “Learning a neural solver for multiple object tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2020, pp. 6247-6257. doi: 10.1109/cvpr42600.202 0.00628.

[5] Y. Du, Z. Zhao, Y. Song et al., “Strongsort: Make deepsort great again,” *IEEE Trans Multimedia*., vol. 25, pp. 8725-8737, 2023, doi:10.1109/ TMM.2023. 3240881.

[6] X. Li, R. Zhu, X. Yu et al., “High-performance detection-based tracker for multiple-object tracking in UAVs,” *Drones*, vol. 7, no. 11, pp. 681, Nov. 2023, doi: 10.3390/drones7110681.

[7] P. Zhu, L. Wen, D. Du et al., “Detection and tracking meet drones challenge,” *IEEE Trans Pattern Anal Mach Intell*., vol. 44, no. 11, pp. 7380-7399, 1 Nov. 2022, doi: 10.1109/TPAMI.2021.3119563.

[8] L. A. Varga, B. Kiefer, M. Messmer et al., “Seadronessee: A maritime benchmark for detecting humans in open water,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis*., 2022, pp. 2260-2270, doi: 10.1109/wac v51458.2022.00374.

[9] Z. Liu et al., “Robust multi-drone multi-target tracking to resolve target occlusion: a benchmark,” *IEEE Trans Multimedia*., vol. 25, pp. 1462-1476, 2023, doi: 10.1109/TMM.2023.3234822.

[10]S. A. H. Mohsan, M. A. Khan, F. Noor et al., “Towards the unmanned aerial vehicles (UAVs): A comprehensive review,” *Drones*, vol. 6, no. 6, pp. 147, Nov. 2022, doi: 10.3390/drones6060147.

[11]A. Li, S. Hou, Q. Cai, Y. Fu and Y. Huang, "Gait recognition with drones: a benchmark," *IEEE Trans Multimedia*., doi: 10.1109/TMM.2023.33 12931.

[12]S. Liu, X. Li, H. Lu and Y. He, "Multi-Object Tracking Meets Moving UAV," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 8866-8875, doi: 10.1109/CVPR52688.2022.00867.

[13]W. Li, Y. Chen, K. Hu et al., “Oriented reppoints for aerial object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2022, pp. 1829-1838, doi: 10.1109/ cvpr52688.2022.00187

[14]Ballasn, L. Yao, C. Pal et al., “Delving deeper into convolutional networks for learning video representations,” 2015, arXiv:1511.06432.

[15]D. Du, Y. Qi, H. Yu et al., “The unmanned aerial vehicle benchmark: Object detection and tracking,” in *Proc. Eur. Conf. Comput. Vis*., 2018, pp. 370-386, doi: 10.1007/978-3-030-01249-6_23.

[16]G. Ciaparrone, F. L. Sánchez, S. Tabik et al., “Deep learning in video multi-object tracking: A survey,” *Neurocomputing*, vol. 381, pp. 61-88, Nov. 2020, doi.org/10.1016/j.neuc om.2019.11.023.

[17]A. Bewley, Z. Ge, L. Ott et al., “Simple online and realtime tracking,” in *Proc. IEEE Int. Conf. Image Process*., Phoenix, AZ, USA, 2016, pp. 3464-3468, doi: 10.1109/icip.2016.7533003.

[18]R. Girshick, “Fast R-CNN,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2015, doi: 10.1109/iccv.2015.169.

[19]R. E. Kalman, “A new approach to linear filtering and prediction problems,” *J. Basic Eng*., vol. 82, pp. 35-45, Nov. 1960, doi: 10.1115/1.3 662552.

[20]H. W. Kuhn, “The Hungarian method for the assignment problem,” *Nav. Res. Logist. Q*., vol. 2, no. 1-2, pp. 83-97, 1955, doi: 10.1002/nav.38000 20109.

[21]N. Wojke, A. Bewley, D. Paulus, “Simple online and realtime tracking with a deep association metric,” in *Proc. IEEE Int. Conf. Image Process*., 2017, pp. 3645-3649, doi: 10.1109/icip.2017.8296962.

[22]S. Lee, E. Kim, “Multiple-object tracking via feature pyramid siamese networks,” *IEEE Access*, vol. 7, pp. 8181-8194, Nov. 2019, doi:10.1109/ access.2018.2889442.

[23]T. Y. Lin, P. Dollár, R. Girshick et al., “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2017, pp. 2117-2125, doi: 10.1109/ cvpr.2017.106.

[24]S. Sun, N. Akhtar, H. S. Song et al., “Deep affinity network for multiple object tracking,” *IEEE Trans Pattern Anal Mach Intell*., vol. 43, no. 1, pp. 104-119, Nov. 2021, doi: 10.1109/TPAMI.2019.2929520.

[25]Z. Wang, L. Zheng, Y. Liu et al., “Towards real-time multi-object tracking,” in *Proc. Eur. Conf. Comput. Vis*., 2020, pp. 107-122, doi: 10.1007/978-3-030-58621-8_7.

[26]J. Redmon, A. Farhadi, “Yolov3: An incremental improvement,” 2018, arXiv: 1804.02767.

[27]A. Kendall, Y. Gal, R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2018, doi: 10.1109/cvpr.2018.00781.

[28]K. Duan, S. Bai, L. Xie et al., “Centernet: Keypoint triplets for object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2019, pp. 6569-6578, doi: 10.1109/iccv.2019.00667.

[29]Y. Zhang, C. Wang, X. Wang et al., “FairMOT: On the fairness of detection and re-identification in multiple object tracking,” *Int. J. Comput. Vis*., 2021: 3069-3087, http://dx.doi.org/10.1007/s11263-021-01513-4.

[30]X. Zhou, V. Koltun, P. Krähenbühl, “Tracking objects as points,” in *Proc. Eur. Conf. Comput. Vis*., 2020, pp. 474-490, doi: 10.1007/978-3-030-58548-8_28.

[31]D. Yuan et al., "Thermal Infrared Target Tracking: A Comprehensive Review," in IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1-19, 2024, Art no. 5000419, doi: 10.1109/TIM.2023.3338701.

[32]D. Yuan, X. Chang, Z. Li, and Z. He, "Learning Adaptive Spatial-Temporal Context-Aware Correlation Filters for UAV Tracking," ACM Trans. Multimedia Comput. Commun. Appl., vol. 18, no. 3, pp. 70:1–70:18, Aug. 2022, doi: 10.1145/3486678.

[33]L. Zhang, Y. Li, R. Nevatia, "Global data association for multi-object tracking using network flows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2008, pp. 1-8, doi: 10.1109/cvpr.2008. 4587584.

[34]Y. Zhang, P. Sun, Y. Jiang et al., "ByteTrack: Multi-object tracking by associating every detection box," in *Proc. Eur. Conf. Comput. Vis*., 2022, pp. 1-21, doi: 10.1007/978-3-031-20047-2_1.

[35]P. Sun, J. Cao, Y. Jiang et al., "Transtrack: Multiple object tracking with transformer," 2020, arXiv:2012.15460.

[36]J. Wu, J. Cao, L. Song et al., "Track to detect and segment: An online multi-object tracker," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2021, pp. 12 352-12361, doi: 10. 1109/cvpr46437.2021.01217.

[37]R. Li, B. Zhang, Z. Teng et al., "An end-to-end identity association network based on geometry refinement for multi-object tracking," *Pattern Recognition*, 2 022, pp. 129: 108738, doi: 10.1016/j.patcog.2022.108738.

[38]L. Zheng, H. Zhang, S. Sun et al., "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2017, pp. 1367-1376, doi: 10.1109/ cvpr.2017.357.

[39]L. Chen, H. Ai, Z. Zhuang et al., "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, doi:10.1109/ icme.2018.8486597.

[40]J. Pang, L. Qiu, X. Li et al., "Quasi-dense similarity learning for multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2021, pp. 164-173, doi: 10.1109/cvpr46437. 2021.00023.

[41]T. Meinhardt, A. Kirillov, L. Leal-Taixe et al., "TrackFormer: multi-object tracking with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2022, pp. 8844-8854, doi: 10.1109/cvpr52688.2022.00864.

[42]F. Zeng, B. Dong, Y. Zhang et al., "MOTR: End-to-end multiple-object tracking with transformer," in *Proc. Eur. Conf. Comput. Vis*., 2022, pp. 659-675, doi: 10.1 007/978-3-031-19812-0_38.

[43]N. Carion, F. Massa, G. Synnaeve et al., "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis*., 2020, pp. 213-229, doi: 10.1007/ 978-3-030-58452-8_13.

[44]S. Liu, L. Qi, H. Qin et al., "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2018, pp. 8759-8768, doi: 1 0.1109/cvpr.2018.00913.

[45]M. Tan, R. Pang, Q. V. Le. "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2020, pp. 10781-10790, doi: 10.1109/ cvpr42600.2020.01079.

[46]E. A. Wan, R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proc. IEEE Adaptive Syst*. Signal Process. Commun. Control Symp., 2000, pp. 153-158, doi: 10.1109/asspcc.2000.882463.

[47]K. Cho, B. Van Merriënboer, C. Gulcehre et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv: 1406.107 8, 2014, doi: 10.3115/v1/d14-1179.

[48]X. Shi, Z. Chen, H. Wang et al., "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Adv Neural Inf Process Syst*., vol. 28, 2015.

[49]P. Tokmakov, J. Li, W. Burgard et al., "Learning to track with object permanence," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2021, pp. 10860-10869, doi: 10.1109/ iccv48922.2021.01068.

[50]A. Milan, L. Leal-Taixé, I. Reid, et al., "MOT16: A benchmark for multi-object tracking," 2016, arXiv:1603.00831.

[51]A. Milan, S. Roth, K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Trans Pattern Anal Mach Intell*, vol. 36, no. 1, pp. 58-72, 2014, doi: 10.1109/tpami.2013.103.

[52]C. Dicle, O. I. Camps, M. Sznaier, "The way they move: Tracking multiple targets with similar appearance," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2013, pp. 2304-2311, doi: 10.1109/iccv.2013.286.

[53]E. Bochinski, V. Eiselein, T. Sikora, "High-speed tracking-by-detection without using image information," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill*., 2017, pp. 1-6, doi: 10.1109/avss.2017.8078516.

[54]L. Greedy, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2011, pp. 1201-1208, doi: 10.1109/cvpr.2011.5995604.

[55]S. H. Bae, K. J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2014, pp. 1218-1225, doi: 10.1109/cvpr.2014.15 9.

[56]Y. Xiang, A. Alahi, S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2015, pp. 4705-4713, doi: 10.1109/iccv.2015.534.

Yuqing Chen received the Ph.D. degree in control theory and control engineering from Dalian University of Technology, Dalian, China, in 2006, where he then joined the College of Marine Electrical Engineering, Dalian Maritime University, Dalian, China. From 2014 to 2015, he was a Visiting Research Fellow with the Robotic Lab, Department of Electrical Engineering, City University of New York, USA. He is currently a Professor with the Department of Automation, College of Marine Electrical Engineering, Dalian Maritime University. His current research interests include nonlinear control theory, intelligent control algorithm and applications in autonomous vessels, unmanned aerial vehicles, autonomous mobile robotics.

Jiayu Wang received the B.S. degree in Automation from Liaoning University of Technology, School of Electrical Engineering in 2021. He is currently pursuing the M.S. degree in Control Science and Engineering at Dalian Maritime University. His current research interests include deep learning, computer vision, multi-target tracking and their application in UAVs.

Qianchen Zhou received the B.S. degree in Electronic Information Engineering from the School of Computer and Information Engineering at Harbin Commercial University in 2022. He is currently pursuing the M.S. degree in Control Science and Engineering at Dalian Maritime University. His current research interests include deep learning, visual servoing, single object tracking, and their applications in robots.

Huosheng Hu (Life Senior Member, IEEE) received the M.S. degree in industrial automation from the Central South University, Changsha, China, in 1982, and the Ph.D. degree in robotics from the University of Oxford, Oxford, U.K., in 1993. He is currently a Professor with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, U.K. His research interests include intelligent systems and autonomous robots.