High Performance Detection-based Tracker for Multiple Object Tracking in UAVs

Xi Li , Ruixiang Zhu , Xianguo Yu * , Yirui Cong , Xiangke Wang

Posted Date: 26 October 2023

doi: 10.20944/preprints202310.1704.v1

Keywords: unmanned aerial vehicles; multi-object tracking; tracking-by-detection; set-membership filter

Article

High Performance Detection-Based Tracker for Multiple Object Tracking in UAVs

Xi Li [1], Ruixiang Zhu [1], Xianguo Yu [2,*], Yirui Cong [2] and Xiangke Wang [2]

[1] Hunan Provincial Key Laboratory of Flexible Electronic Materials Genome Engineering, Changsha University of Science and Technology, Changsha 410004, China; xili@csust.edu.cn (X.L.); ruixiangzhu@foxmail.com(R.Z.)

[2] College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China; congyirui11@nudt.edu.cn (Y.C.); xkwang@nudt.edu.cn (X.W.)

* Correspondence: yuxianguo11@nudt.edu.cn;

Abstract: Multi-Object Tracking (MOT) is a key technology for Unmanned Aerial Vehicles (UAVs). Traditional tracking-by-detection methods firstly employ an object detector to retrieve targets in each image and then track them based on a matching algorithm. Recently, the popular multi-task learning methods has been dominating this area since they can detect targets and extract Re-Identification (Re-ID) features in a computationally efficient way. However, the detection task and the tracking task have conflicting requirements on image features, leading to the poor performance of the joint learning model compared to separate detection and tracking methods. The problem is more severe when it comes to UAV images due to the presence of irregular motion of large number of small targets. In this paper, we propose a balanced Joint Detection and Re-ID learning (JDR) network to address the MOT problem in UAV vision. To better handle the non-uniform motion of objects in UAV videos, the Set-Member Filter is applied which describes object state as a bounded set. An appearance matching cascade is then proposed based on target state set. Furthermore, a Motion-Mutation module is designed to address the challenges posed by the abrupt motion of UAV. Extensive experiments on the VisDrone-MOT2019 dataset demonstrate that our proposed model, termed as SMFMOT, outperforms state-of-the-arts by a large margin and achieves superior performance on the MOT tasks in UAV videos.

Dataset: https://github.com/VisDrone/VisDrone-Dataset

Keywords: unmanned aerial vehicles; multi-object tracking; tracking-by-detection; set-membership filter

1. Introduction

Multi-Object Tracking (MOT) has been a long-term task of computer vision[1–5]. The goal of MOT is to estimate the trajectories of the objects of interest in videos. A successful solution to MOT can greatly help accomplish many applications such as autonomous driving, intelligent video analysis, human activity recognition[6], and even social computing. Recently, MOT in Unmanned Aerial Vehicles (UAVs) has attracted a lot of interest of researchers as the convenience and flexibility of UAV.

MOT methods [3,5,7] typically follow the tracking-by-detection paradigm. It firstly obtain potential boxes and appearance information for each object in each frame, then the matching algorithm based on motion cues [8] and appearance information [2] is used to associate objects across adjacent frames.

With the popularity of multi-task learning, the Joint Detection and Re-Identification (JDR) method [3,5,9,10] is dominant in this paradigm since it allows object detection and Re-Identification (Re-ID) feature extraction to be accomplished by a single network. It effectively reduces the computational cost of network. However, due to the conflicting requirements of detection tasks and Re-ID tasks on image features, the joint learning model performs poorly compared to separate detection and tracking methods. In the UAV scene, the existence of a large number of small objects and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1704.v1

the irregular movement coupled by movement of UAV and objects make the MOT model suffer from more severe challenges. In addition, the UAV undergoes abrupt motion during flight, which makes the matching blocked.

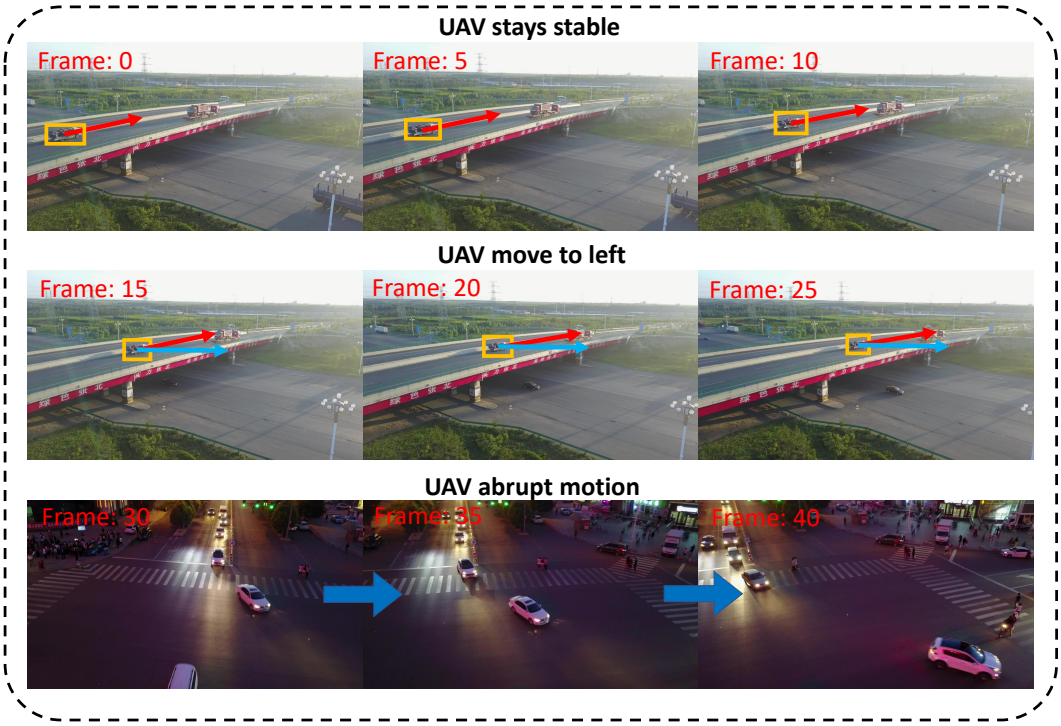

Figure 1. Irregular motion of objects in UAV video sequences. From frame 0 to frame 10, the UAV was stationary and the object was moving in the direction of the red arrow. From frame 15 to frame 25, the UAV was moving to the left while the object was moving in the direction of both the red and blue arrows in the image. From frame 30 to frame 40, the UAV was moving rapidly from left to right result in significant changes in the state of object.

In this paper, we propose a novel multi-object tracker, SMFMOT, to accomplish MOT tasks in UAV videos. SMFMOT introduces an anchor-based JDR network, which was built by a symmetric two-branch architecture to balance the detection task and Re-ID task, and attached to the same backbone that extracts multi-level features from the input image. Since the anchors improve the recall capability of network, the JDR model based on anchor-based detector can effectively deal with UAV video scenes with a large number of small objects. However, previous work has pointed out that anchor-based JDR network faces the problem that anchor and re-id feature cannot correspond one by one. Therefore, the network we design learns Re-ID features for every anchor.

In the matching algorithm, to better deal with the irregular coupled movement of objects in frame, we apply Set-Membership filter (SMF) [11] to predict and update the state of trajectory, where the state is described by the bounded set. Based on this characteristic of SMF, we designed an appearance matching cascade (AMC) module that improves the matching accuracy by selecting the matching objects in the prior range of trajectory. The actual state of the object may exist at any position within the prior range. Fortunately, we discovered that this module can effectively address matching errors caused by similar appearances of different targets, as strict screening of candidates based on state information was utilized. Additionally, to address the abrupt motion of UAV, we devised a motion-mutation filter (MMF) module that incorporates diverse matching strategies contingent on the UAV motion state. When the UAV is in an abnormal motion state, we adopted a state set match cascade method, which utilizes different techniques at multiple scales to cope with the varying scale of motion mutations in UAV.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1704.v1

We evaluated our model on the UAV dataset Visdrone-MOT2019 [12] via the evaluation server. The experimental results certify that our model can effectively complete the MOT task in UAV visions. The main motivation of this work derives from the particularity of object motion in unmanned aerial vehicle video sequences. The main contributions of this work are summarized as follows:

1. We proposed a JDR model that jointly learns an anchor-based detector and finds corresponding Re-ID features for multi-object tracking. Fairnesss between the two tasks is achieved by a symmetric design of the framework.
2. We employ the SMF for predict and update trajectory state, which describes the state of the trajectory with a bounded set to address the irregular motion of object. Based on the bounded set, the AMC is designed to accurately select candidates for appearance matching, thereby reducing false matches.
3. We proposed a MMF module to address the abrupt movement of the UAV, which determines the matching strategy based on the UAV motion state.

2. Related Work

The most effective methodologies for MOT typically rely on a tracking-by-detection approach [1–3,5,10,13], which involves detecting objects in each frame and associating them using a matching algorithm. As such, this section primarily focuses on recent advancements in multi-object detectors, Re-ID method, matching methods, and motion model in matching methods.

2.1. Detection Method

The popular object detectors can be classified into two categories: anchor-free and anchor-based. The anchor-based object detector generates a wide range of anchors using predefined rules that effectively cover a large proportion of the image. The predicted box is selected by matching the anchor to a positive sample. Faster R-CNN [14] generates anchors with varying sizes and proportions at each point of the feature graph, before filtering these anchors through the Region Proposal network (RPN). YOLOv1 [15] applies a clustering method to obtain anchors of different sizes in each grid of the divided image, before obtaining prediction boxes using a sample-sample matching strategy. As anchors improve the ability of the network to detect small objects with increased accuracy, anchor-based object detection is a powerful approach for identifying small objects in UAV videos.

Anchor-free object detectors directly perform object classification and localization on the image without anchors. CornerNet [16] utilizes the Keypoint-based method to detect objects by locating their key points and then generating bounding boxes based on those points. CornerNet-Lite [17] is proposed as an optimization, which includes the introduction of CornerNet-Saccade and CornerNet-Squeeze to improve the speed of the detection process.

2.2. Re-ID Method

With the development of multi-task learning, the JDR model has received increasing attention, as it completes both detection and Re-ID tasks within one network. Wang and Voigtlaender et al. [9,10] extract detection and Re-ID features from a single network for improving inference efficiency. MOTS [9] adds a Re-ID head in Mask R-CNN [18] and extracts the Re-ID feature while regressing to the bounding box. JDE [10] is deployed on YOLOv3 and inference with high efficiency. FairMOT [3] adds a Re-ID head to CenterNet [19], which achieved good performance. This method can effectively reduce the computational cost in the MOT task, as the detection subtask and the Re-ID subtask are completed in a single network.

2.3. Matching Method

Data association is a critical step in MOT, especially in the tracking-by-detection paradigm, which is achieved through matching methods. It associates objects across adjacent frames and assigns the

 doi:10.20944/preprints202310.1704.v1

same ID to the same object. In general, matching algorithms mainly rely on the appearance information and motion patterns of objects.

2.3.1. Matched based on Appearance Information

There is a large body of literature attempting to use the appearance information for tracking [2,3,5,9,10], which feed the object region into the Re-ID network to obtain appearance feature vectors. They then calculate the correlation between the trajectories and the detections. References [20–22] have focused on enhancing physical features for more reliable tracking. Bae et al. [21] propose an online appearance learning method to handle appearance variations. Tang et al. [22] utilize body postures to enhance physical features. However, in scenarios where a large number of similar objects are present in a single frame, relying solely on appearance information may become less reliable.

2.3.2. Matched based on Motion Sign

SORT is a classic algorithm in MOT. It first uses a Kalman filter [23] to predict the state of trajectories, then computes the Intersection over Union (IoU) with detections. The Hungarian algorithm is used for assigning detections to trajectories. IOU-Tracker [24] directly calculates the IoU between the tracklet and detection without predicting the prior state of the trajectory by the motion model. Both SORT and IOU-Tracker are widely used because of the simplicity. However, these two methods do not perform well in crowded objects or fast-moving objects. Therefore, matching methods based solely on location and movement information cannot address extreme scenarios.

2.3.3. Matched based on Appearance Information and Motion Sign

Many studies [2,3,5,10,25] have attempted to enhance the performance of matching methods in multi-object tracking by incorporating both motion models and appearance information. The DeepSORT improves upon the SORT by implementing a cascade matching process that takes advantage of appearance information. The MOTDT [25] proposes a multi-level association strategy that utilizes the IoU method to match detection and tracklets when appearance features are not reliable. Additionally, the UAVMOT [5] adopts an adaptive motion filter to improve the multi-object tracking performance under the UAV background. These methods utilize Kalman filters to predict the prior state of the trajectory, which is considered to conform to a Gaussian distribution. However, as the motion of the object in the image is often coupled with the motion of the UAV, the use of a Gaussian distribution with a mean of 0 to describe the error in state representation is unreliable.

2.4. Motion Model

The Kalman filter [23] is a widely-used algorithm for state estimation and prediction in various scientific and engineering applications. In the context of multi-object tracking, the Kalman filter is often employed to predict the positions, sizes, and velocities of trajectories, and match them with objects. In existing tracking algorithms, such as SORT and DeepSORT, the state of the trajectory is assumed to follow a Gaussian distribution of which the mean is considered as the state of the trajectory. However, the trajectory should lie within a bounded range rather than a single value.

SMF [11] is a class of nonlinear filtering algorithms that process measurements and predict the state of a system by enclosing it within a prior range, which is represented in this paper with a zonotope[26]. SMF can perform state estimation under both Gaussian and non-Gaussian measurement noise, making them suitable for a wide range of applications. In this article, the SMF is utilized to predict and update possible trajectory states, and the detection can be matched with the trajectory by determining whether it falls within the predicted set.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1704.v1

3. Methodology

3.1. Overall Framework

In this paper, our proposed model consists of two main parts. The first part is the balanced anchor-based JDR network, which extracts appearance feature vectors of detected objects when detecting. The second part is the matching method, which was improved by DeepSORT. It includes a SMF module, an MC module, and a MMF module. Given a sequence $\{I_t \in \mathbb{R}^{W\times H\times 3}\}_{t=1}^{T}$ of UAV videos in moving, our proposed model aims to associate objects of the same category in adjacent sequences with the aid of their position and appearance information, assigning them a unique ID. The overall framework of our model is illustrated in Figure 2. We input the images I_t and I_{t+1} of adjacent frames into the JDR network. The detection head outputs the categories $\{C\}_{i=1}^{N}$ and boxes $\{B\}_{i=1}^{N}$ of N objects, while the Re-ID head outputs the appearance vectors $\{ID\}_{i=1}^{N}$ corresponding to objects. We predict the next state of the objects using the SMF module, which describes the state of the object with a set in UAV videos.

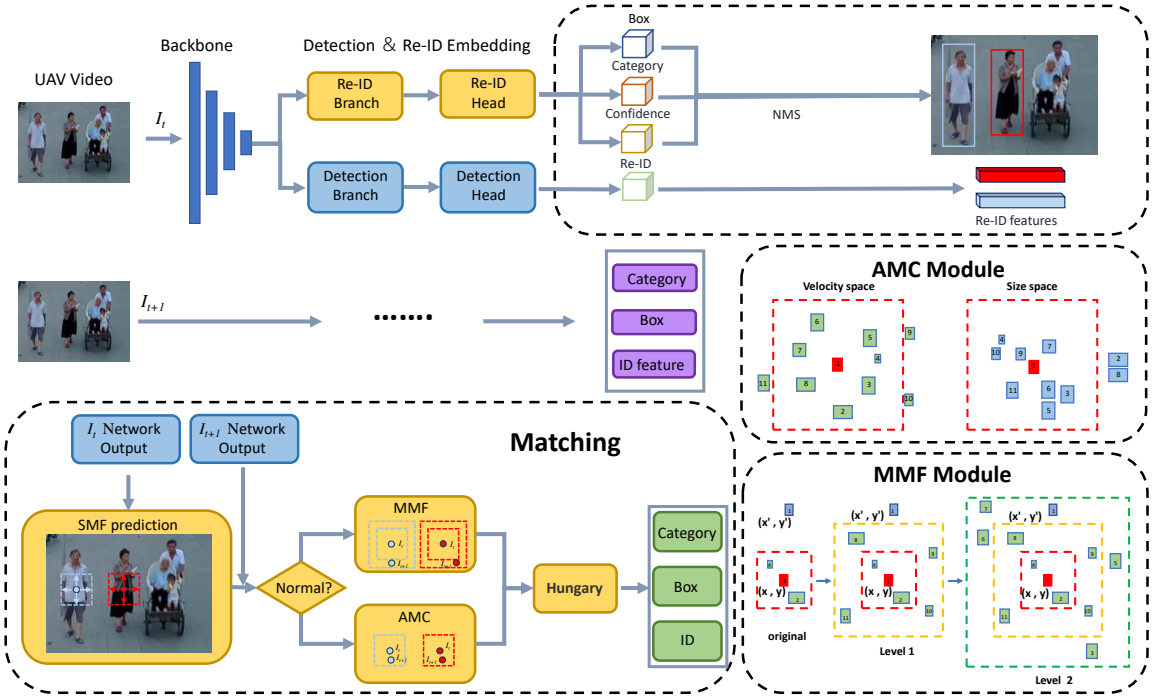

Figure 2. Overview of the proposed model. We utilize the JDR network to obtain the category, box, and appearance feature vector of objects from the adjacent frames of input images I_t and I_{t+1}. In the tracking stage, we designed the SMF module to predict trajectory states, the AMC module to select eligible candidates for appearance matching, and the MMF module to tackle the challenge of rapid motion of UAV. In AMC module and MMF module, rectangular boxes with numerical values representing the size of the object or trajectory, where the red one is the trajectory and the others are detection.

3.2. JDR Network

We designed an anchor-based JDR model to accomplish both detection and Re-ID tasks which consist of a backbone, a detection branch, and a Re-ID branch. The overview of network is shown in the top of Figure 2. It adopts the backbone of the YOLOv7 detector which is released in YOLOv7, YOLOv7-tiny, and YOLOv7-W6. To strike a good balance between accuracy and speed, we employe the version of YOLOv7. With the addition of E-ELAN, an enhanced version of ELAN, the backbone can improve the learning ability of the network through expanding, shuffling, and merging cardinality

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1704.v1

without destroying the original gradient path. This is conducive to ensuring the quality of Re-ID feature vectors while executing the detection tasks. The multiple uses of E-ELAN allow the network to learn a variety of features at different levels.

The detection branch consists of a detection neck and three output heads corresponding to three levels of feature maps output from backbone. The neck is used for feature transformation and feature fusion. Each output head contains a box head, a category head and a confidence head. The box head aim to determine the location and size of the object. The sizes of objects corresponding to the three segregated levels are different. The higher the level is, the richer the semantics and the larger the corresponding detection targets are. The category head outputs the belonging category for each object. The confidence head represents the probability that the anchor serves as a positive sample, which determines whether the anchor contains the correct target.

In order to ensure the performance balance between detection tasks and Re-ID tasks, the Re-ID branch is designed as similar to the detection branch. Feature graphs output from different layers in the backbone correspond to three different scale of targets. We extracted the appearance feature for each anchor in the same level. The Re-ID features are learned through a classification task. All objects with the same id are treated as the same category. For each predicted target, we extract its feature vector E^i and map it to a distribution vector $P = \{p(k), k \in [1, K]\}$ through a fully connected layer and softmax operation. We adapted the IDs of targets into one-hot form and denoted it as $L^i(k)$. We use a focal loss to calculate the Re-ID loss, as shown in the following.

$$L_{id} = -\sum_{i=1}^{N}\sum_{k=1}^{K} L^i(k) \log p(k) \tag{1}$$

Where K is the total number of identities in a single sequence and N is the total number of sequences.

3.3. Matching Model

3.3.1. Motion Module

In our model, the SMF[11] is used as a motion model to predict and update the state of the trajectory. For each trajectory, our tracking scene is defined in an eight-dimensional space $(u, v, \gamma, h, \dot{x}, \dot{y}, \dot{\gamma}, \dot{h})$, which includes the central point position (u, v) of the object, the aspect ratio r, the length h of the object box and their corresponding velocities $(\dot{x}, \dot{y}, \dot{\gamma}, \dot{h})$. Set $\mathbf{x}$ represents the range of each trajectory in the eight-dimensional space, which shown in the following.

$$\exists(\mathbf{G}, \mathbf{c}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n : \mathbf{x} = \{\mathbf{G}\xi + \mathbf{c} : \|\xi\|_\infty \leq 1\} \tag{2}$$

Where the vector $\mathbf{c}$ is the center of set, the columns of $\mathbf{G}$ are the generators of set.

The motion of the trajectory is modeled as a linear system, in which the noise is independent of the initial state. It can be described by

$$\mathbf{x}_{k+1} = A\mathbf{x}_k + B\mathbf{w}_k \tag{3}$$

$$\mathbf{y}_k = C\mathbf{x}_k + D\mathbf{v}_k \tag{4}$$

Where state transition matrix $A \in \mathbb{R}^{n\times n}$, control matrix $B \in \mathbb{R}^{n\times p}$, measurement matrix $C \in \mathbb{R}^{m\times n}$, and control matrix $D \in \mathbb{R}^{m\times q}$. Each trajectory in the system is required to follow the following steps:

- **Initialization.** Set the initial prior range $[\![\mathrm{x}_0]\!]$.
- **Prediction.** For $k \in \mathbb{Z}_+$, the prior range is

$$[\![\mathbf{x}_k \,|y_{0:k-1}]\!] = A[\![\mathbf{x}_{k-1}|\, y_{0:k-1}]\!] \oplus B[\![\mathbf{w}_{k-1}]\!], \tag{5}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1704.v1

where $\oplus$ stands for the Minkowski sum. The prior ranges will be used for matching with the objects.

- **Update.** For $k \in \mathbb{N}_0$, given y_k, the posterior range is

$$[\![\mathbf{x}_k \,|\, y_{0:k}]\!] = \mathcal{X}_k\left(C, y_k, D[\![\mathbf{v}_k]\!]\right) \bigcap [\![\mathbf{x}_k \,|\, y_{0:k-1}]\!], \tag{6}$$

here we define $[\![\mathrm{x}_0]\!] := [\![\mathrm{x}_0 \mid y_{0:-1}]\!]$ for consistency, and $\mathcal{X}_k\left(C, y_k, D[\![\mathbf{v}_k]\!]\right) = \{x_k : y_k = Cx_k + Dv_k, v_k \in [\![\mathbf{v}_k]\!]\}$. The update stage is performed after matching trajectories with objects, the prior range $[\![\mathbf{x}_k \mid y_{0:k-1}]\!]$ is assumed to be infinitely large.

3.3.2. Appearance Matching Cascade Module

Since SMF describes the state of the trajectory as a set, we designed the AMC module based on this peculiarity to select candidates for appearance matching in a precise manner. Only the detections contained in the predicted state set are considered for calculating the appearance cost matrix and performing the matching. Compared to the DeepSORT, which uses a threshold based on the distance between the center points of the trajectories and detection to determine the appearance matching candidates, our method more accurately reflects the true state of the trajectory to reduce mismatching. As shown in Figure 3, DeepSORT selects candidates for appearance matching by judging that their location is within the circular region. It only takes into account that the location has boundaries, and ignores that there are also boundaries in size and speed.

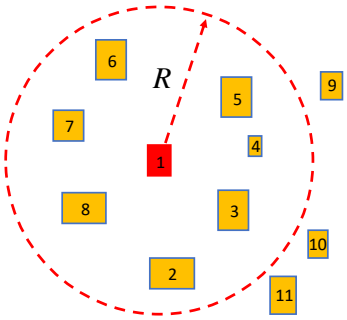

Figure 3. Overview of selecting candidates for appearance matching in DeepSORT. It use positional threshold to select objects as the criterion for appearance matching. The candidates selected as appearance matches only need to have their center positions exist within a circular range of radius R around trajectory 1.

In the AMC module, we only select objects whose position, size and corresponding velocity are contained within the trajectory prior range as candidates for appearance matching. As shown in Figure 2, there are 7 appearance matching candidates in the position space of trajectory 1 which is represented by a red box. However, some of these objects have aspect ratios that are not within the state space of trajectory 1, and therefore these objects should be excluded. Similarly, trajectories that are not within the other four velocity component spaces will also be excluded. Moreover, this module can alleviate the problem of misidentification between similar objects as they frequently have diverse states. The overall AMC module can be summarized as Algorithm 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1704.v1

Algorithm 1 AMC algorithm

Input: Track State $\mathcal{T} = \{track_1, \ldots, track_N\}_t^T$, $track_i$ is represented by Eq. 4, Detection $\mathcal{D} = \{det_1, \ldots, det_M\}_t^T$, $det_j = (u_j, v_j, \gamma_j, h_j, \dot{x}_j, \dot{y}_j, \dot{\gamma}_j, \dot{h}_j)$
Output: Matched trajectories and detections $\mathcal{M} = \{(track_n, det_m)|n \in N, m \in M\}$
Initialize candidate matrix $C \leftarrow \emptyset$
for $track \in \mathcal{T}$ **do**
  Initialize flag matrix $F \leftarrow \emptyset$
  for $det \in \mathcal{D}$ **do**
    if $det \in track$ **then**
      $F \leftarrow True$
    else
      $F \leftarrow False$
  end for
  $C \leftarrow \{f \in F | f = True\}$
end for
Calculate the Euclidean distance E of appearance vectors between trajectories and candidates
Matching using Hungarian algorithm with E, get $\mathcal{M}$

3.3.3. Motion-mutation Filter Module

In UAV video sequences, as the possibility of sudden changes in UAV motion, the position of objects in the adjacent frames can experience significant changes. The use of solely a SMF module is insufficient to deal with such extreme motion as the state of the objects has changed dramatically accordingly. Therefore, we propose a MMF module to address the challenges posed by abrupt motion in UAV.

Based on the motion of the UAV, objects in UAV videos can be classified into two modes: normal mode and anomaly mode. In normal mode, the drone flies smoothly in the sky, the coupling motion between the object in the image and the UAV remains in a stable state. We calculate the matching quantity by AMC module. When the matching quantity falls below a certain threshold P, it is considered to be entering abnormal mode, where the drone performs fast movements with different amplitudes, resulting in significant variations in the position and speed of the object in the video sequence. Luckily, we found that the characteristic of describing states as sets with SMF allowed us to address this challenge by dynamically challenging the range of sets. Therefore, we adopts a state set match cascade to enlarge the state space at each level to cope with the rapid movements of drones at different scales. For the predicted trajectory state $\widehat{\mathbf{x}}$ through SMF, there exists a ΔG_k at level k such that the trajectory state set as shown in follow.

$$\widehat{\mathbf{x}} = \{(\mathbf{G} + \mathbf{\Delta G})\xi + \mathbf{c} : \|\xi\|_\infty \leq 1\}. \tag{7}$$

At each level, a matching process is performed between the trajectory and object by the AMC module, and the matching result of the level with the best performance is selected as the matching result for the current frame. In Figure 2, we demonstrate an example of MMF on the position space of objects. The red box 1 represents the trajectory to be matched, where the (x, y) is the position of the trajectory in the image. The blue box 1 represents the ground truth that trajectory 1 should match where location is repesented by (x', y'). Since the scale variance of the UAV motion, the objects that should truly match the trajectory are at the level 2 of position, requiring a two-level state space expansion for correct matching. In the same way as the position space, changes in the size and velocity space also occur. The reason for changes in the velocity space is that the fast movement of the drone coupled with the object motion, while changes in the size space are due to the slight variation in object size caused by the movement of UAV and angle of capture. However, compared to the changes in position and velocity, the variation in size space is minimal. Therefore, the position space and its corresponding velocity space are the primary focuses of our attention.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1704.v1

4. Experiments

4.1. Datasets and Metrics

The data set used in this work is Visdrone-MOT2019, which has training test and verification sets. It provides both box and identity annotations. Therefore, we can train both branches in this dataset. The VisDrone-MOT2019 dataset contains ten categories: pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle. During the multi-object tracking evaluation, we only consider five object categories, i.e., pedestrian, car, van, bus, and truck. Following UAVMOT, both the training set and the validation set of the Visdrone-MOT2019 were used to train the model, and evaluated our approach on the testing set. We evaluate detection results by Average Precision (AP). After obtaining the detection results, we will extract corresponding Re-ID features to facilitate the use of the tracking stage. To evaluate our model with other state-of-the-art approaches, we adopt a variety of metrics to assess the performance of tracking, such as multiple object tracking accuracy (MOTA), multiple object tracking precision (MOTP), Identification F-Score (IDF1), ID switching (IDs) and other metrics.

4.2. Implementation Details

We adopt our published JDR model based on anchor-based detector YOLOv7 as the detection model and extract appearance feature vectors. The model parameters pre-trained on the COCO dataset [27] are used to initialize the model. We train the model with the SGD [28] optimizer for 30 epochs with a starting learning rate of 10^{-2}. The learning rate decays to 10^{-3} at 15 epochs. The batch size is set to 6. We used standard data enhancement techniques such as left-right inversion, zooming, and color transformation. During the training, the input image is resized to 1280×1280. The training took about 20 hours on two RTX 3090 GPUs.

4.3. Comparison with State-of-the-arts

We compared our approach to the previous methods on the VisDrone-MOT2019 dataset. We trained our model on the training set and verification set. We tested it on the test set using the official VisDrone MOT toolkit. As illustrated in Table 1, our method achieves 43.6% on MOTA and 54.9% on IDF1. The MOTP, MT, ML and FN also achieved the best performance. These experimental data shown in Table 1 are quoted from ref. [5], where the input image is resized to 1920×1080 during the training. Our approach achieves better performance with lower-resolution input images.

Table 1. Quantitative comparisons between our method and other methods for MOT task on VisDrone-MOT2019 test-dev set.

Dataset	Method	MOTA↑(%)	MOTP↑(%)	IDF1↑(%)	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓
Visdrone	MOTDT	-0.8	68.5	21.6	87	1196	44548	185453	1437	3609
	SORT	14.0	73.2	38.0	506	545	80845	112954	3629	4838
	IOUT	28.1	74.7	38.9	467	670	36158	126549	2393	3829
	GOG	28.7	76.1	36.4	346	836	**17706**	144657	1387	**2237**
	MOTR	22.8	72.8	41.4	272	825	28407	147937	**959**	3980
	TrackFormer	25	73.9	30.5	385	770	25856	141526	4840	4855
	UAVMOT	36.1	74.2	51.0	520	574	27983	115925	2775	7396
	Ours	**43.6**	**76.4**	**54.9**	**656**	**469**	32599	**86654**	2113	4042

4.4. Ablation Study

In this section, we performed a series of ablation experiments on VisDrone-MOT2019 validation to validate our model. In ablation experiments, we used our proposed JDR model and DeepSORT as the baseline model. As shown in Table 2, there are three core components in our model, the SMF module, the AMC module and the MMF module. We report the results of four critical performance

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1704.v1

metrics for each module on the Visdrone-MOT2019 test set. The baseline gets 42.3% on MOTA ,75.3% on MOTP and 53.9% on IDF1. When the motion model was replaced by the SMF filter, the MOTA increased to 42.5%, the MOTP increased to 76.6% and IDF1 get 49.7%. When the motion model is replaced by the SMF filter and the AMC module is applied to the baseline model, the MOTA increased to 43.3%, the MOTP achieves 76.4% and IDF1 achieves 53.9%. Add the SMF module, AMC module and MMF module to the baseline model, the MOTA increased to 43.6%, the MOTP achieves 76.4% and IDF1 achieves 54.9%. Due to the interdependency among the modules we designed, the ablation experiments of each module below may be dependent on one or two modules.

Table 2. Ablation study on VisDrone-MOT2019 test set.

Baseline	SMF	AMC	MMF	MOTA↑(%)	MOTP↑(%)	IDF1↑(%)	IDs↓	IDP↑(%)	IDR↑(%)
✓				42.3	75.3	53.9	2314	61.7	47.9
✓	✓			42.5	**76.6**	49.7	3200	60.0	42.4
✓	✓	✓		43.3	76.4	53.9	2155	63.1	47.0
✓	✓	✓	✓	**43.6**	76.4	**54.9**	**2113**	**64.1**	**48.0**

4.4.1. Effectiveness of SMF Module

The SMF module can effectively predict the state of the trajectory and improve tracking accuracy. To evaluate the effectiveness of the SMF module, we list four critical indicators(Pren, FP) on the baseline model and the baseline +SMF model, respectively. As shown in Table 3, Prcn increased from 77.9% to 81.2% and FP from 36868 decrease to 28614. The experimental results demonstrate that SMF has a positive impact on the matching stage, enabling accurate prediction of the object motion state, reducing the number of tracking errors, and improving tracking accuracy.

Table 3. Ablation study on VisDrone-MOT2019 test set.

	Prcn↑(%)	FP↓(%)
Baseline	77.9	36868
Baseline+SMF	**81.2**	**28614**

4.4.2. Effectiveness of AMC Module

The AMC module is used to filter out objects that do not meet the trajectory state and prevent mismatches between similar objects. To evaluate the effectiveness of AMC, we list ID association indicators (IDF1, IDS, IDP, IDR) on the baseline+SMF model and baseline+SMF+AMC model, respectively. As illustrated in Table 4, The IDs from 3200 decreases to 2155. The IDF1, IDP and IDR increase from 49.7%, 60.0% and 42.4% to 53.9%, 63.1% and 47.0%, respectively. The experimental results demonstrate that AMC has a significant impact on the matching stage, and its primary contribution stems from a considerable reduction in ID switches.

Table 4. Ablation study on VisDrone-MOT2019 test set.

	IDF1↑(%)	IDs↓	IDP↑(%)	IDR↑(%)
Baseline+SMF	49.7	3200	60.0	42.4
Baseline+SMF+AMC	**53.9**	**2155**	**63.1**	**47.0**

4.4.3. Effectiveness of MMF Module

The MMF module can automatically switch tracking modes based on the UAV motion state. To evaluate the effectiveness of MMF, we list ID association indicators (IDF1, IDs, IDP, IDR) on the baseline+SMF+AMC model and baseline+SMF+AMC+MMF model, respectively. As illustrated in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1704.v1

Table 5, The IDs from 2155 decrease to 2113. The IDF1, IDP and IDR increase from 53.9%, 63.1% and 47.0% to 54.9%, 64.1% and 48.0%, respectively. The experimental results demonstrate that MMF has a significant impact on the matching stage, it alleviated the influence of ID switching caused by mutation motion of UAV.

Table 5. Ablation study on VisDrone-MOT2019 test set.

	IDF1↑(%)	IDs↓	IDP↑(%)	IDR↑(%)
Baseline+SMF+AMC	53.9	2155	63.1	47.0
Baseline+SMF+AMC+MMF	**54.9**	**2113**	**64.1**	**48.0**

4.5. Visualization

In this section, we will showcase the tracking results of our model on UAV video sequences from the Visdone-MOT2019 dataset more intuitively. As depicted in Figure 4, the UAV was in normal motion from farm10 to farm20, while it underwent rapid motion from farm20 to farm30. Our model can effectively track multiple objects in UAV environments with normal and fast movements. The visualization results demonstrate that our model can accomplish most multi-object tracking tasks in the UAV domain.

Figure 4. Visualization of tracking results on Visdrone-MOT2019.

5. Conclusions

In conclusion, a novel MOT model is proposed to MOT task in UAV, which follows the track-by-detection paradigm. In the model, we utilized a symmetric anchor based JDR model to accomplish both detection and Re-Identification (Re-ID) tasks. It can achieve high-performance detection results and high-quality Re-ID features, while balancing the detection task and Re-ID task effectively. We utilized the SMF to predict the set of trajectory states, which describe the range of trajectory existence accurately. The AMC module is proposed to optimize the appearance-based matching strategy using object status. The module eliminates objects that are not in the trajectory state set and reduces the effects of appearance similarity. We proposed an MMF module to address the sudden movement issue of UAV, which determines matching strategies based on the UAV motion state. We conducted a series of experiments on the VisDrone-MOT2019 dataset and compared our approach with other methods. The results show that our method achieves state-of-the-art performance in UAV multi-object tracking tasks.

Author Contributions: Conceptualization, X.Y., X.W. and Y.C.; methodology, X.L., R,Z., X.Y. and Y.C.; software, X.Y. and R.Z.; validation, X.Y. and Y.C.; formal analysis, X.L. X.Y. and Y.C.; investigation, X.Y. and R.Z.; resources, X.Y. and X.W.; data curation, X.Y. and R.Z.; writing—original draft preparation, R.Z.; writing—review and editing, X.L., X.Y. and R.Z; visualization, R.Z.; supervision, X.Y., Y.C. and X.W.; project administration, X.Y. and X.W; funding acquisition, X.L. and Y.C.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China grant number 52001029. This work was supported by the National Natural Science Foundation of China grant number 61973309 and the Natural Science Foundation of Hunan Province grant number 2021JJ20054.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to thank the Editor-in-Chief, Editor, and anonymous reviewers for their valuable reviews.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of this study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

MOT	Multi-Object Tracking
UAVs	Unmanned Aerial Vehicles
Re-ID	Re-Identification
JDR	Joint Detection and Re-ID learning
SMF	Set-Membership filter
AMC	Appearance matching cascade
MMF	Motion-mutation filter
RPN	Region Proposal network
IoU	Intersection over Union
AP	Average Precision
MOTA	Multiple object tracking accuracy
MOTP	Multiple object tracking precision
IDF1	Identification F-Score
IDs	ID switching

References

1. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 3464–3468. https://doi.org/10.1109/ICIP.2016.7533003.
2. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962.
3. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision* **2021**, *129*, 3069–3087.
4. Yu, F.; Li, W.; Li, Q.; Liu, Y.; Shi, X.; Yan, J. POI: Multiple Object Tracking with High Performance Detection and Appearance Feature. In Proceedings of the Computer Vision – ECCV 2016 Workshops; Hua, G.; Jégou, H., Eds., Cham, 2016; pp. 36–42.
5. Liu, S.; Li, X.; Lu, H.; He, Y. Multi-Object Tracking Meets Moving UAV. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 8876–8885.
6. Wang, C.; Wang, Y.; Yuille, A.L. An Approach to Pose-Based Action Recognition. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2013.
7. Braso, G.; Leal-Taixe, L. Learning a Neural Solver for Multiple Object Tracking. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.
8. Huang, C.; Wu, B.; Nevatia, R. Robust Object Tracking by Hierarchical Association of Detection Responses. In Proceedings of the Computer Vision – ECCV 2008; Forsyth, D.; Torr, P.; Zisserman, A., Eds., Berlin, Heidelberg, 2008; pp. 788–801.
9. Voigtlaender, P.; Krause, M.; Osep, A.; Luiten, J.; Sekar, B.B.G.; Geiger, A.; Leibe, B. MOTS: Multi-Object Tracking and Segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

10. Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; Wang, S. Towards Real-Time Multi-Object Tracking. In Proceedings of the Computer Vision – ECCV 2020; Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J.M., Eds., Cham, 2020; pp. 107–122.
11. Cong, Y.; Wang, X.; Zhou, X. Rethinking the Mathematical Framework and Optimality of Set-Membership Filtering. *IEEE Transactions on Automatic Control* **2022**, *67*, 2544–2551. https://doi.org/10.1109/TAC.2021.3082508.
12. Wen, L.; Zhu, P.; Du, D.; Bian, X.; Ling, H.; Hu, Q.; Zheng, J.; Peng, T.; Wang, X.; Zhang, Y.; et al. VisDrone-MOT2019: The Vision Meets Drone Multiple Object Tracking Challenge Results. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Oct 2019.
13. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. ByteTrack: Multi-object Tracking by Associating Every Detection Box. In Proceedings of the Computer Vision – ECCV 2022; Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G.M.; Hassner, T., Eds., Cham, 2022; pp. 1–21.
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **2015**, *28*.
15. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
16. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), September 2018.
17. Law, H.; Teng, Y.; Russakovsky, O.; Deng, J. CornerNet-Lite: Efficient Keypoint Based Object Detection. *CoRR* **2019**, *abs/1904.08900*, [1904.08900].
18. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.
19. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.
20. Chen, L.; Ai, H.; Zhuang, Z.; Shang, C. Real-Time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), 2018, pp. 1–6. https://doi.org/10.1109/ICME.2018.8486597.
21. Sadeghian, A.; Alahi, A.; Savarese, S. Tracking the Untrackable: Learning to Track Multiple Cues With Long-Term Dependencies. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.
22. Tang, S.; Andriluka, M.; Andres, B.; Schiele, B. Multiple People Tracking by Lifted Multicut and Person Re-Identification. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.
23. Basar, T., A New Approach to Linear Filtering and Prediction Problems. In *Control Theory: Twenty-Five Seminal Papers (1932-1981)*; 2001; pp. 167–179. https://doi.org/10.1109/9780470544334.ch9.
24. Bochinski, E.; Eiselein, V.; Sikora, T. High-Speed tracking-by-detection without using image information. In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1–6. https://doi.org/10.1109/AVSS.2017.8078516.
25. Chen, L.; Ai, H.; Zhuang, Z.; Shang, C. Real-Time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), 2018, pp. 1–6. https://doi.org/10.1109/ICME.2018.8486597.
26. Scott, J.K.; Raimondo, D.M.; Marseglia, G.R.; Braatz, R.D. Constrained zonotopes: A new tool for set-based estimation and fault detection. *Automatica* **2016**, *69*, 126–136. https://doi.org/https://doi.org/10.1016/j.automatica.2016.02.036.
27. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision – ECCV 2014; Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T., Eds., Cham, 2014; pp. 740–755.
28. Robbins, H.; Monro, S. A stochastic approximation method. *The annals of mathematical statistics* **1951**, pp. 400–407.

 doi:10.20944/preprints202310.1704.v1