Multi-Object Tracking Meets Moving UAV

Shuai Liu[†1], Xin Li[†2], Huchuan Lu[1,2], You He[*3]
[1]Dalian University of Technology, [2]Peng Cheng Laboratory, [3]Naval Aeronautical University
[1]Dalian, [2]Shenzhen, [3]Yantai, China
lshuai@mail.dlut.edu.cn, xinlihitsc@gmail.cn, lhchuan@dlut.edu.cn, youhe_nau@163.com

Abstract

Multi-object tracking in unmanned aerial vehicle (UAV) videos is an important vision task and can be applied in a wide range of applications. However, conventional multi-object trackers do not work well on UAV videos due to the challenging factors of irregular motion caused by moving camera and view change in 3D directions. In this paper, we propose a UAVMOT network specially for multi-object tracking in UAV views. The UAVMOT introduces an ID feature update module to enhance the object's feature association. To better handle the complex motions under UAV views, we develop an adaptive motion filter module. In addition, a gradient balanced focal loss is used to tackle the imbalance categories and small objects detection problem. Experimental results on the VisDrone2019 and UAVDT datasets demonstrate that the proposed UAVMOT achieves considerable improvement against the state-of-the-art tracking methods on UAV videos.

1. Introduction

Multi-object tracking (MOT) is a fundamental task in computer vision and is widely used in numerous applications [30, 33], such as autonomous driving, intelligent transportation system, and advanced video analysis. MOT methods [5, 45] typically follow the tracking by detection paradigm which includes two steps: detection and data association. The detection step generates potential box predictions of the target objects in every frame while the data association step matches the predicted boxes of the same target across frames based on appearance and motion cues [18]. Recently, multi-object tracking in UAV views has aroused the keen interest of researchers [1, 31, 39, 51] due to the convenience and dexterity of unmanned aerial vehicle (UAV) [9].

Despite the progress made in conventional multi-object tracking (usually tested on the videos captured in a static view), multi-object tracking under the moving UAV views is still challenging. As illustrated in Fig. 1, there are two critical problems need to be solved urgently in the detection and data association stages. In the detection stage, there are usually multiple categories of objects in a moving UAV view and the object numbers of each category are extremely imbalanced, which makes the training of the detection model difficult. In addition, most of the objects in UAV videos are small due to the high altitude of UAV, which further aggravates the difficulty of the detection task. In the data association stage, the challenge lies in the inconsistent appearance and motion information of the target objects which is caused by irregular and fast camera motion and usually results in ID switches. The object motion in UAV videos is a superposition of object movements and the motion of UAV which is irregular and hard to be modelled by the traditional Kalman filter.

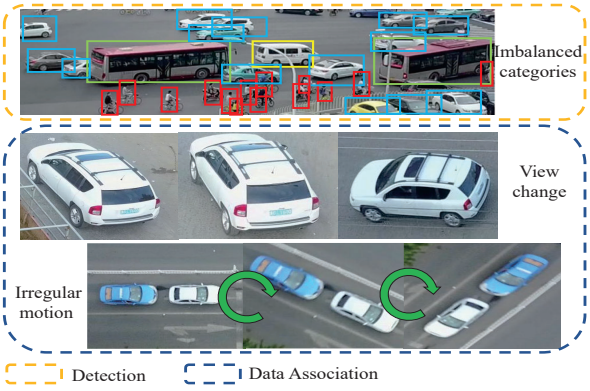

Figure 1. Challenges of MOT on UAV videos. Detection stage: the categories of objects in UAV videos are imbalanced and most of the targets under a UAV view are small. Data association stage: appearance and motion of objects change irregularly and rapidly due to the moving of the UAV camera.

In this paper, we propose a novel multi-object tracker, named UAVMOT network, for multi-object tracking on moving UAV videos. To enhance ID embedding features of objects, we construct an ID feature update (IDFU) mod-

[†] Equal contribution, ∗ corresponding author

ule, where the correlation technology [12] is used to relevant adjacent frame features and the ID embedding features would be updated with the UAV views changing. To tackle the issues caused by UAV motions, we develop an adaptive motion filter (AMF), where an motion mode is used to judge the UAV motion mode and applies different tracking strategies according to the motion mode. Particularly, a local relation filter is specially designed to handle the irregular motions of UAV, which grasps the invariant characteristics that do not change with the UAV moving. Furthermore, to alleviate the issues of imbalanced categories classification and small-scale objects detection, we propose a gradient balanced focal (GBF) loss to supervise the heatmaps learning. The GBF loss combines the equalization loss [29] to balance the imbalanced categories and enhances the small-scale objects detection ability.

We conduct experiments on two public benchmark datasets, i.e., VisDrone2019 dataset [52] and UAVDT dataset [13] to evaluate the proposed algorithm. The experimental results demonstrate that the proposed UAVMOT can accurately track multiple objects in the view of UAV. The key motivation of this work is that a novel multi-object tracker is specifically designed for UAVs. It fully considers the object characteristics in UAV video perspectives, and makes corresponding improvements for multi-object tracking task. The main contributions of this article are summarized as follows:

- We propose an ID feature update module to enhance object ID embedding features, which could update ID features adaptively with UAV changing views.
- We develop an adaptive motion filter for complex motion tracking of objects in UAV videos, which adaptively switches motion filters to adapt to the movement of UAV.
- We design a novel gradient balanced focal loss to supervise the learning of objects' heatmaps, which not only considers the imbalanced categories but also focuses on the small-scale objects in UAV videos.

2. Related work

In this section,we discuss the recent multi-object tracking methods and studies on the data association problem.

Multi-object tracking. The early MOT algorithms follow the two-stage framework of tracking by detection paradigm. The first step is to detect all targets in each video frame and the second step is to associate these detected objects. For example, SORT [3] uses Fast RCNN [15] to detect targets in each frame image, and then uses Kalman filter and Hungarian matching algorithm to complete multiple objects data association. Deep SORT [40] is improved on the basis of SORT, and the idea of cascade matching is proposed to further improve the accuracy of multi-object tracking. To balance the accuracy and speed of MOT, researchers begin to propose single-stage multi-object tracking algorithms. The main framework of the single-stage multi-object tracking algorithm is to add an embedding vector on the detector's head for ReID learning, and this embedding vector is used for multi-object data association in the later stage. For example, JDE [37] extracts a feature vector from feature maps of YOLOv3 [26] for the first time. FairMot [46] adds the learning of embedded vector on the basis of CenterNet [50], to form a multi-target tracker, and achieves good accuracy and speed. CenterTrack [49] directly predicts the displacement of the target's center point.

Recently, transformer technology begins to apply in computer vision and has a good performance on various vision tasks. As for multi-object tracking, researchers regard each tracked target as a query, which contains its ID features and geometric information [8,22,28,43]. For example, Sun et al. [28] propose TransTrack, which applies the transformer technology to the MOT task firstly and builds on the DETR [6] detector. Zeng et al. [43] propose the MOTR to achieve an end-to-end multi-object tracker, which correlates times association in several frames implicitly. Chu et al. [8] propose TransMOT to combine transformer and graphs.

Data association. Data association [44] is a critical step in MOT, especially in the tracking by detection paradigm. It associates the detected objects between two different frames and gives the same object the same ID number. Generally speaking, the data association mainly follows two critical clues: object features and motion laws. For object appearance, similar to the ReID task [47], researchers extract each object's features to distinguish different objects. For example, JDE predicts an ID embedding vector to represent the object ID features. For object motion laws, various filtering methods are used to track the objects, such as the Kalman filter [38], Particle filter [20].

Besides, some researchers covert the data association problem into graph matching problem [21, 27, 32, 35]. Fist, the multi-object tracking process is built into one graph, where each detected object as a node and edges indicate the relation between two detected objects. Then, the graph matching problem can be solved by the min cost global optimization. For example, He [16] et al. propose a novel learnable graph matching method for multiple crowds tracking, which focus on the relationship in intra frame and achieves end-to-end optimization. Wang [36] et al. propose a method that combines the graph networks avoiding the additional data association. Although the graph matching technology can effectively solve the matching problem, it consumes huge computing resources. The proposed adaptive motion filter also considers the relationship between objects, but its form is more concise and the amount of calculation is less.

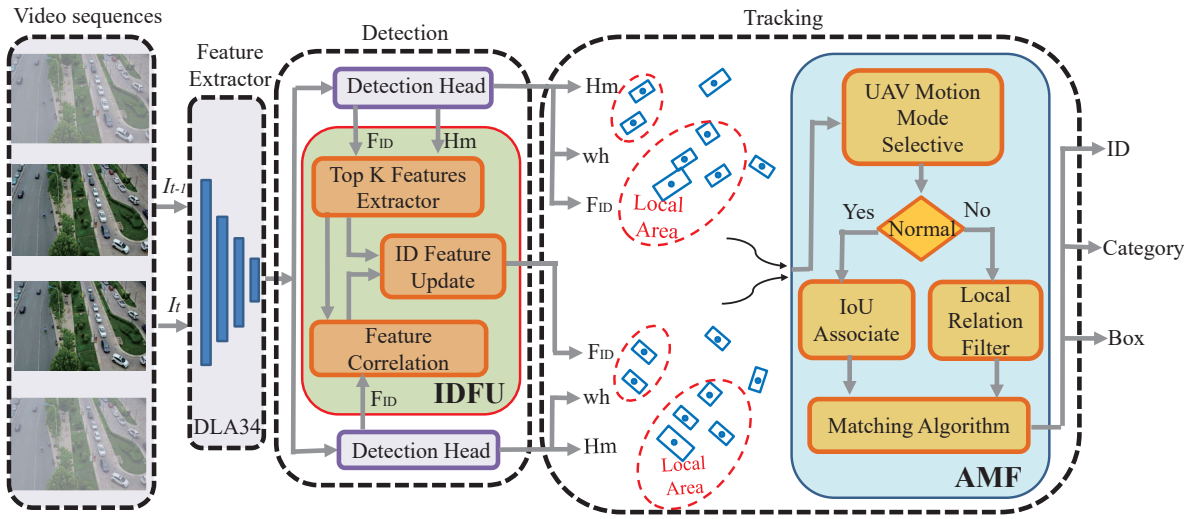

Figure 2. An overview of the proposed UAVMOT. In a UAV video sequence $\{I_t \in \mathbb{R}^{W\times H\times 3}\}_{t=1}^{T}$, take two adjacent frame images I_t and I_{t-1} as input, the UAVMOT uses DLA34 as backbone to extract object features. In object detection head, three parallel branches are built for object bounding boxes sizes (width and height) wh, objects' heatmaps Hm, and tracking ID embedding features F_{ID}, respectively. In the ID embedding feature branch, an ID feature update (IDFU) module is proposed to enhance the ID features learning. In the tracking stage, we design an adaptive motion filter (AMF) to track the objects according to the moving of UAV adaptively. Besides, we propose a gradient balanced focal loss to alleviate the imbalanced categories problem and enhance the small objects detection capacity.

Long tailed object distribution. Long tailed object distribution is a common phenomenon in the real word. The head categories have a large number instances but the tail categories have a few instances. The long trailed distribution brings huge difficulties for object classification, because the network pays more attention to the head categories while neglecting the tailed categories in network training.

To tackle the imbalance of categories, many researchers propose a series of approaches in the literature. On the one hand, some researchers consider it from the perspective of the loss function. For example, Feng [14] et al. utilize mean classification score to indicate the classification learning status and propose an equilibrium loss to balance the classification. Wang [34] et al. propose an adaptive class suppression loss avoiding complex manual grouping. Hsieh [17] et al. propose an adaptive DropLoss for object instance segmentation. On the other hand, some researchers consider the long trailed problem from the perspective of training strategy. For example, Yu et al. [42] propose a dual sampler to perform biased sampling on object proposals for tail and head classes respectively. Zhou et al. [48] propose a novel cumulative learning strategy for classification.

3. UAVMOT Network

3.1. Overall Framework

Given a video sequences $\{I_t \in \mathbb{R}^{W\times H\times 3}\}_{t=1}^{T}$ captured by the moving UAV, our UAVMOT aims to propose the categories $\{C\}_{i=1}^{N}$, bounding boxes $\{B\}_{i=1}^{N}$ and tracking identification $\{ID\}_{i=1}^{N}$ of N objects. The overall framework of UAVMOT is illustrated in Fig. 2. We fed two adjacent frame images I_{t-1} and I_t into UAVMOT network. The two adjacent frames go through the shared feature extraction network and detection head to finish the object detection. The detection head consists of object bounding box size wh, heatmaps Hm, and tracking ID embedding features F_{ID}. We propose an ID feature update (IDFU) module to strengthen the ID embedding features connection between two adjacent frames. We build an adaptive motion filter (AMF) to tackle the objects' complex motions in moving UAV videos. Besides, to alleviate the imbalanced categories and enhance small-scale objects detection capacity, we propose a gradient balanced focal (GBF) loss to supervise the learning of objects' heatmaps.

3.2. ID Feature Update

In UAVMOT, the ID embedding features are used to identify the ID information of each object and are critical for the data association. However, The characteristics of the objects will change with the UAV moving, which are not conducive to ID embedding features learning and harmful to later feature association. To enhance objects features association, inspired by correlation layer in [12,19], we propose an ID feature update (IDFU) module for ID embedding features learning in two adjacent frames. The IDFU module extracts the previous frame object features to associate with current frame features, which can adaptively update the ID embedding features in various UAV views.

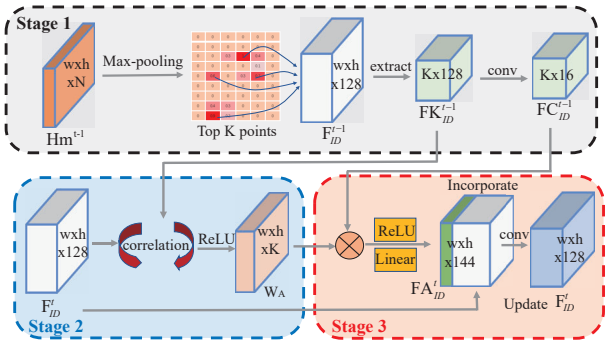

Figure 3. Architecture of IDFU module. The IDFU module consists of three stages: First stage, previous frame features extraction. Second stage, two adjacent frame features correlation. Third stage, ID embedding features update.

As illustrated in Fig. 3, the IDFU module consists of three stages. First, we extract the object ID embedding features F_{ID}^{t-1} in previous frame I_{t-1}. To simplify the feature computing, we only extract top K keypoints ID embedding features FK_{ID}^{t-1} by selecting corresponding top K points in the heatmaps. And the FK_{ID}^{t-1} is compressed from 128 to 16 dimensions to obtain the compress ID features FC_{ID}^{t-1} for the following feature update. Second, we get feature enhance attention weights W_A via two adjacent frames features correlation operation. This correlation attention weights W_A guide the network where should be focused on in current frame. The W_A are future combined with FC_{ID}^{t-1} by multiplication and obtain the previous frame attention features FA_{ID}^{t-1} through a series operations. Finally, the attention features FA_{ID}^{t-1} incorporate current frame ID embedding features F_{ID}^{t}, and through convolution to finish object ID embedding features update.

3.3. Adaptive Motion Filter

In UAV video sequences, the object movement is no longer a linear motion, but a nonlinear motion formed by the coupling of the motion of the UAV and the object itself. The traditional Kalman filter is difficult to deal with this irregular motion, we propose an adaptive motion filter (AMF) to handle the complex UAV movement. The AMF module adaptively switches different filters according to different motion modes of UAV, which can accurately complete the object ID association.

UAV motion mode selective. According to the movement of UAV, the motion of objects in UAV videos can be roughly divided into two modes: normal mode and abnormal mode. In normal mode, the UAV flies smoothly and normally in the sky, and the objects' movement in the video can be regarded as approximate linear motion; In abnormal mode, the UAV rotates or accelerates suddenly, and the objects movement in UAV videos presents a kind of nonlinear motion. Particularly, we perform Kalman filter on the objects between two adjacent frames and compute the objects matching number. When the matching number is higher than a certain threshold p, we consider it as normal mode, and vice versa. The AMF module adopts IoU association and local relation filer in two motion modes, respectively.

Local relation filter. The local relation filter aims to create a filter, avoiding being affected by the external motion of the UAV. Fortunately, we notice that the positional relationship between objects stays basically invariant in a local area between two adjacent frames. To make good use of this permanent characteristic of local positional relation, we propose a local relation filter.

The local relation filter designs a relative relation vector v to describe the positional relationship between the object and the surrounding objects in a local area. There are many relative positional relationships around each object, to simplify the calculation, the relative relation vector v only consists of three elements: the length l_{max} from the farthest object, the length l_{min} from the nearest object and include angle θ between these two objects in the local area. As illustrated in Fig. 4, we draw two frames of detected objects distribution and each dot represents a detected object. Take the red dot P_1 as an example, take it as the center point and the circle with the radius of R as the local area of P_1. In the local area, find the nearest point P_2 and the farthest point P_8. We present the relative relation vector $v = [\theta, l_{max}, l_{min}]$ and $v' = [\theta', l'_{max}, l'_{min}]$ in two adjacent frames, respectively. Obviously, the relative relation state vector remains invariant basically without affecting by UAV moving.

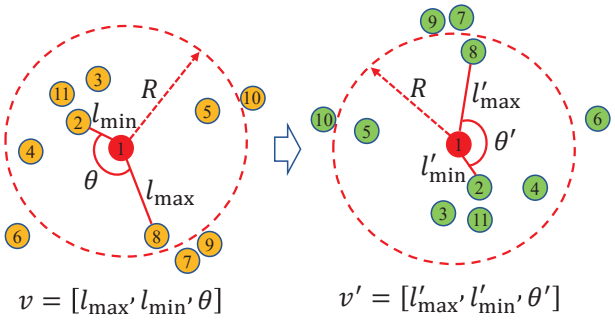

Figure 4. Schematic diagram of relative relation vector.

When switching to the abnormal mode, we obtain m detected objects and compute its relative relation vector v in videos frames I_t. These relative relation vectors are used to construct the cosine similarity matrix with n tracked objects to measure the position similarity, and we obtain a $m \times n$ location similarity matrix M_L. Then, the ID embedding features are used to construct the ID feature similarity matrix M_F, we fuse the M_L with the M_F, and get the last similarity matrix M for matching algorithm. The overall

UAVMOT algorithm can be summarized as Algorithm 1.

Algorithm 1 UAVMOT algorithm

Input: An UAV video sequences $\{I_t \in \mathbb{R}^{W\times H\times 3}\}_{t=1}^{T}$
Output: The tracked objects $T_t = \{B_t, C_t, ID_t\}$

while (t<T) **do**
  Input two adjacent frames images I_{t-1}, I_t.
  Finish object detection, obtain two frames detected objects $O_{t-1} = \{B_{t-1}, C_{t-1}\}$, $O_t = \{B_t, C_t\}$.
  Obtain ID embedding features F_{ID}^{t-1}, F_{ID}^{t}, finish the feature association.
  Kalman filter and judge the object motion mode.
  if normal mode **then**
    IoU association.
  else
    Local relation filter.
  end if
  Matching algorithm, get ID_t.
end while

3.4. Gradient Balanced Focal Loss

The environment in UAV videos is far more complex than the traditional multi-object tracking on crowds, in which two prominent problems affect the performance of detection: imbalanced categories and small-scale objects detection. To tackle these two problems, we propose a gradient balanced focal (GBF) loss to supervise objects' heatmaps learning. The GBF loss not only alleviates the imbalance between categories but also pays attention to small-scale objects.

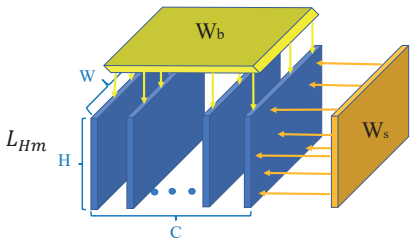

Figure 5. Schematic diagram of GBF loss.

Particularly, the GBF loss is improved on the original Cross-Entropy loss L_{Hm} and two adaptive wights are designed to reweight the loss for objects heatmaps learning: categories balanced weights W_b and small-scale object attention weights W_s. As illustrated in Fig. 5, W_b is used to balance the categories and W_s is used to focus on the small-scale objects. These two adaptive loss weights adjust its self considering on the imbalanced categories and the sizes of objects, respectively. The GBF loss is defined as:

$$GBF = W_b \cdot W_s \cdot L_{Hm} \tag{1}$$

The small-scale object attention weights W_s focus on the small-scale objects, and give the small object a larger weight. Particularly, we measure the size of objects by the area of the bounding box, so the W_s is defined as:

$$W_s = e^{-(w\cdot h-\mu)} + 1 \tag{2}$$

where w and h indicate the width and height of object bounding box, respectively. The $\mu = 5$ in this article.

The categories balanced weights W_b give different weights to positive samples and negative samples according the corresponding gradients, and the W_b is defined as:

$$W_b = pos\_w \cdot Hm + neg\_w \cdot (1 - Hm) \tag{3}$$

where pos_w and neg_w indicate the weights of positive samples and negative samples, respectively. They will update adaptively with the network training and the specific update process can be referred to [29].

4. Experiments

4.1. Dataset and Metrics

Dataset. To validate the effectiveness of UAVMOT, we conduct a series of experiments on VisDrone2019 dataset and UAVDT dataset.

VisDrone2019 dataset [52] is used for tracking and detection in UAV views. In MOT task, the VisDrone2019 dataset consists of training set (56 sequences), validation set (7 sequences) and test set (33 sequences (test-challenge: 16 sequences, test-dev: 17 sequences)). In each frame, every object is annotated by bounding box, category and tracking ID. The VisDrone2019 dataset includes 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., car, bus, truck, pedestrian, and van.

UAVDT dataset [13] is specially used for vehicle object detection and tracking, and it consists of three categories: car, truck, and bus. In MOT task, it divides into training set (30 sequences) and test set (20 sequences). And it only considers a single category car. The video images have a resolution of 1080 × 540 pixels and includes various common scenes, such as squares, arterial streets and toll stations.

Metrics. To evaluate UAVMOT with other state-of-the-arts approaches, we adopt multiple metrics to measure the performance of tracking [23], such as multiple object tracking accuracy (MOTA), multiple object tracking precision (MOTP), ID switching (IDs) and and other metrics.

$$MOTA = 1 - \frac{FP + FN + IDs}{GT} \tag{4}$$

where FP, FN and GT are the number of false positive samples, false negative samples and ground truth.

Dataset	Method	MOTA↑(%)	MOTP↑(%)	IDF1↑(%)	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓
VisDrone2019	MOTDT [7]	-0.8	68.5	21.6	87	1196	44548	185453	1437	3609
	SORT [3]	14.0	73.2	38.0	506	**545**	80845	**112954**	3629	4838
	IOUT [4]	28.1	74.7	38.9	467	670	36158	126549	2393	3829
	GOG [25]	28.7	**76.1**	36.4	346	836	**17706**	144657	1387	**2237**
	MOTR [43]	22.8	72.8	41.4	272	825	28407	147937	**959**	3980
	TrackFormer [22]	25	73.9	30.5	385	770	25856	141526	4840	4855
	Ours	**36.1**	74.2	**51.0**	**520**	574	27983	115925	2775	7396
UAVDT	CEM [24]	-6.8	70.4	10.1	94	1062	64373	298090	1530	**2835**
	SMOT [10]	33.9	72.2	45.0	524	367	57112	166528	1752	9577
	GOG [25]	35.7	72	0.3	627	374	62929	153336	3104	5130
	IOUT [4]	36.6	72.1	23.7	534	357	42245	163881	9938	10463
	CMOT [2]	36.9	**74.7**	57.5	**664**	351	69109	144760	1111	3656
	SORT [3]	39	74.3	43.7	484	400	**33037**	172628	2350	5787
	DSORT [40]	40.7	73.2	58.2	595	338	44868	155290	2061	6432
	MDP [41]	43.0	73.5	61.5	647	324	46151	147735	541	4299
	Ours	**46.4**	72.7	**67.3**	624	**221**	66352	**115940**	**456**	5590

Table 1. Quantitative comparisons between UAVMOT and other methods for MOT task on VisDrone2019 test-dev set and UAVDT test set.

4.2. Implementation Details

Training. We utilize random crop, random scaling (between 0.6 to 1.3) as data augmentation. We use multiple loss functions for careful supervision and the initial learning rate sets to 7e-5. We train the network 30 epochs in total and the learning rate decays 10 times at 10 epochs and 20 epochs, respectively. We conduct experiments on two GeForce RTX 2080Ti GPUs with batch size 4. In multiple loss functions, L1 loss is used to supervise the object width and height. Crossentrpy loss and triplet loss [11] are used to deal with the object ID. Besides, we use the proposed GBF loss to supervise the object heatmaps.

Inference. The UAVMOT follows the tacking by detection paradigm. At the detection phase, the detection score threshold is set to 0.4 and the number K is set to 100 in IDFU module. At the tracking phase, the threshold p is set to 0.6 in the AMF module.

4.3. Comparison with State-of-the-arts

VisDrone2019 dataset. We compare our method with previous methods on VisDrone2019 dataset for MOT task. We train the training set together with the validation set and evaluate our approach on VisDrone2019 test-dev set using the official VisDrone MOT toolkit. As illustrated in Tab. 1, our method achieves 36.1% on MOTA and 51.0% on IDF1, which outperform the exiting approaches on VisDrone2019 test-dev set.

UAVDT dataset. We also compare our method with other methods on UAVDT test set for the MOT task. We train the UAVMOT network using UAVDT training set and evaluate our approach on UAVDT test set. we list a series of indicators such as MOTA, MOTP and IDF1 to compare the performance of our method with other methods. As illustrated in Tab. 1, our method achieves 46.4% on MOTA and 67.3% on IDF1, and gets significantly better results against existing methods.

4.4. Ablation Study

In this section, we conduct a series of ablation experiments on VisDrone2019 validation set and test-dev set to verify each module of UAVMOT. In ablation experiments, we use FairMot as the baseline model and the DLA-34 as the backbone network.

Baseline	IDFU	AMF	GBF	MOTA↑(%)	IDs↓	IDF1↑(%)
✓				20.1	2079	40.6
✓	✓			23.3	1974	43.8
✓	✓	✓		23.7	**867**	45.5
✓	✓	✓	✓	**26.7**	969	**45.8**

Table 2. Ablation study on VisDrone2019 validation set.

As illustrated in Tab. 2, there are three core components in UAVMOT, IDFU module, AMF module and GBF loss, we report three critical metrics of each module on the VisDrone2019 validation set. The baseline model gets 20.1% on MOTA, 40.6% on IDF1 and 2079 on IDs. Adding the IDFU module to the baseline model, the MOTA improves to 23.3%, the IDs decreases to 1974 and achieves 43.8% on IDF1. Adding on the IDFU module and AMF module to the baseline model, the MOTA improves to 23.7%, the IDF1 improves to 45.5% and the IDs decreases from 1974 to 867. Adding all three modules, our UAVMOT model achieves 26.7% on MOTA and 45.8% on IDF1.

Effectiveness of IDFU module. The IDFU module enhances the ID embedding features association, which can effectively adapt to the change of UAV view. To evaluate the effectiveness of IDFU module, we list four critical ID

association indicators (IDS, IDF1, IDP, IDR) on the baseline model and the baseline+IDFU model, respectively. As illustrated in Tab. 3, The IDs from 2079 decreases to 937. The IDF1, IDP and IDR increase from 40.6%, 53.2% and 32.8% to 43.8%, 57.9% and 35.3%, respectively. The results demonstrate that the IDFU model has a good effect on the data association, which can grasp the objects' characters accurately in moving UAV videos.

	IDs↓	IDF1↑(%)	IDP↑(%)	IDR↑(%)
Baseline	2079	40.6	53.2	32.8
Baseline+IDFU	**937**	**43.8**	**57.9**	**35.3**

Table 3. Analysis of the effectiveness of IDFU module. We report the IDs, IDF1, IDP and IDR on VisDrone2019 validation set.

Effectiveness of AMF module. the AMF module can automatically switch the tracking filter mode according to the motion of UAV. To evaluate the effectiveness of AMF, we list ID association indicators (IDS, IDF1) and detection indicators (recall rate, precision rate) on the baseline model and the baseline+ADA model, respectively. As illustrated in Tab. 4, The IDs from 2079 decreases to 1048 and the IDF1 increases from 40.6% to 44.1%. Besides, the recall increases from 41.5% to 46.5% and the precision slightly decreases from 67.4% to 66.6%. The results demonstrate that the AMF module has a good effect on the data association, and the contribution mainly comes from the improvement of recall rate.

	IDs↓	IDF1↑(%)	Recall↑(%)	Precision↑(%)
Baseline	2079	40.6	41.5	**67.4**
Baseline+AMF	**958**	**44.1**	**46.5**	66.6

Table 4. Analysis of the effectiveness of AMF module. We report the IDs, IDF1, recal and precision on VisDrone2019 validation set.

Effectiveness of GBF loss. To verify the effectiveness of gradient balanced focal loss, we compare the MOTA of each category between the baseline model and after using GBF loss. As illustrated in Fig. 6, each category in VisDrone2019 test-dev set has a great improvement on MOTA after the baseline using GBF loss, especially the tail categories, ie, the van from 4.6% improves to 11.7% on MOTA, the truck from 16.3% improves to 25% on MOTA. Besides, the small-scale category (pedestrian) from 14.2% improves to 20.2% on MOTA. These results demonstrate that the GBF loss can effectively improve the MOTA of small number categories and small-scale objects.

4.5. Case Study

To better prove the advantages of UAVMOT in moving UAV videos, we analyze three UAV special motion cases:

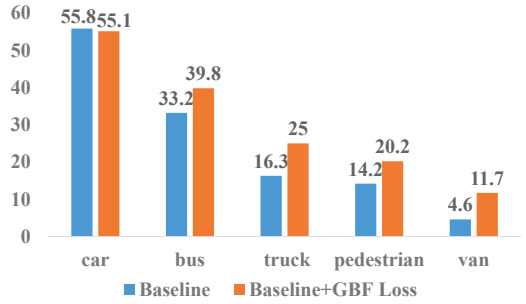

Figure 6. Analysis of the effectiveness of GBF loss. We report the MOTA of each category in VidsDrone2019 test-dev set.

UAV hovers in the sky, turns left and right, moves up and down suddenly.

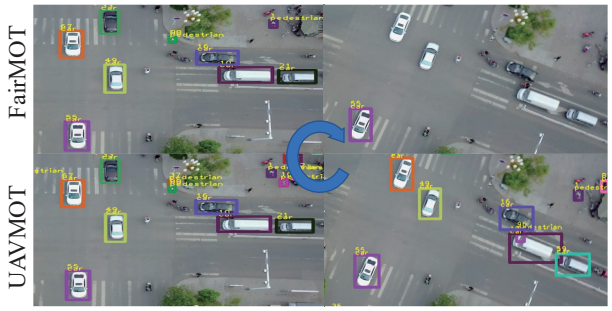

Figure 7. Analysis of the special case: UAV hovers in the sky.

UAV hovers in the sky. When the UAV hovers in the sky, the position of objects captured in the UAV videos will rotate with the UAV hovering. we compare the visualization results of UAVMOT with the FairMOT on this special case, As illustrated in Fig. 7, the FairMOT could not track the cars when the UAV rotates quickly in the sky, but UAVMOT can accurately track the cars without being affected by the rotation of UAV.

Figure 8. Analysis of the special case: UAV moves up and down.

UAV moves up and down. When the UAV moves up suddenly, the objects sizes in UAV videos will become small

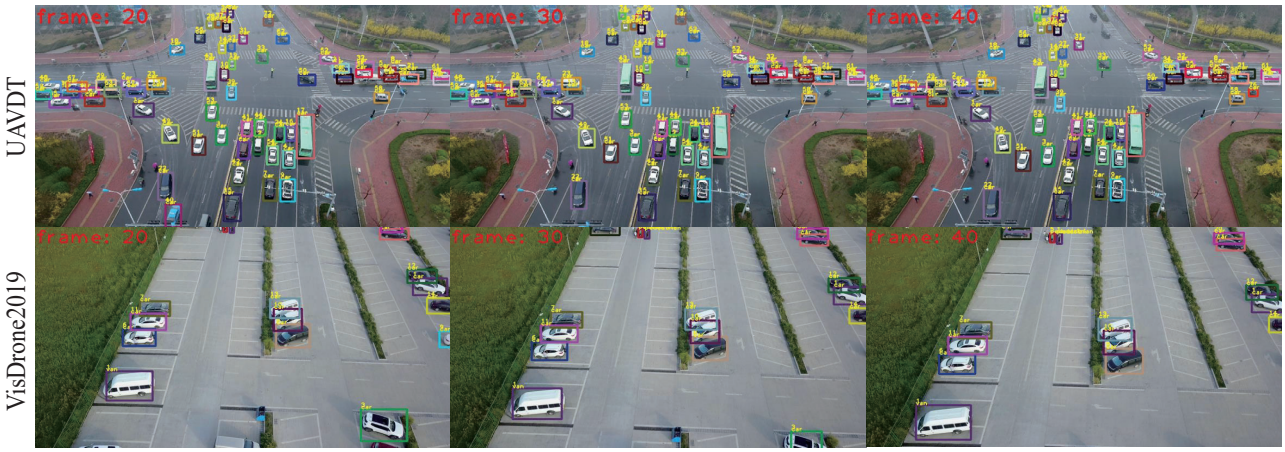

Figure 9. Visualization of tracking results on Visdrone2019 and UAVDT datasets.

and are difficult to be detected. As illustrated in Fig. 8, the UAV moves up suddenly and the objects in the video become small, especially the pedestrian and the cars in the distance, these objects are difficult to be tracked in the FairMOT but are accurately tracked in UAVMOT.

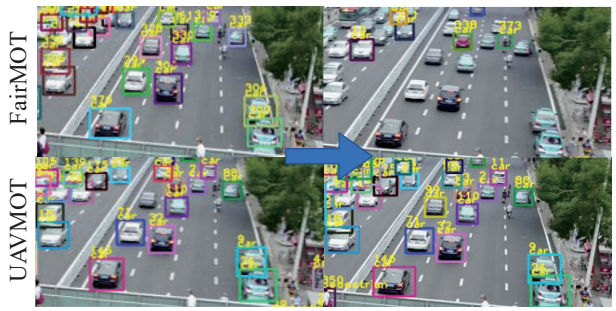

Figure 10. Analysis of the special case: UAV turns left and right.

UAV turns left and right. When the UAV turns left and right suddenly, the captured images in UAV videos will become blurred and the original motion trend law will be broken. As illustrated in Fig. 10, the FairMOT could not track the cars when the UAV turns right quickly, but UAVMOT can accurately track the cars without being affected by the turns of UAV.

4.6. Visualization

To show the effectiveness of our method more intuitively, we draw the tracking results on VisDrone2019 test-dev set and UAVDT test set. As illustrated in Fig. 9, the UAVMOT can well adapt to the moving UAV environment, small-scale objects are accurately detected and the multiple objects tracking results are not affected by UAV motion.The visualization results demonstrate that UAVMOT can well complete MOT task on UAV videos.

4.7. Limitations

The UAVMOT network can complete MOT task effectively in moving UAV videos. Due to the MOT algorithm is loaded in the moving UAV and limited hardware of mobile equipments, the parameters and running speed of the algorithm should be matched with the UAV equipment. The UAVMOT network performs 12 FPS on DLA34 backbone network with a video resolution of 1920×1080, we will explore a smaller parameters model to obtain real-time running speed on the mobile UAV equipment in the future.

5. Conclusions

This paper proposes a novel UAVMOT network for multi-object tracking in UAV videos. In UAVMOT, an ID feature update module is designed to enhance the ID embedding features learning. To adapt to the complex UAV motions, the adaptive motion filter gives different motion filters to different motion modes. Besides, a gradient balanced focal loss is proposed to supervise the objects' heatmap learning, which not only considers the imbalanced categories but also focus more attention on small-scale objects. We conduct a series of experiments on VisDrone2019 and UAVDT datasets, and compare UAVMOT with other methods. The results demonstrate that our method achieves state-of-the-art performance on UAV videos for MOT task.

6. Acknowledgments

The paper is supported in part by China Postdoctoral Science Foundation (Grant No.2021M701803), the National Key R&D Program of China (Grant No.2018AAA0102001), National Natural Science Foundation of China (Grant No.61725202, U1903215, 61829102, 62022092), and Dalian Innovation leader's support Plan (Grant No.2018RD07).

References

[1] Seyed Majid Azimi, Maximilian Kraus, Reza Bahmanyar, and Peter Reinartz. Multiple pedestrians and vehicles tracking in aerial imagery: A comprehensive study. *arXiv preprint arXiv:2010.09689*, 2020. 1

[2] Seung-Hwan Bae and Kuk-Jin Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1218–1225, 2014. 6

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 2, 6

[4] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017. 6

[5] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6247–6257, 2020. 1

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2

[7] Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *2018 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2018. 6

[8] Peng Chu, Jiang Wang, Quanzeng You, Haibin Ling, and Zicheng Liu. Transmot: Spatial-temporal graph transformer for multiple object tracking. *arXiv preprint arXiv:2104.00194*, 2021. 2

[9] Ibrahim Delibasoglu. Uav images dataset for moving object detection from moving cameras. *arXiv preprint arXiv:2103.11460*, 2021. 1

[10] Caglayan Dicle, Octavia I Camps, and Mario Sznaier. The way they move: Tracking multiple targets with similar appearance. In *Proceedings of the IEEE international conference on computer vision*, pages 2304–2311, 2013. 6

[11] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 459–474, 2018. 6

[12] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2, 3

[13] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 370–386, 2018. 2, 5

[14] Chengjian Feng, Yujie Zhong, and Weilin Huang. Exploring classification equilibrium in long-tailed object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3417–3426, 2021. 3

[15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[16] Jiawei He, Zehao Huang, Naiyan Wang, and Zhaoxiang Zhang. Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5299–5309, 2021. 2

[17] Ting-I Hsieh, Esther Robb, Hwann-Tzong Chen, and Jia-Bin Huang. Droploss for long-tail instance segmentation. *arXiv preprint arXiv:2104.06402*, 2021. 3

[18] Chang Huang, Bo Wu, and Ramakant Nevatia. Robust object tracking by hierarchical association of detection responses. In *European Conference on Computer Vision*, pages 788–801. Springer, 2008. 1

[19] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 3

[20] Yonggang Jin and Farzin Mokhtarian. Variational particle filter for multi-object tracking. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 2

[21] Qiankun Liu, Qi Chu, Bin Liu, and Nenghai Yu. Gsm: Graph similarity model for multi-object tracking. In *IJCAI*, pages 530–536, 2020. 2

[22] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. *arXiv preprint arXiv:2101.02702*, 2021. 2, 6

[23] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 5

[24] Anton Milan, Stefan Roth, and Konrad Schindler. Continuous energy minimization for multitarget tracking. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):58–72, 2013. 6

[25] Hamed Pirsiavash, Deva Ramanan, and Charless C Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR 2011*, pages 1201–1208. IEEE, 2011. 6

[26] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[27] Samuel Schulter, Paul Vernaza, Wongun Choi, and Manmohan Chandraker. Deep network flow for multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6951–6960, 2017. 2

[28] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 2

[29] Jingru Tan, Xin Lu, Gang Zhang, Changqing Yin, and Quanquan Li. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1685–1694, 2021. 2, 5

[30] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8797–8806, 2019. 1

[31] Leon Amadeus Varga, Benjamin Kiefer, Martin Messmer, and Andreas Zell. Seadronessee: A maritime benchmark for detecting humans in open water. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2260–2270, 2022. 1

[32] Gaoang Wang, Renshu Gu, Zuozhu Liu, Weijie Hu, Mingli Song, and Jenq-Neng Hwang. Track without appearance: Learn box and tracklet embedding with local and global motion patterns for vehicle tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9876–9886, 2021. 2

[33] Gaoang Wang, Xinyu Yuan, Aotian Zheng, Hung-Min Hsu, and Jenq-Neng Hwang. Anomaly candidate identification and starting time estimation of vehicles from traffic videos. In *CVPR Workshops*, pages 382–390, 2019. 1

[34] Tong Wang, Yousong Zhu, Chaoyang Zhao, Wei Zeng, Jinqiao Wang, and Ming Tang. Adaptive class suppression loss for long-tail object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3103–3112, 2021. 3

[35] Yongxin Wang, Kris Kitani, and Xinshuo Weng. Joint object detection and multi-object tracking with graph neural networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13708–13715. IEEE, 2021. 2

[36] Yongxin Wang, Kris Kitani, and Xinshuo Weng. Joint object detection and multi-object tracking with graph neural networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13708–13715. IEEE, 2021. 2

[37] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 107–122. Springer, 2020. 2

[38] Greg Welch, Gary Bishop, et al. An introduction to the kalman filter. 1995. 2

[39] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Detection, tracking, and counting meets drones in crowds: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7812–7821, 2021. 1

[40] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 2, 6

[41] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE international conference on computer vision*, pages 4705–4713, 2015. 6

[42] Weiping Yu, Taojiannan Yang, and Chen Chen. Towards resolving the challenge of long-tail distribution in uav images for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3258–3267, 2021. 3

[43] Fangao Zeng, Bin Dong, Tiancai Wang, Cheng Chen, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. *arXiv preprint arXiv:2105.03247*, 2021. 2, 6

[44] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[45] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864*, 2021. 1

[46] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, pages 1–19, 2021. 2

[47] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017. 2

[48] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[49] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490. Springer, 2020. 2

[50] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2

[51] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *arXiv preprint arXiv:2001.06303*, 2020. 1

[52] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Qinghua Hu, and Haibin Ling. Vision meets drones: Past, present and future. *arXiv preprint arXiv:2001.06303*, 2020. 2, 5