2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

DC-MOT: Motion Deblurring and Compensation for Multi-Object Tracking in UAV Videos

Song Cheng, Meibao Yao*, and Xueming Xiao

***Abstract*—In this paper, we propose a multi-object tracking framework for videos captured by UAVs, considering motion imperfection in the following two aspects: 1) motion blurring of objects due to high-speed motion of the UAV and the objects, deteriorating the performance of the detector; 2) motion coupling of the global movement of the UAV camera with the object motion, resulting in the nonlinearity of objects trajectories in adjacent frames and further more difficult to predict. For motion blurring, this paper proposes a hybrid deblurring module that deals with the blurred frames while retaining the clear frames, trading off between video tracking performance and spatio-temporal consistency. For motion coupling, we proposed a motion compensation module to align adjacent frames by feature matching, and the corrected target position is obtained in the next frame to alleviate the interference of camera movement with tracking. We evaluate the proposed methods on VisDrone dataset and validate that our framework achieves new state-of-the-art performance on UAV-based MOT systems.**

***Index Terms*—Multi-object Tracking, Motion Deblurring, Global Motion Compensation, UAV Vision**

I. Introduction

Multi-object tracking algorithm based on computer vision has gradually become an attractive research field. With the increasing popularity and rise of commercial UAVs, efficient multi-object tracking algorithms based on the UAV platform provide a new technical perspective for autonomous driving, collaborative navigation, urban safety and disaster relief.

Although promising, it is still challenging for most current MOT algorithms to cope with UAV video due to high-speed camera movements and drastic object changes. There are two main paradigms for multi-object tracking methods: tracking-by-detection (TBD) [1]–[6] and joint detection and tracking (JDT) [7]–[13]. On the one hand, although the JDT paradigm can simultaneously perform both detection and tracking in a forward operator, most re-identification (ReID) modules and detectors share the same backbone network, and ReID loss and detection loss are commonly contradictory [7]–[9]. If the detection loss is to be reduced, then the ReID loss will be increased, which will inevitably affect the detection performance. On the other hand, TBD paradigm treats detection and tracking as two separate tasks. In the detection phase, the detector workflow is not affected by the tracking module, and the detector becomes the cornerstone of tracking. Therefore TBD is more efficient and becomes the mainstream solution for UAV-based MOT systems. [14].

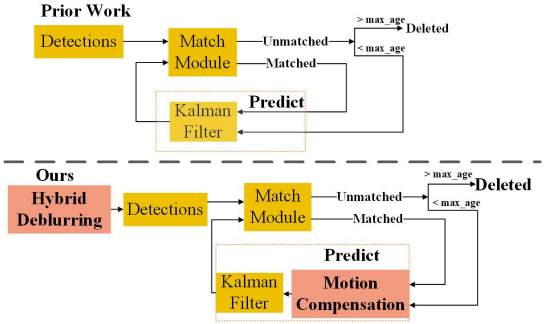

Fig. 1: (**Top**) Prior work: Cascaded matching of detected objects in each frame to obtain objects associated with the trajectory, new objects or lost trajectories. The Kalman filter is used to predict the linear motion of existing trajectories and new objects. (**Bottom**) Ours: A multi-object tracking framework with motion deblurring and compensation modules. Two main innovations: (1) the video has been processed by a hybrid deblurring module before detection to clarify the blurred frame caused by the high-speed movement of the UAV; (2) in the prediction stage, global motion compensation is performed to mitigate the prediction error caused by camera movement.

Although TBD methods can balance the accuracy of the detector and tracker, and achieve good results in object detection of video captured by static cameras, their performance is much less satisfactory with UAV video. There are two main challenges with mainstream TBD methods for UAV tracking: (1) Most TBD methods [1]–[6] are validated by videos collected by fixed cameras, where motion ambiguity of the object is usually processed using the detector. Since both the camera and the observed object are moving, it is very common to cause more motion blurring in videos captured by UAVs. Some UAV MOT systems [14]–[16] focus on small objects and rely only on deepening the number of convolutional network layers to improve detection capability.

This work was sponsored by the National Natural Science Foundation of China (NSFC) through grants No. 62103163 and No. 62003055, Natural Science Foundation of Jilin Province through grant No. YDZJ202101ZYTS033. We thank the above mentioned funds for their financial support.

Song Cheng and Meibao Yao are with the Intelligent Robotics Lab (IRL), School of Artificial Intelligence, Jilin University, Changchun 130012, China; Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, Ministry of Education, China. Email: chengsong20@mails.jlu.edu, meibaoyao@jlu.edu.cn. Xueming Xiao is with the CVIR lab, Changchun University of Science and Technology, 130022, China; Key Lab of Optoelectronic Measurement and Optical Information Transmission Technology, Ministry of Education, China, alexcapshow@gmail.com

*Meibao Yao is Corresponding Author.

However, in the blurred frame, the features of the objects change greatly, which is difficult to identify through the detection network. Therefore, it is difficult to achieve accurate detection by relying on the detector alone, and motion deblurring of the video is a key supplement. (2) Motion prediction in most TBD methods is achieved by Kalman filtering, where the relative motion of an object is considered linear. However, UAV video is the coupling of two parts: the motion of the camera and objects. Therefore, linear motion prediction is not appropriate in the UAV MOT system. Some tracking methods [3], [17] use motion compensation, but due to the inaccurate feature matching method, the deviation is larger when applied to the UAV MOT system. Therefore, in the UAV MOT system, accurate global motion compensation is necessary to offset camera motion and lead to more accurate prediction of object motion. In this paper, we propose a new UAV MOT framework, which employs extra modules, see Fig.1: Hybrid **D**eblurring and AdaLAM-based Global Motion **C**ompensation, named DC-MOT.

Fig. 2: Comparison of tracking performance before and after deblurring video frames: (a)(b) Positive effect: after deblurring, the undetected objects in the blurred frame become clear and can be detected by the detector. (c)(d) Negative effect: inappropriate deblurring will lead to spatio-temporal inconsistency, i.e. adding unrelated features in locations between adjacent frames, and further misleads the detector.

Through extensive experiments, we observe that the motion deblurring of each frame is a double-edged sword in more accurate MOT of UAV videos, as shown in Fig.2. 1) Positive effect (Fig.2 (a)(b)): after deblurring, optical flow information for adjacent frames is used to restore the blurred scene for the detector to disclose more objects and thus enhance the tracking accuracy; 2) Negative effect (Fig.2 (c)(d)): The commonly used optical flow methods in MOT add spatio-temporal feature information, and can better overcome the interrupted tracking, leading a lower identity switches (IDs). However, the generated deblurring frames add some pixels in the blurred regions, resulting in a larger FPs [18] then a lower $MOTA$ [18]. Therefore, deblurring all video sequences by applying the existing deblurring methods cannot guarantee enhanced tracking performance. To this end, we propose a hybrid deblurring strategy to select frames that are more suitable for subsequent deblurring and tracking by an encoder, with PSNR thresholds designed for different sequences according to the GTs [18], where both spatio-temporal consistency and detection accuracy are considered to achieve better overall tracking performance.

The main contributions of this paper are as follows:

- In this paper, we proposed a new MOT framework for UAV video, which uses a hybrid deblurring module to improve detection quality and obtain better tracking results. To the best of our knowledge, this work is the first to use motion deblurring to improve UAV-based MOT performance.
- We proposed a feature matching method as a filter for global motion compensation, which can counteract horizontal, vertical, and rotational camera motion and enables more accurate prediction of object motion.
- Extensive experiments demonstrate that our framework with motion deblurring and compensation can improve multi-object tracking performance on UAV videos.

II. Related Work

A. TBD Multi-Object Tracking

In TBD methods, multi-object tracking is usually decomposed into two independent stages: the bounding box is first obtained by object detection in each frame; then the association module adds the target to an existing trajectory or creates a new trajectory. As a representative of TBD methods, DeepSort [2] is proposed using CNN to extract the depth feature information of objects, and significantly reduced the number of IDs without losing too much tracking speed and accuracy. However, due to the simple structure of the detector network, object recognition capability is limited, and FPs remains large. To solve this issue, Tracktor [3] is proposed with Faster RCNN [19] as a detector to perform object detection in video frames, which has the highest efficiency in object detection at that time. In addition, a motion compensation model for low-speed camera motion is proposed. However, the above TBD methods can barely cope with UAV video scenes and datasets with fast movement of the cameras.

To this end, we propose a new framework for UAV video MOT system to improve the accuracy of motion compensation. Both local spatial consistency and the method of global geometry verification are considered in motion compensation by feature matching. Our tracking method is based on DeepSort and uses Yolov5 [20] as a detector, fitting with both accuracy and speed.

B. Video Deblurring

With the success of adopting CNNs in image restoration [21], deblurring networks [22] [23] [24] based on encoder and decoder architecture are widely used in video deblurring. An end-to-end solution [22] is presented to train a deep neural network to learn deblurring given adjacent frames. To make better use of spatial and temporal information, Kim [23] developed an optical flow estimation method for aligning and

aggregating information between adjacent frames to recover the underlying spatio-temporal flow information. Pyramid, cascade and deformable convolutions are used in [24] to achieve better alignment performance of two frames. A deep CNN model [25] with temporal and spatial attention strategies is used to discover potential targets. It is noted that the success of these algorithms in video deblurring is due to the use of large detection models, but their generalizability in practical applications is limited, which is likely to cause overfitting and deteriorate the deblurring performance. Different from these methods, CDVD [26] is more compact and lightweight, and combines the advantages of optical flow method. Instead of expanding the capacity of network models, video deblurring enhances the generalization ability of the model and greatly improves the performance of downstream tasks such as object detection and tracking.

In this work, we propose a deblurring module for UAV MOT system based on CDVD, with a hybrid strategy to select frames for deblurring to achieve the most satisfactory tracking performance.

C. Global Motion Compensation

For global motion compensation, we aim to align the global features of adjacent frames, so we can formulate it as an image matching problem. The classical image matching methods, such as SIFT [27] and ORB [28], extract local features from two images, match them to find the correspondence to find the corresponding feature points set. Then, the homography matrix is calculated by selecting the data from the points set. RANSAC [29] and its improved methods [30]–[32] play a part by repeatedly selecting a random subset of the data. However, it is inefficient to perform random search without filtering. Some methods have been proposed to alleviate this problem. Simple low-level filters based only on descriptors, such as ratio test [27], can filter out ambiguous matches and outliers pruned by Hamming distance threshold. The heuristic algorithms are efficient and easy to implement, but they are likely to cause many outliers.

To achieve more accurate matches, the following two methods are commonly used by recent studies in feature matching: (1) Local spatial consistency filter [33]–[36] based on the observation that correct matches should be consistent with other correct matches in its vicinity, while an incorrect match should be vice versa. Methods performed at this level can also be very efficient and can select more valuable information than simple filters. (2) The method of global geometric verification methods [31], [37], [38] is based on global transformations on which the correct correspondence must be consistent. This can be done by fitting a global transformation to the set of all matches, which can also eliminate outliers and obtain satisfactory matches.

To achieve more accurate motion compensation, we propose a module based on AdaLAM [39], which combines local spatial consistency and global geometric verification, and then use RANSAC [29] to obtain the homography matrix. Experiments show that our global motion compensation module outperforms other feature compensation schemes.

III. Methods

A. Pipline

We follow the methodology of TBD MOT system. Given a video consisting of consecutive frames, the objects are detected frame by frame, and predict the position of the objects in the next frame. The same object is associated with forming a trajectory, and the newly detected objects are initialized to new trajectories. A detection box is described by $d = (f_{id}, cl, h_x, h_y, w_x, w_y)$, where f_{id} denotes the ID of the frame; cl denotes the class information of the detection; (h_x, h_y, w_x, w_y) denotes the position of the detection box. To achieve more accurate multi-object tracking for UAV videos, we propose an algorithm named DC-MOT. DC-MOT includes a hybrid deblurring module, a global motion compensation module and Yolov5 [20] based DeepSort module. The framework is shown in Fig.3. In the hybrid deblurring module, the input video sequence is deblurred by CDVD algorithm, and we design an encoder to select deblurred frames for subsequent tracking. The Yolov5-based detector is then used to detect objects in each frame. Ahead of Kalman prediction, we integrate a global motion compensation module based on AdaLAM, which aims to find the best matching feature points of adjacent frames, and then a RANSAC module is added to obtain the homography matrix. The position of existing objects can be processed by cascading matching based on DeepSort [2] to obtain new objects and trajectories of the current frame. We proceed to describe the details of the DC-MOT core modules are described below.

B. Hybrid deblurring Model

UAV motion is caused by its own rotation, forward or backward motion. The fast forward and reverse movement of the UAV is likely to cause blurring in consecutive frames, but there also exist frames where both the UAV and the objects move slowly, and most of these frames are very clear. Simply applying the deblurring module to each frame before deblurring has limited improvement in tracking results (see Fig.2).

We propose a hybrid deblurring module with an encoder (Fig.4) to obtain a hybrid deblurring sequence. Frames with a high PSNR [41] value are encoded as 1 in the initial sequence, and other frames are encoded as 0. For frames with their PSNR values lower than $MOT\_PSNR(j)$(Eq.1), they are considered as blurred and require to restore clarity based on CDVD. These frames are encoded as 1 in the deblurring sequence, and other frames are encoded as 0 in the deblurring sequence. The final hybrid deblurring sequence is composed of the initial sequence and the frame encoded as 1 in the deblurring sequence. By this method, all frames are considered in the new hybrid deblurring sequence, and the advantages of deblurring algorithm can be maximized.

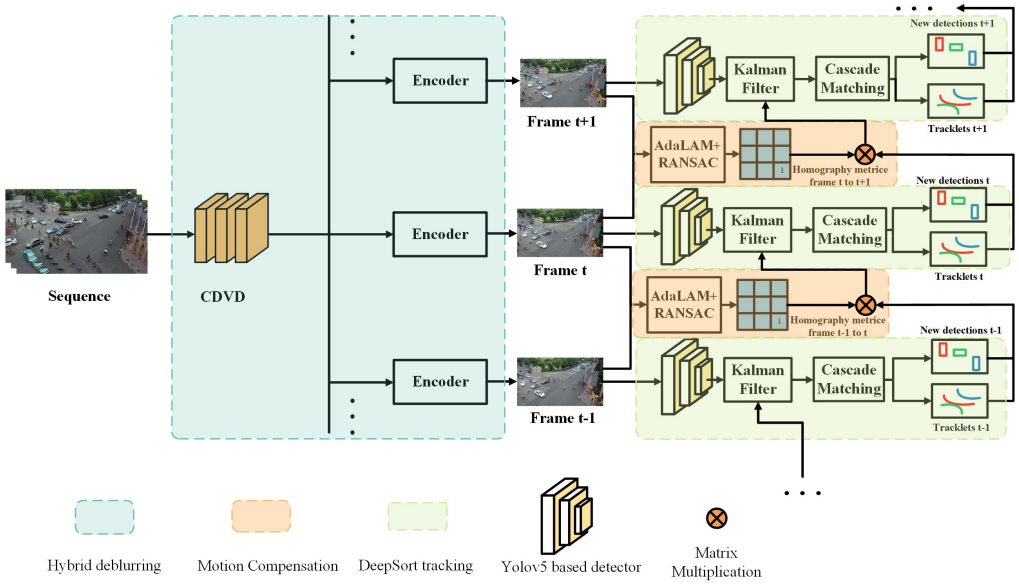

Fig. 3: The framework of DC-MOT, following TBD workflow, has three main components: (1) Hybrid deblurring module: an encoder is used to find suitable frames in the initial sequence and the deblurring sequence to form a hybrid deblurring sequence, which can not only repair blurred frames, but also ensure spatio-temporal consistency. (2) Global Motion Compensation: Feature points are extracted based on AdaLAM, and the homography transformation matrix of adjacent frames is obtained by RANSAC to achieve global motion compensation. (3) TBD tracking: The output of Kalman filter [40] is cascaded matching based on DeepSort [2] to obtain new objects and trajectories of the current frame.

We set PSNR=MOT_PSNR as the threshold for each sequence. Frames with value below this threshold are selected for deblurring, and then input to the tracking algorithm. Frames with value above this threshold are considered clear. Through experiments, the selection of fixed PSNR as a threshold is not well performed. This is because GTs for various sequences is different. For sequences with larger GTs, its PSNR stays lower at the same level of motion blur. Therefore, it is unreasonable to define a uniform threshold to determine whether the frames are clear or not. To this end, we design a new threshold, MOT_PSNR, taking into account the GTs of different sequences, to generate adaptive thresholds for different sequences:

$$MOT\_PSNR(j) = a - b \times \frac{GT(j)}{\sum\limits_{1 \le i \le n} GT(i)} \tag{1}$$

where a and b are constant parameters; $GT(j)$ denotes the GTs of the j_{th} sequence and $\sum\limits_{1 \le i \le n} GT(i)$ denotes the GTs of all sequences in the data set. For frames in each sequence with their PSNR values higher than $MOT\_PSNR(j)$, deblurring is not required to prevent unnecessary pixel information.

C. Global Motion Compensation

Different from the fixed-camera video datasets, for UAV video, camera motion is highly dynamic, and the traditional linear motion assumption in MOT can hardly be applied in this scene. The goal of global motion compensation is to align the features of adjacent frames, and decouple the motion of objects, so that the motion of objects closely compiles with linearity in adjacent frames, so as to better predict the position of objects in the current frame.

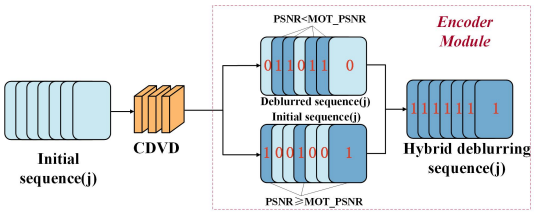

Fig. 4: Hybrid deblurring module, MOT_PSNR threshold of each sequence is used to encode each frame, and appropriate frames are selected by an encoder to constitute a new sequence for deblurring.

Based on this, the problem of global motion compensation can be reformulated into image feature matching, and the transformation matrix can be calculated by selecting appropriate pairs of feature points. We use AdaLAM [39] for feature matching, which can comprehensively select feature points. Since AdaLAM performs local and global screening, feature points satisfy local and global geometric constraints, making matching more accurate:

$$\Phi(t-1, t) = AdaLAM(f_{t-1}, f_t) \tag{2}$$

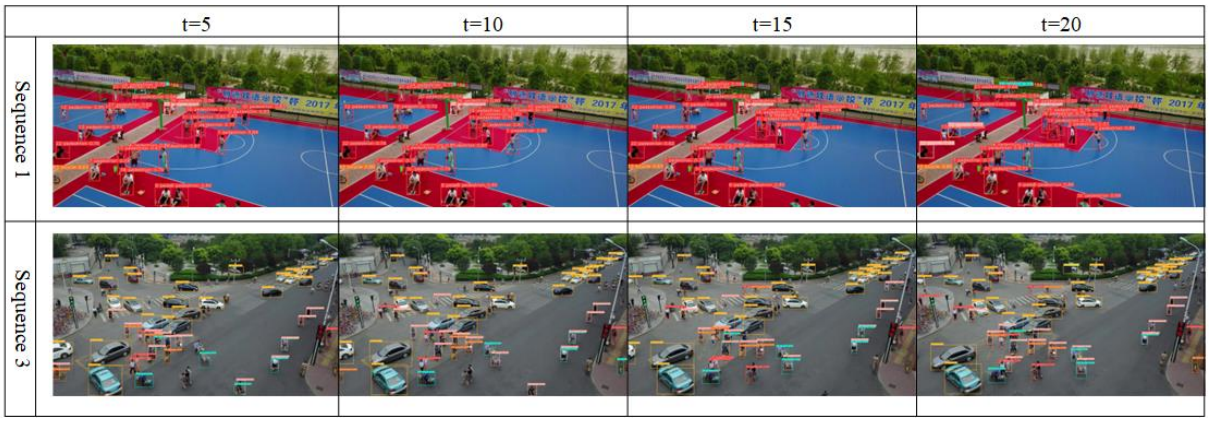

Fig. 5: Visualization of our detection and tracking results on two sequences of the VisDrone2019-MOT-val dataset.

where f_{t-1} and f_t denote frame $t-1$ and frame t, respectively, $\Phi()$ denotes the set of matching point pairs. After the matching, we use the feature points generated by AdaLAM as the input to RANSAC [29], and set appropriate hyperparameters to obtain the feature points, and calculate the homography matrix of adjacent frames. Finally, the global motion compensation result is obtained by the homography matrix:

$$\boldsymbol{GMC(t-1,t) = Homography(\Phi(t-1,t))} \times (x_t, y_t, 1)^T \tag{3}$$

where $\boldsymbol{Homography(t-1,t)}$ denotes calculating the homography matrix from frame $t-1$ to frame t by RANSAC method; $\boldsymbol{GMC(t-1,t)}$ denotes the global motion compensation on frame t; $(x_t, y_t) = Detector(f_t)$ denotes the coordinate of the object detected in frame t, which is transformed from the position of the detection box. With the homography matrix, the new objects and trajectories coordinates detected in the previous frame become the previous state estimation of the Kalman filter [40] after the matrix multiplication with the homography matrix, which achieves global motion compensation and the Kalman filter can predict with a near-linear motion condition.

D. TBD Multiple Object Tracking

DC-MOT uses Yolov5 [20] as the detector and DeepSort [2] as the tracker. Yolov5 adopts a multi-level feature fusion strategy and uses different heads to predict and classify objects of different scales, which is advantageous to detect objects accurately in UAV videos. DeepSort uses cascade matching, which not only takes into account the location information of the tracked objects, but can also associate the same object with different frames through the feature information.

As mentioned above, we utilize a general TBD framework, which indicates that our proposed hybrid deblurring and global motion compensation modules can be generalized and applied to other MOT algorithms.

IV. Experiments Setup

Datasets. We evaluate our approach on VisDrone2019-MOT. The dataset used for training the detector is VisDrone2019-det. In the MOT, we classified objects simultaneously and tracked objects of all categories (pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, motor). In order to make a fair comparison with previous work, we calculated the evaluation index on VisDrone2019-MOT-val.

Evaluation Metrics. We use standard CLEAR MOT metrics [18] and $IDF1$ metrics [43] for evaluation. In CLEAR MOT, MOTA is considered one of the most important metrics, but it only considers the counts that the tracker makes wrong decisions. In more cases, the number of frames to track the target can also indicate the tracking quality of the tracker, so we proposed $IDF1$. $IDF1$ focuses on the quality of the current tracking object. Combining the two metrics can better evaluate the performance of a tracker.

Implementation Details. We train the detector based on Yolov5l model [20]. CDVD [26] is used for deblurring, and the deblurring dataset [22] was used to train the CDVD and obtain the training model. Through experimental verification, we select the parameters $a = 100$ and $b = 305$ in Eq.1. In global motion compensation, since feature points are randomly selected for subsequent extraction by RANSAC, the results of each experiment may have slight differences. The computing resource for simulation is two NVIDIA Quadro RTX6000 GPUs. The trajectory tracking method is DeepSort with OSNet, and the hardware environment of this experiment is characterized by Intel Core I5 CPU 2.4GHz.

Evaluating Multi-Object Tracking. We summarize the results of Visdrone-MOT, listed in Tab.I and marked the **best**. Across most metrics (e.g., primary metrics MOTA and IDF1), we achieved the best performance on the Visdrone-MOT dataset. Since the $MOTA$ metric emphasizes more on the detection, our high $MOTA$ performance shows that our hybrid deblurring model improves the detector performance. Furthermore, our $IDF1$ performance shows that our method is able to decouple the objects motion from the UAV motion

TABLE I: Comparative results of the algorithms on the VisDrone-MOT dataset

Method	MOTA(%)↑	MOTP(%)↑	IDF1(%)↑	FP↓	FN↓	IDs↓
SORT [1]	18.1%	65.1%	32.2%	104453	78467	3342
Flow-Tracker [42]	26.4%	**78.1%**	41.9%	**9987**	43766	-
HDHNet(CNN) [14]	27.6%	74.3%	35.2%	153487	48980	2489
IOUT [16]	28.1%	74.7%	38.9%	126549	36158	2393
GOG [15]	28.7%	76.1%	36.4%	144657	**17706**	1387
DeepSort [2]	32.4%	75.9%	45.1%	12829	65797	1153
HDHNet(CascademaskRCNN) [14]	32.5%	75.2%	40.9%	79788	39743	**1042**
HDHNet(HTC) [14]	32.9%	76.9%	42.3%	80454	35686	1056
DC-MOT(ours)	**33.5%**	76.1%	**45.4%**	12594	64856	1139

through global motion compensation, making the objects less likely to be lost. The general Yolov5 detector and DeepSort can reflect that our proposed method has great help in improving the performance of MOT in UAV videos, and can be ported to other MOT algorithms. To visually verify our results, we also provide qualitative results of our methods in Fig.5.

V. Ablation Studies

We conducted ablation studies on VisDrone-MOT-val, with different parameters and feature matching and deblurring methods.

Effect of training size. Image size training for the detector plays a role in tracking performance. Although larger size is good for detecting normal size objects, it is likely to produce a large FNs on small objects. However, there are more small objects in UAV videos than in other datasets. As can be seen from Tab. II, the best performance is achieved when the training size is 1920 pixels. A global motion compensation module with the same hyperparameter has been adopted for all experiments.

TABLE II: Ablation study of training size

imgsz(pixels)	MOTA(%)↑	MOTP(%)↑	IDF1(%)↑	IDs↓
640	27.09%	75.01%	42.98%	1283
1280	30.64%	75.02%	**46.57%**	1613
1920	**32.94%**	76.15%	44.44%	1205
2560	31.69%	**76.45%**	41.28%	**1002**

Selection of hyperparameter in global motion compensation. ϵ denotes the tolerable error between the point of the original image after transformation and the corresponding point on the target image. ϵ ranges from 1 to 10. It is an important hyperparameter in global motion compensation. Comprehensive comparison (Tab.III) shows that global motion compensation achieves the best performance when $\epsilon = 3$.

TABLE III: Ablation study of hyperparameter in global motion compensation

ϵ	MOTA(%)↑	MOTP(%)↑	IDF1(%)↑	IDs↓
1.0	32.92%	76.16%	43.94%	1200
3.0	**33.24%**	**76.18%**	**45.56%**	**1183**
5.0	32.94%	76.15%	44.44%	1205
7.0	32.84%	76.13%	44.40%	1216

Different feature matching methods. The quality of the feature matching will affect the tracking effect. We compared the traditional methods, ORB [28] and ECC [3] used in previous work, with our proposed method, and found that our method performed best (Tab.IV).

TABLE IV: Ablation study of different feature matching methods

Method	MOTA(%)↑	MOTP(%)↑	IDF1(%)↑	IDs↓
DeepSort	32.46%	75.91%	45.41%	**1153**
DeepSort+ORB	32.35%	75.88%	45.28%	1167
DeepSort+ECC	32.87%	76.05%	44.51%	1166
DeepSort+AdaLAM	**33.24%**	**76.18%**	**45.56%**	1183

Different deblurring methods. The tracking quality is compared with no deblurring module, deblurring all frames (DeepSort+CDVD), and our hybrid deblurring module (Tab.V). It is observed that hybrid deblurring strategy can take maximum advantages of the deblurring and help improve tracking performance. When we select the parameters $a = 100$ and $b = 305$ in Eq.1, the proportion of deblurred frames in the dataset is 70.7%.

TABLE V: Ablation study of different deblurring methods

Method	MOTA(%)↑	MOTP(%)↑	IDF1(%)↑	IDs↓
DeepSort	32.46%	75.91%	45.41%	1153
DeepSort+CDVD	32.44%	75.71%	45.40%	1159
DC-MOT	**33.47%**	**76.14%**	**45.44%**	**1139**

VI. Conclusions

We propose a new MOT system framework for UAV video. To alleviate the motion blur caused by the high-speed movement of the UAV and the targets, we propose a hybrid deblurring module to accurately locate and deblur the selected frames while ensuring spatio-temporal consistency, laying a foundation for accurate detection by the detector. To mitigate the influence of UAV motion on trajectory prediction, we propose a global motion compensation method based on feature matching, which improves the accuracy of object prediction and association. Experiments show that the proposed methods can improve tracking accuracy, and the final tracking results outperform the state-of-the-art methods on the VisDrone dataset.

REFERENCES

[1] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*, pp. 3464–3468, IEEE, 2016.

[2] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*, pp. 3645–3649, IEEE, 2017.

[3] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 941–951, 2019.

[4] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3539–3548, 2017.

[5] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, "Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism," in *Proceedings of the IEEE international conference on computer vision*, pp. 4836–4845, 2017.

[6] S. Pang, D. Morris, and H. Radha, "3d multi-object tracking using random finite set-based multiple measurement models filtering (rfs-m 3) for autonomous vehicles," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13701–13707, IEEE, 2021.

[7] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3415–3424, 2017.

[8] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, "Mots: Multi-object tracking and segmentation," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 7942–7951, 2019.

[9] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *European Conference on Computer Vision*, pp. 107–122, Springer, 2020.

[10] Z. Lu, V. Rathod, R. Votel, and J. Huang, "Retinatrack: Online single stage joint detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14668–14678, 2020.

[11] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *European Conference on Computer Vision*, pp. 474–490, Springer, 2020.

[12] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3069–3087, 2021.

[13] Y. Wang, K. Kitani, and X. Weng, "Joint object detection and multi-object tracking with graph neural networks," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13708–13715, IEEE, 2021.

[14] W. Huang, X. Zhou, M. Dong, and H. Xu, "Multiple objects tracking in the uav system based on hierarchical deep high-resolution network," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13911–13929, 2021.

[15] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *CVPR 2011*, pp. 1201–1208, IEEE, 2011.

[16] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information," in *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pp. 1–6, IEEE, 2017.

[17] Y. Du, Y. Song, B. Yang, and Y. Zhao, "Strongsort: Make deepsort great again," *arXiv preprint arXiv:2202.13514*, 2022.

[18] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[20] G. Jocher, "ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements." https://github.com/ultralytics/yolov5, Oct. 2020.

[21] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

[22] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1279–1288, 2017.

[23] T. H. Kim, M. S. Sajjadi, M. Hirsch, and B. Scholkopf, "Spatio-temporal transformer network for video restoration," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 106–122, 2018.

[24] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[25] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "Tdan: Temporally-deformable alignment network for video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3360–3369, 2020.

[26] J. Pan, H. Bai, and J. Tang, "Cascaded deep video deblurring using temporal sharpness prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3043–3051, 2020.

[27] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[28] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*, pp. 2564–2571, Ieee, 2011.

[29] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[30] D. P. Capel, "An effective bail-out test for ransac consensus scoring.," in *BMVC*, vol. 1, p. 2, Citeseer, 2005.

[31] O. Chum and J. Matas, "Matching with prosac-progressive sample consensus," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 220–226, IEEE, 2005.

[32] O. Chum and J. Matas, "Optimal randomized ransac," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1472–1482, 2008.

[33] T. Sattler, B. Leibe, and L. Kobbelt, "Scramsac: Improving ransac's efficiency with a spatial consistency filter," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 2090–2097, IEEE, 2009.

[34] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, and M.-M. Cheng, "Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4181–4190, 2017.

[35] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," *International Journal of Computer Vision*, vol. 127, no. 5, pp. 512–531, 2019.

[36] K. Ni, H. Jin, and F. Dellaert, "Groupsac: Efficient consensus in the presence of groupings," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 2193–2200, IEEE, 2009.

[37] D. Barath and J. Matas, "Graph-cut ransac," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6733–6741, 2018.

[38] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, "Usac: A universal framework for random sample consensus," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 2022–2038, 2012.

[39] L. Cavalli, V. Larsson, M. R. Oswald, T. Sattler, and M. Pollefeys, "Adalam: Revisiting handcrafted outlier detection," *arXiv preprint arXiv:2006.04250*, 2020.

[40] G. Welch, G. Bishop, *et al.*, "An introduction to the kalman filter," 1995.

[41] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*, pp. 2366–2369, IEEE, 2010.

[42] W. Li, J. Mu, and G. Liu, "Multiple object tracking with motion and appearance cues," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[43] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European conference on computer vision*, pp. 17–35, Springer, 2016.