DroneMOT: Drone-based Multi-Object Tracking Considering Detection Difficulties and Simultaneous Moving of Drones and Objects

Peng Wang, Yongcai Wang and Deying Li

***Abstract*— Multi-object tracking (MOT) on static platforms, such as by surveillance cameras, has achieved significant progress, with various paradigms providing attractive performances. However, the effectiveness of traditional MOT methods is significantly reduced when it comes to dynamic platforms like drones. This decrease is attributed to the distinctive challenges in the MOT-on-drone scenario: (1) objects are generally small in the image plane, blurred, and frequently occluded, making them challenging to detect and recognize; (2) drones move and see objects from different angles, causing the unreliability of the predicted positions and feature embeddings of the objects. This paper proposes DroneMOT, which firstly proposes a Dual-domain Integrated Attention (DIA) module that considers the fast movements of drones to enhance the drone-based object detection and feature embedding for small-sized, blurred, and occluded objects. Then, an innovative Motion-Driven Association (MDA) scheme is introduced, considering the concurrent movements of both the drone and the objects. Within MDA, an Adaptive Feature Synchronization (AFS) technique is presented to update the object features seen from different angles. Additionally, a Dual Motion-based Prediction (DMP) method is employed to forecast the object positions. Finally, both the refined feature embeddings and the predicted positions are integrated to enhance the object association. Comprehensive evaluations on VisDrone2019-MOT and UAVDT datasets show that DroneMOT provides substantial performance improvements over the state-of-the-art in the domain of MOT on drones. The code will be available at https://github.com/PenK1nG/DroneMOT.**

I. INTRODUCTION

Multi-object tracking (MOT) is a critical task in computer vision, which has a wide range of applications in autonomous driving [1] and video surveillance [2]. The goal of MOT is to find the trajectories of objects through continuous observations by cameras. MOT methods can be broadly categorized into two paradigms: tracking-by-detection [3]–[7] and tracking-by-regression [8]–[10]. Currently, due to the great success of deep learning-based object detection [11]–[13], tracking-by-detection methods [14] [4], which firstly detect objects in each frame and then associate the detections with the trajectories, have a leading performance in MOT.

MOT has shown impressive performance for static cameras [15]–[17]. However, when applied to drones or unmanned aerial vehicles, the performance of existing MOT methods decreases significantly [18]. This decrease in performance is attributed to the difficulty in accurately detecting objects and associating them with their trajectories. These challenges are inherent to MOT-on-drone scenarios, as illustrated in Fig. 1. At first, the elevated altitude at which drones operate often results in smaller apparent scales of the objects in the footage. Additionally, the swift movement of drones can introduce motion blur and occlusion into the video frames. Combining these factors makes it challenging to detect objects and extract meaningful feature embeddings [19] [20]. Furthermore, when the drone and the objects move simultaneously, there can be significant shifts in the pixel positions of the same object across consecutive frames. Such irregular movement might also cause objects near the camera's edge to appear discontinuously. With the drone in motion, the same object can be viewed from multiple angles, leading to inconsistent features. Therefore, data association based on the coherence of target pixel positions and the consistency of target features tends to perform poorly under the dynamic conditions of drones.

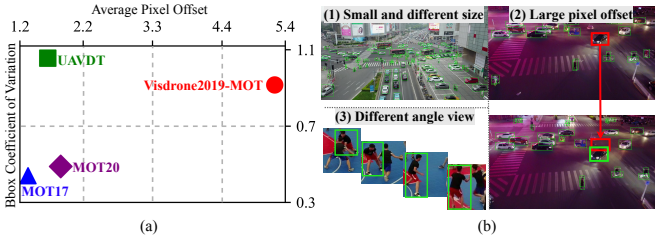

Fig. 1. **Challenges of MOT on drones.** (a) comparisons between conventional MOT datasets(MOT17/20) and drone-based MOT datasets(Visdrone2019-MOT and UAVDT). The x-axis represents the average change in the pixel position of the same object in adjacent frames. In contrast, the y-axis represents the coefficient of variation (variance/mean) of the object's bbox size. (b) Visualization of these challenges, encompassing small-scale objects, large pixel offsets, and varying angle views.

Given the significant importance of drone-based object detection and tracking in various applications [19] [21], several methods have emerged. One dominant approach adheres to the tracking-by-detection paradigm, emphasizing enhancements in drone-based object detection and feature embedding. For instance, Wang et al. [22] modified YOLOv3 [23] to utilize three different resolution feature maps for vehicle detection and tracking in UAV videos. UAVMOT [24] leverages the correlation layer between two adjacent frames to reinforce ID embedding based on features. Some methods have been reported to address association issues. Zhang et al. [25] employ the TNT network [26] for detection and directly calculate the Semi-Direct Visual Odometry by Multi-View Stereo for data association. Other studies [27]–[29] utilize

All authors are with the Department of Computer Science, School of Information, Renmin University of China, Beijing 100872, China. Corresponding author: Yongcai Wang. Email {peng.wang, ycw, deyingli}@ruc.edu.cn

Dr. Li is supported in part by the National Natural Science Foundation of China Grant No. 12071478. Dr. Wang is supported in part by the National Natural Science Foundation of China Grant No. 61972404, Public Computing Cloud, Renmin University of China, and the Blockchain Lab, School of Information, Renmin University of China.

RTK, IMU, or GPS to directly compute the drone's poses, aiming to boost the performance of drone MOT. However, these methods require additional equipment.

In this work, we rely solely on the image information and propose **DroneMOT**, which not only enhances object detection and feature embedding but also considers simultaneous motions of the drone and the objects to improve the **robustness** of the data-association. In particular, in the detection module, we introduce a *Dual-Domain Integrated Attention (DIA)*, which integrates Spatial Attention and Heatmap-Guided Temporal Attention to achieve more accurate and comprehensive detections with embedding. In the data-association module, we propose an innovative *Motion-Driven Association (MDA)* scheme considering the simultaneous movement of the drone and the objects. In MDA, we first present a *Adaptive Feature Synchronization (AFS)* module that refines trajectory appearance by dynamically adjusting the feature weights based on the detection scores and preserving key historical features from different angles of the same object. Then, we introduce the *Dual Motion-based Prediction (DMP)* module. Instead of solely focusing on the target motion, DMP also takes the drone motion into account. We decompose the drone's motion into three primary components: hovering, translation, and rotation. Combining the motion of the drone and the motions of the objects, the trajectory's pixel position in the subsequent frame is more accurately predicted. The key contributions are summarized as follows:

- Dual-Domain Integrated Attention (DIA) is proposed to enhance the detection and feature embedding of small-sized, blurred, and occluded objects in videos captured by drone.
- Motion-Driven Association (MDA) is proposed for robust data association, which includes AFS to refine the trajectory appearance and DMP to predict the object position considering the simultaneous motions of the drone and the objects.
- Extensive evaluations on the Visdrone2019-MOT [30] and UAVDT [31] datasets demonstrate that DroneMOT outperforms the state-of-the-art methods for multi-object tracking on drones.

II. Related Work

Multi-Object-Tracking on Drone. MOT algorithms are usually divided into tracking-by-detection paradigms [6], [7], [32]–[35] and tracking-by-regression paradigms [5], [9], [36]–[40]. Due to the unpredictable and irregular properties of the simultaneous movement of drones and objects, MOT for drones [25], [41] typically adopts the tracking-by-detection paradigm. This approach first uses a network to detect objects in each frame and then associates these detections with the stored trajectories. PAS Tracker [42] uses an additional ReID network to obtain object features and combines position, appearance, and size information jointly to make full use of the object representations. UAVMOT [24] utilizes the correlation layer [43]–[46] between two adjacent frames to strengthen the embedding features, and develops an adaptive motion filter to complete the object ID association accurately. GLOA [47] proposes a global-local awareness detector to extract scale variance feature information from the input frames for the frequent occluded objects. FOLT [48] adopts a light-weight optical flow extractor to extract object detection features and motion features at a minimum cost to improve the detection of small objects. Although some research has begun to focus on data association, the drone-based MOT methods are still focused on building powerful detectors. In this work, we present an integrated framework tailored not only for the enhanced detection of small, blurred, and occluded objects but also for data-association strategies specifically designed to accommodate the motion of drones.

Data-Association. Early MOT approaches, such as SORT [7], [32], adopt the data-association method. These methods employ a Kalman filter [49]–[52] to predict an object's trajectory position in the subsequent frame, serving as the motion model. Concurrently, a network [53] is utilized to obtain the object feature embedding, acting as the appearance model. By integrating both the motion and the appearance models, data-association is achieved using the Hungarian algorithm [54]. BoT-SORT [55] utilizes an enhanced Kalman filter and compensates for camera motion to achieve a more accurate motion model. OC-SORT [56] uses object observations to compute a virtual trajectory to correct the error accumulation of the Kalman filter during the occlusion period. Meanwhile, some researchers have focused on the appearance model to get effective and comprehensive features. CorrTracker [57] uses the correlation layer [58] to calculate the spatio-temporal correlation of features between adjacent frames, thereby obtaining more accurate object feature embedding. GHOST [59] analyzes MOT failure cases and proposes a combination method of proxy appearance features with a simple motion model, leading to strong tracking results. In addition, ByteTrack [60] employs a multi-level data-association method. The trajectories are first matched with the detections that have high detection scores, and the remaining trajectories are matched with the detections that have low detection scores. In this work, we adopt these advanced data-association methods and further consider the motion patterns of drones to specifically design motion and appearance models for data association on drones.

III. METHOD

DroneMOT is primarily split into two modules: the network module (III-A) for detection and feature embedding, and the data-association module (III-B) based on the result of the network module. The image $I_t \in \mathbb{R}^{W\times H\times 3}$ captured by the moving drone at the t-th frame is fed into the network along with the previous frame image I_{t-1}. The results of the network module, represented by $\mathcal{O}_t = \{o_1, o_2, \cdots, o_i, \cdots, o_M\}$ consist of M detections where $o_i = (b_i, s_i, f_i)$. Here, b_i represents the bounding box (x, y, w, h), s_i is the detection score, and f_i is the feature embedding vectors. The data association module takes the detections $\mathcal{O}_t$ and all N stored trajectories of the objects $\mathcal{T}_{t-1} = \{T_1, T_2, \cdots, T_j, \cdots, T_N\}$ as inputs,

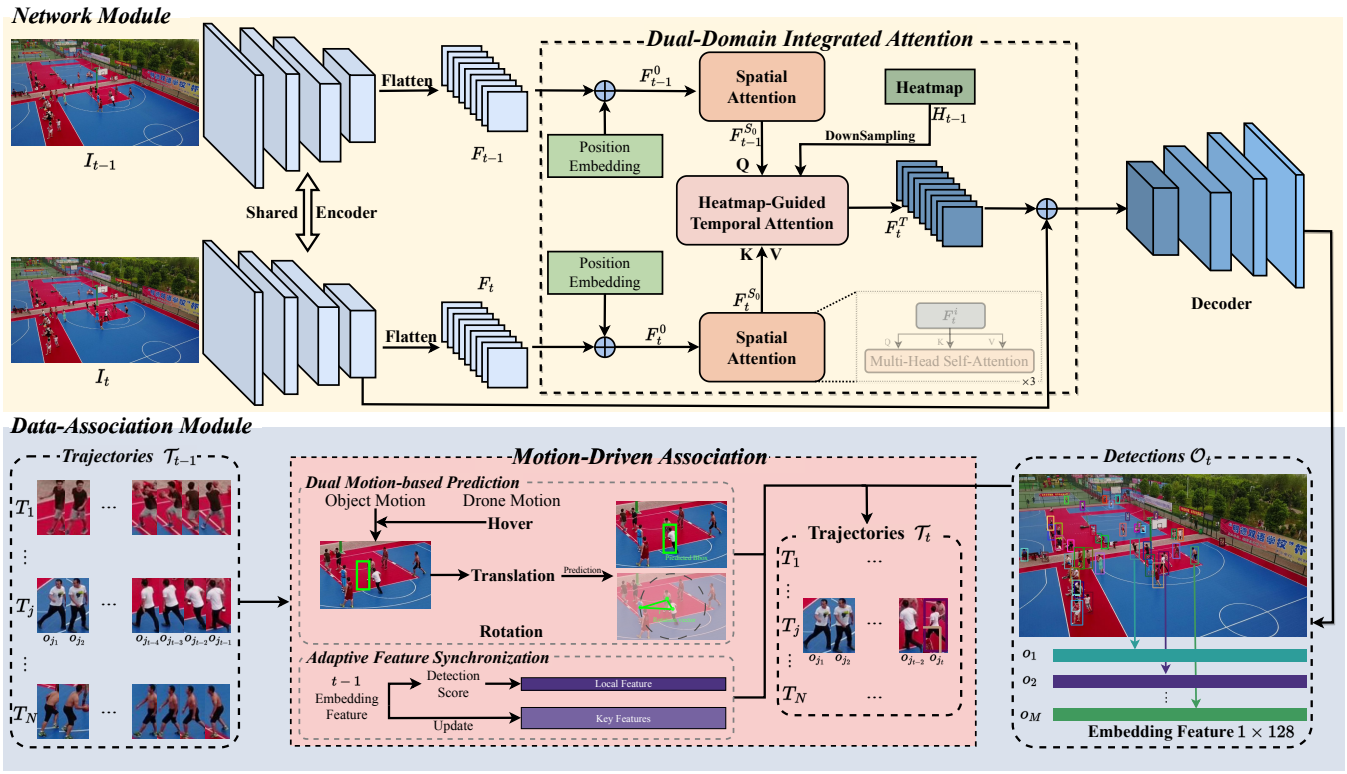

Fig. 2. **The overall architecture of DroneMOT.** It primarily consists of two modules: the network module (III-A) for online detection and feature embedding and the data-association module (III-B) to associate detections with stored trajectories of objects.

where $T_j = \{o_{j_1}, o_{j_3}, \cdots, o_{j_{t-1}}\}$, and $o_{j_{t-1}}$ represents the detection associated with the trajectory j in the $t-1$-th frame. The goal of the data association module is to match each detection with a trajectory, treat the unmatched detections as the new trajectories, and ultimately produce the final tracking results $\mathcal{T}_t$. An overview of the proposed **DroneMOT** is presented in Fig. 2.

A. Network Module

In the network module, we utilize the DLA34 [61] network as the backbone, which is an encoder-decoder architecture. The encoder uses shared convolution layers to extract local features from images $\{I_{t-1}, I_t\}$. After flattening the local features, these features denoted as $\{F_{t-1}, F_t\}$ serve as inputs to the **Dual-Domain Integrated Attention (DIA)** module. DIA module consists of two parts: *Spatial Attention* and *Heatmap-Guided Temporal Attention*.

Spatial Attention. The Spatial Attention layer aims to augment object features with spatial positional information and the relationships between objects, enabling the network to distinguish different small-scale objects easily. The effectiveness of the Spatial Attention is illustrated in Fig. 4(a). To achieve this goal, we firstly add the flattened local features F_{t-1}, F_t with a 2D extension of the standard position encoding [62] to make the features cognizant of their global positions within the entire 2D image feature space:

$$F_t^0 = F_t + \text{PosEncod}. \tag{1}$$

Then we adopt three multi-head self-attention layers separately to enhance the spatial relationships and object interactions within the feature maps, thereby crafting a more spatially aware representation feature $F_{t-1}^{S_0}, F_t^{S_0}$:

$$\begin{aligned} F_t^{i+1} &= \text{Norm}(F_t^i + \text{MultiHead}(F_t^i, F_t^i, F_t^i)), i = 0, 1, \\ F_t^{S_0} &= \text{Norm}(F_t^2 + \text{MultiHead}(F_t^2, F_t^2, F_t^2)). \end{aligned} \tag{2}$$

where t can be replaced by $t-1$, "MultiHead" refers to the multi-head attention [63] following the query, key, and value, and "Norm" represents the layer normalization.

Heatmap-Guided Temporal Attention. The temporal attention layer focuses on the evolution of features for the same object over successive time steps. In aerial tracking, the presence of motion blur or occlusion often leads to ineffective temporal contexts. To filter out the regions without objects and to heighten the feature's focus on the objects affected by motion blur and occlusion, we propose to use the heatmap of the $t-1$-th frame as the filter's attention. As illustrated in Fig. 4(b)(c), this heatmap-guided filter leads to a more context-aware interpretation of the blurred and occluded objects detected from the visual sequence.

Specifically, given the adjacent spatial enhanced feature $F_{t-1}^{S_0}, F_t^{S_0}$, and the heatmap H_{t-1} obtained from the $t-1$-th frame, we acquire the output feature $F_t^{S_2}$ of the stacked multi-head attention layer in the t-th frame:

$$\begin{aligned} F_t^{S_1} &= \text{Norm}(F_t^{S_0} + \text{MultiHead}(F_{t-1}^{S_0}, F_t^{S_0}, F_t^{S_0})), \\ F_t^{S_2} &= \text{Norm}(F_t^{S_1} + \text{MultiHead}(F_t^{S_1}, F_t^{S_1}, F_t^{S_1})). \end{aligned} \tag{3}$$

Then as presented in Fig. 3, the feature representation $\hat{F}_{t-1}^{S_0}$ is generated by concatenating the heatmap with the resized convolutional features, followed by a 1×1 convolution. A heatmap-guided weight W_{t-1} is derived via Global Average Pooling (GAP) and a feed-forward network (FFN). This

weight is then multiplied with the feature $F_t^{S_2}$, creating a refined feature representation F_t^f guided by the heatmap. Finally, F_t^T is obtained by the multi-head attention:

$$\begin{aligned} \hat{F}_{t-1}^{S_0} &= \mathcal{F}(\text{Cat}(H_{t-1}, \text{Resize}(F_{t-1}^{S_0}))), \\ W_{t-1} &= \text{FFN}(\text{GAP}(\hat{F}_{t-1}^{S_0})), \\ F_t^f &= F_t^{S_2} + F_t^{S_2} \times W_{t-1}, \\ F_t^T &= \text{Norm}(F_t^f + \text{MultiHead}(F_t^f, F_t^f, F_t^f)). \end{aligned} \tag{4}$$

where $\mathcal{F}$ represents a convolution layer, and FFN means a feed-forward network.

The local feature F_t at the t-th frame, combined with the results F_t^T from the DIA module, is utilized as the input to the decoder, resulting in the Detection Head. Following [34], the detection head applies successive convolutional operations to obtain the heatmap H_t of the objects, which can be used as the input to the network of the t+1-th frame, along with the corresponding width, height, and feature embedding. These form the object detection results and their feature embeddings, i.e., $\mathcal{O}_t = \{o_1, o_2, \cdots, o_M\}$ for the t-th frame.

B. Motion-Driven Association

Motion-Driven Association (MDA) takes detections $\mathcal{O}_t$ in the t-th frame and trajectories $\mathcal{T}_{t-1}$ from the $t-1$-th frame as inputs. Considering the simultaneous movements of both the drone and the objects, MDA consists of two primary components: (1) Adaptive Feature Synchronization (AFS) and (2) Dual Motion-based Prediction(DMP). Finally, both the refined feature embeddings and the precise predicted positions are integrated to enhance the object association to get the trajectory $\mathcal{T}_t$ for the t-th frame.

Adaptive Feature Synchronization. In previous work [32], [34], the appearance feature vectors of a trajectory only consider the local feature, which is updated by an Exponential Moving Average (EMA) of the current feature vector and the historical feature vector. EMA typically requires a fixed weight coefficient α to control the contribution of the historical feature vectors.

As an appearance model for data-association, AFS categorizes the features of trajectories into local and key features. To obtain more accurate local features, we dynamically adjust the weight coefficient α based on the detection score of the current frame. In addition, to address scenarios with sudden changes in target angles or extended occlusions, we preserve a subset of historical features as key features.

For the local feature, we use the detection score s_t as the proxy to dynamically adjust the weight coefficient α in EMA, which is defined as

$$\begin{aligned} f_t^{local} &= \alpha f_{t-1}^{local} + (1-\alpha) f_t, \\ \alpha &= \alpha_f + (1-\alpha_f) e^{(\theta - s_t)}. \end{aligned} \tag{5}$$

where α_f is a fixed value, usually set to 0.9, s_t represents the object detection score, and θ is a detection confidence threshold to filter out noisy detections. For high-confidence detections, α approaches α_f, increasing its impact on the local feature.

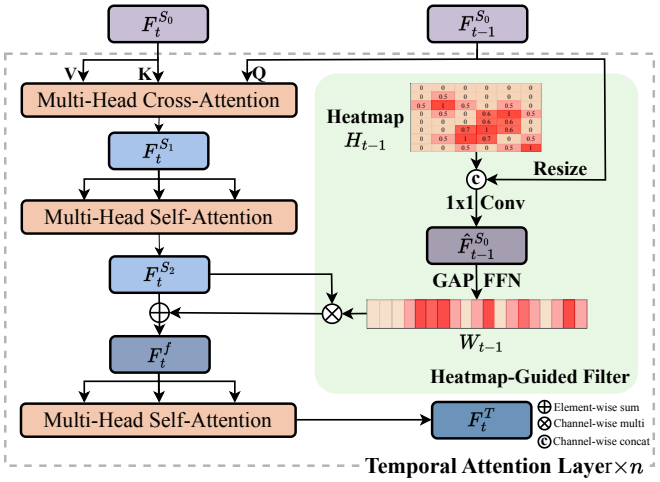

Fig. 3. Structure of Heatmap-Guided Temporal Attention.

As for the key features, AFS retains a portion of historical features for every trajectory. The key features are typically updated by employing the least recently used algorithm to store the ten key features.

Dual Motion-based Prediction. Unlike existing methods such as [34], [55] that only consider the movement of objects, DMP also incorporates the drone's motion. We classify the drone's movements into three fundamental types: hovering, translation, and rotation. When the drone hovers, the camera can be approximated as a fixed camera. We can utilize the Kalman filter [49] for fixed cameras to predict the trajectories $\mathcal{T}_{t-1}$'s position in the t-th frame. When the drone undergoes translation or rotation, we compensate separately for the movements of the drone to improve the object-trajectory association.

Regarding the translation, following [55], we calculate the affine matrix between two frames and subsequently determine the position of the trajectories after the affine transformation. This method, termed Camera Motion Compensation, effectively compensates for the impact of translation of the drone on MOT. For rotation, we observed that the shape of the triangle formed by the object and its surrounding objects in adjacent frames is almost congruent. Therefore, the rotation vector of an object can be effectively captured using the intrinsic features of a triangle: $v_t = [\alpha_t, \beta_t, l_t]$ for an object in the t-th frame. Here, α, β denote the two smallest angles of the triangle, while l represents the side length opposite the largest angle. The triangle is formed by the object, the farthest object, and the nearest object within a radius of R pixels.

Finally, by integrating the drone's hovering and translation with the objects' movement, we can predict the trajectories' positions in the t-th frame. This integration enables us to compute the IOU cost matrix I_C between the predicted object positions (bounding box with positions) and the detected object positions. Moreover, we evaluate the cosine similarity between the rotation vector of the trajectories and that of the detections, resulting in the rotation cost matrix R_C. On the other hand, the AFS module efficiently calculates the appearance cost matrix A_C based on the minimal cosine

TABLE I

QUANTITATIVE COMPARISONS BETWEEN DRONEMOT AND OTHER METHODS ON VISDRONE2019-MOT TEST-DEV AND UAVDT TEST SET. METHODS IN BLUE BLOCK ARE MOT METHODS SPECIFICALLY FOR THE DRONE. THE BEST RESULTS ARE MARKED IN **BOLD**.

Dataset	Method	Pub&Year	IDF1↑	MOTA↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDs↓
VisDrone2019-MOT	SiamMOT [64]	CVPR2021	48.3	31.9	73.5	-	-	24123	142303	862
	MOTR [38]	ECCV2022	41.4	22.8	72.8	272	825	28407	147937	959
	ByteTrack [60]	ECCV2022	40.8	25.1	72.4	446	1099	34044	194984	1590
	OC-SORT [56]	CVPR2023	50.4	39.6	73.3	-	-	**14631**	123513	986
	STDFormer [65]	TCSVT2023	57.1	**45.9**	77.9	684	538	21288	101506	1440
	UAVMOT [24]	CVPR2022	51	36.1	74.2	520	574	27983	115925	2775
	FOLT [48]	MM2023	56.9	42.1	**77.6**	-	-	24105	107630	**800**
	GLOA [47]	J-STARS2023	46.2	39.1	76.1	581	824	18715	158043	4426
	DroneMOT	Ours	**58.6**	43.7	71.4	**689**	**397**	41998	**86177**	1112
UATDT	DeepSORT [32]	ICIP2017	58.2	40.7	73.2	595	338	44868	155290	2061
	SiamMOT [64]	CVPR2021	61.4	39.4	76.2	-	-	46903	176164	190
	ByteTrack [60]	ECCV2022	59.1	41.6	79.2	-	-	**28819**	189197	296
	OC-SORT [56]	CVPR2023	64.9	47.5	74.8	-	-	47681	148378	288
	UAVMOT [24]	CVPR2022	67.3	46.4	72.7	624	221	66352	115940	456
	FOLT [48]	MM2023	68.3	48.5	**80.1**	-	-	36429	155696	338
	GLOA [47]	J-STARS2023	68.9	49.6	79.8	626	220	55822	115567	433
	DroneMOT	Ours	**69.6**	**50.1**	74.5	**638**	**178**	57411	**112548**	**129**

value discerned between the feature of the detections and both the local feature and the key features of the trajectories. Therefore, the final cost matrix is typically formulated by combining the three cost matrices, represented as:

$$C = I_C + w_a A_C + w_r R_C \tag{6}$$

By using a linear sum assignment [54], each detection can uniquely correspond to a trajectory. Unmatched targets are treated as new trajectories, yielding the trajectories $\mathcal{T}_t$ for the t-th frame.

IV. EXPERIMENTS

A. Experimental Setup

Dataset. We evaluate the proposed methods using two multi-object tracking datasets for drones: (1) VisDrone2019-MOT [30] and (2) UAVDT [31]. They are both developed for multi-category tracking using drones. The VisDrone2019-MOT dataset [30] is divided into three parts: a training set (56 sequences), a validation set (7 sequences), and a test set (33 sequences). It encompasses ten categories: pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle. The UAVDT dataset [31] is explicitly designed for vehicle object tracking. It is split into two parts: a training set and a test set, covering three categories: car, truck, and bus. The video images in this dataset offer a resolution of 1080 × 540 pixels and showcase various illumination conditions, including sunshine, fog, and rain.

Metrics. We adopt IDF1 [66], MOTA [67], and ID switching (IDs) [67] as the primary evaluation metrics to evaluate our proposed DroneMOT with other state-of-the-arts approaches. MOTA is computed based on FP, FN, and IDs, which focus more on the detection performance. And IDF1 evaluates the identity association accuracy of the tracking results.

Training Details. We train DroneMOT for 30 epochs on six NVIDIA GeForce RTX 2080ti GPUs with batch size 12. In the multiple loss functions, we modify the EQ-Loss v2 [68] to supervise the heatmap. Furthermore, L1 loss and Triplet loss [69] are separately used to deal with the width and height of the object and the object ID.

Tracking Details. At the data-association stage, we follow ByteTrack [60] to set the high detection score threshold to 0.6 and the low detection score threshold to 0.1. In Dual Motion-based Prediction, w_a, w_r in Equation. 6 are set to 0.5 and 0.1, respectively. Furthermore, R in AFS module is set to 100 pixels.

B. Comparison with the state-of-the-art methods

We compare DroneMOT with state-of-the-art (SOTA) trackers, including those specifically tailored for MOT on drones including UAVMOT [24], FOLT [48], GLOA [47] and the generic ones including SiamMOT [64], MOTR [38], ByteTrack [60], OC-SORT [56], and STDFormer [65]. The performance results on the two drone-based MOT datasets are presented in the following sections.

Visdrone2019-MOT. In this dataset, we train using all categories. However, we adhere to the official VisDrone toolkit for evaluation, which focuses on five categories: car, bus, truck, pedestrian, and van—consistent with other trackers. Our results on the VisDrone2019 test-dev set are presented in Table I. DroneMOT stands out, achieving the highest IDF1 score of 58.6%, which is a marked improvement over competing methods. This score underscores DroneMOT's effectiveness in correctly identifying and matching object identities. Furthermore, DroneMOT excels in detection capabilities, recording the lowest FN count of 86,177. Moreover, it boasts the highest MT while registering the fewest ML, emphasizing its precision and consistency in maintaining trajectory IDs.

UAVDT. The UAVDT dataset presents a more pronounced bbox variation compared to VisDrone2019-MOT, as evidenced in Fig. 1. This characteristic implies that UAVDT is more challenging in terms of both detection and embedding tasks. When evaluated on the official server, our results for the UAVDT benchmarks can be seen in Table I. DroneMOT continues to set the benchmark, achieving an unrivaled IDF1 score of 69.6% and a commendable MOTA of 50.1%. Additionally, DroneMOT outperforms by registering a minimal

TABLE II

ABALATION STUDY ON VISDRONE2019-MOT VALIDATION SET.

Baseline	DIA	MDA	MOTA(%)	IDs	IDF1(%)
✓			29.7	1509	38.3
✓	✓		33.4	1407	45.1
✓		✓	32.4	406	48.9
✓	✓	✓	34.3	218	53.4

TABLE III

ANALYSIS OF THE EFFECTIVENESS OF MDA MODULE. THE BASELINE USES THE KALMAN FILTER AND EMA TO UPDATE THE FEATURE.

Motion model	Appearance model	IDs	IDF1	IDP	IDR
-	-	1407	45.1	48.6	42.1
DMP	-	229	52.8	57.8	48.6
-	AFS	690	46.5	52.8	41.5
DMP	AFS	218	53.4	43.0	52.8

129 ID switches, underscoring its expertise in consistently preserving object identities across sequences.

C. Ablation Study

The baseline model we compared against is FairMOT [34], which uses DLA34 as its backbone and has the same loss settings as DroneMOT.

Dual-Domain Integrated Attention. The DIA module, powered by spatial attention and heatmap-guided temporal attention, significantly refines feature representation, bolstering robustness and accuracy. As evidenced in Table II, including the DIA module enhances the MOTA and IDF1 scores to 20.4% and 45.1%, respectively. Furthermore, it results in a decrease in IDs, dropping from 1509 to 1407. The proficiency of the DIA module is visually represented in Fig. 4, which underscores its effectiveness in assisting the network to recognize small-sized, blurred, or occluded objects.

Motion-Driven Association Module. The integration of the MDA module plays a pivotal role in enhancing tracking performance, as evident in Table II. Specifically, we observe improvements of 4.7% in MOTA and 10.6% in IDF1. Moreover, IDs are significantly reduced, plummeting from 1509 to 406. Delving deeper into the MDA module's components in Table III, we find that the DMP component substantially curtails ID switches, bringing them down from 1407 to 229. Further synergizing DMP with AFS elevates the IDF1 score to 53.4%, underscoring the combined strength of both components in refining tracking accuracy.

D. Visualization

To showcase the efficacy of DroneMOT, we present a tracking visualization compared to UAVDT. Particularly during drone rotations, DroneMOT consistently retains the trajectory ID of targets, ensuring no loss or mismatch of IDs, as evidenced in Fig. 5. Even under challenging foggy conditions, exemplified in Fig. 6, DroneMOT's DIA module proves instrumental in accurately identifying targets — even the minute ones obscured by fog cover as the drone ascends. These visual representations highlight how adeptly DroneMOT adapts to diverse and dynamic conditions, excelling in the MOT task on drone footage.

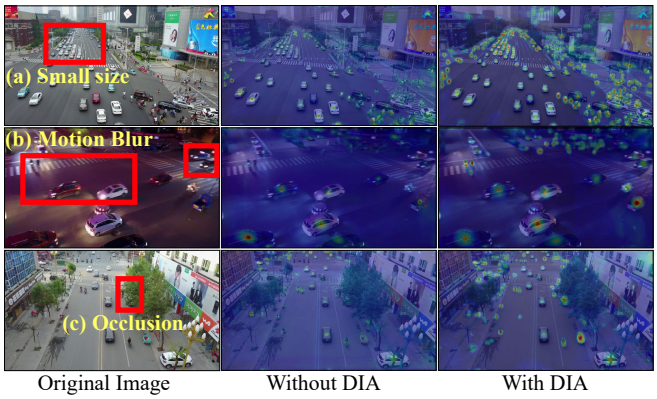

Fig. 4. Feature map comparison between without DIA and with DIA.

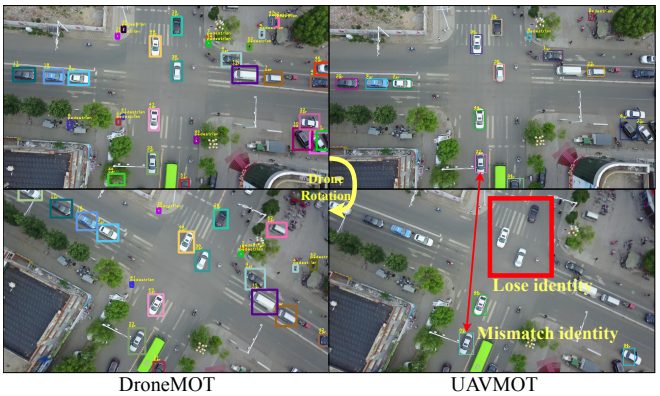

Fig. 5. Visualization of tracking results on the **Visdrone2019-MOT** dataset when the drone is rotating rapidly.

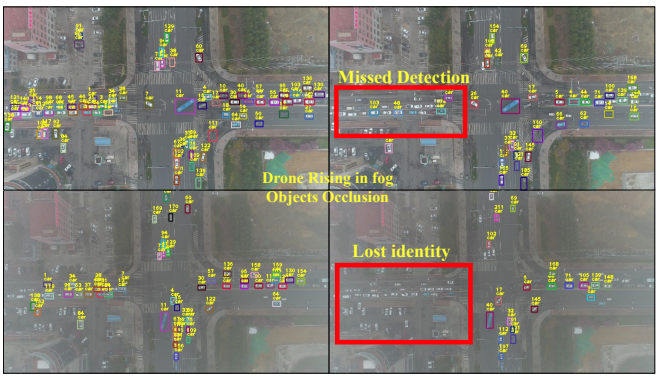

Fig. 6. Visualization of tracking results on the **UAVDT** dataset when the drone raises in foggy conditions, and the target is obscured by the fog.

V. CONCLUSIONS

In this paper, we introduced DroneMOT, a novel approach tailored specifically for the challenges presented by drone-based multiple object tracking. By integrating the proposed Dual-Domain Integrated Attention, DroneMOT excels in object detection and feature embedding, capitalizing on spatial nuances and leveraging heatmap-guided temporal insights. Moreover, our Motion-Driven Association scheme delivers a robust data association method, recognizing the combined movement of drones and objects. This is further enriched by our innovative Adaptive Feature Synchronization (AFS) and Dual Motion-based Prediction modules. Empirical results validate DroneMOT's superiority over existing methods for drone-based MOT.

REFERENCES

[1] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 853–17 862.

[2] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *CVPR 2011*. IEEE, 2011, pp. 3153–3160.

[3] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *2008 IEEE Conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.

[4] Z. Sun, J. Chen, L. Chao, W. Ruan, and M. Mukherjee, "A survey of multiple pedestrian tracking based on tracking-by-detection framework," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1819–1833, 2020.

[5] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *European conference on computer vision*. Springer, 2020, pp. 474–490.

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

[7] 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)*. IEEE, 2016, pp. 3464–3468.

[8] X. Wan, J. Cao, S. Zhou, J. Wang, and N. Zheng, "Tracking beyond detection: learning a global response map for end-to-end multi-object tracking," *IEEE Transactions on Image Processing*, vol. 30, pp. 8222–8235, 2021.

[9] J. Cai, M. Xu, W. Li, Y. Xiong, W. Xia, Z. Tu, and S. Soatto, "Memot: Multi-object tracking with memory," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8090–8100.

[10] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8844–8854.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] 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.

[13] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[14] Y.-H. Wang, J.-W. Hsieh, P.-Y. Chen, and M.-C. Chang, "SMILEtrack: SiMIlarity LEarning for Multiple Object Tracking," Nov. 2022.

[15] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *arXiv preprint arXiv:1504.01942*, 2015.

[16] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[17] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, "Mot20: A benchmark for multi object tracking in crowded scenes," *arXiv preprint arXiv:2003.09003*, 2020.

[18] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.

[19] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 445–461.

[20] I. Kalra, M. Singh, S. Nagpal, R. Singh, M. Vatsa, and P. Sujit, "Dronesurf: Benchmark dataset for drone-based face recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–7.

[21] K. Abdulrahim and R. A. Salam, "Traffic surveillance: A review of vision based vehicle detection, recognition and tracking," *International journal of applied engineering research*, vol. 11, no. 1, pp. 713–726, 2016.

[22] J. Wang, S. Simeonova, and M. Shahbazi, "Orientation-and scale-invariant multi-vehicle detection and tracking from unmanned aerial videos," *Remote Sensing*, vol. 11, no. 18, p. 2155, 2019.

[23] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[24] S. Liu, X. Li, H. Lu, and Y. He, "Multi-object tracking meets moving uav," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8876–8885.

[25] H. Zhang, G. Wang, Z. Lei, and J.-N. Hwang, "Eye in the sky: Drone-based object tracking and 3d localization," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 899–907.

[26] G. Wang, Y. Wang, H. Zhang, R. Gu, and J.-N. Hwang, "Exploit the connectivity: Multi-object tracking with trackletnet," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 482–490.

[27] E. Schreiber, A. Heinzel, M. Peichl, M. Engel, and W. Wiesbeck, "Advanced buried object detection by multichannel, uav/drone carried synthetic aperture radar," in *2019 13th European Conference on Antennas and Propagation (EuCAP)*. IEEE, 2019, pp. 1–5.

[28] H. Hosseinpoor, F. Samadzadegan, and F. DadrasJavan, "Pricise target geolocation and tracking based on uav video imagery," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 41, pp. 243–249, 2016.

[29] S. Wang, F. Jiang, B. Zhang, R. Ma, and Q. Hao, "Development of uav-based target tracking and recognition systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3409–3422, 2019.

[30] P. Zhu, L. Wen, D. Du, X. Bian, Q. Hu, and H. Ling, "Vision meets drones: Past, present and future," *arXiv preprint arXiv:2001.06303*, vol. 1, no. 2, p. 8, 2020.

[31] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 370–386.

[32] 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)*. IEEE, 2017, pp. 3645–3649.

[33] B. Shuai, A. G. Berneshawi, D. Modolo, and J. Tighe, "Multi-object tracking with siamese track-rcnn," *arXiv preprint arXiv:2004.07786*, 2020.

[34] 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, pp. 3069–3087, 2021.

[35] B. Yan, Y. Jiang, P. Sun, D. Wang, Z. Yuan, P. Luo, and H. Lu, "Towards grand unification of object tracking," in *European Conference on Computer Vision*. Springer, 2022, pp. 733–751.

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

[37] Z. Qin, S. Zhou, L. Wang, J. Duan, G. Hua, and W. Tang, "Motiontrack: Learning robust short-term and long-term motions for multi-object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 939–17 948.

[38] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "Motr: End-to-end multiple-object tracking with transformer," in *European Conference on Computer Vision*. Springer, 2022, pp. 659–675.

[39] P. Chu, J. Wang, Q. You, H. Ling, and Z. Liu, "Transmot: Spatial-temporal graph transformer for multiple object tracking," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4870–4880.

[40] K. Liu, S. Jin, Z. Fu, Z. Chen, R. Jiang, and J. Ye, "Uncertainty-aware unsupervised multi-object tracking," *arXiv preprint arXiv:2307.15409*, 2023.

[41] 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, pp. 13 911–13 929, 2021.

[42] D. Stadler, L. W. Sommer, and J. Beyerer, "Pas tracker: Position-, appearance-and size-aware multi-object tracking in drone videos," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 604–620.

[43] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4282–4291.

[44] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8971–8980.

[45] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6182–6191.

[46] Z. Cao, Z. Huang, L. Pan, S. Zhang, Z. Liu, and C. Fu, "Tctrack: Temporal contexts for aerial tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 798–14 808.

[47] L. Shi, Q. Zhang, B. Pan, J. Zhang, and Y. Su, "Global-local and occlusion awareness network for object tracking in uavs," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[48] M. Yao, J. Wang, J. Peng, M. Chi, and C. Liu, "Folt: Fast multiple object tracking from uav-captured videos based on optical flow," *arXiv preprint arXiv:2308.07207*, 2023.

[49] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.

[50] S. J. Julier and J. K. Uhlmann, "New extension of the kalman filter to nonlinear systems," in *Signal processing, sensor fusion, and target recognition VI*, vol. 3068. Spie, 1997, pp. 182–193.

[51] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P.-J. Nordlund, "Particle filters for positioning, navigation, and tracking," *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 425–437, 2002.

[52] G. L. Smith, S. F. Schmidt, and L. A. McGee, *Application of statistical filter theory to the optimal estimation of position and velocity on board a circumlunar vehicle*. National Aeronautics and Space Administration, 1962, vol. 135.

[53] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[54] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[55] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "Bot-sort: Robust associations multi-pedestrian tracking," *arXiv preprint arXiv:2206.14651*, 2022.

[56] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-centric sort: Rethinking sort for robust multi-object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9686–9696.

[57] Q. Wang, Y. Zheng, P. Pan, and Y. Xu, "Multiple Object Tracking with Correlation Learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 3875–3885.

[58] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[59] J. Seidenschwarz, G. Brasó, V. C. Serrano, I. Elezi, and L. Leal-Taixé, "Simple Cues Lead to a Strong Multi-Object Tracker," Apr. 2023.

[60] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *European Conference on Computer Vision*. Springer, 2022, pp. 1–21.

[61] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2403–2412.

[62] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[64] B. Shuai, A. Berneshawi, X. Li, D. Modolo, and J. Tighe, "Siammot: Siamese multi-object tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 372–12 382.

[65] M. Hu, X. Zhu, H. Wang, S. Cao, C. Liu, and Q. Song, "Stdformer: Spatial-temporal motion transformer for multiple object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[66] 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*. Springer, 2016, pp. 17–35.

[67] 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.

[68] J. Tan, X. Lu, G. Zhang, C. Yin, and Q. 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*, 2021, pp. 1685–1694.

[69] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 459–474.