drones

MDPI

Article

Enhancing Online UAV Multi-Object Tracking with Temporal Context and Spatial Topological Relationships

Changcheng Xiao [1], Qiong Cao [2,*], Yujie Zhong [3], Long Lan [4,*], Xiang Zhang [4,5], Huayue Cai [1] and Zhigang Luo [1]

1 School of Computer Science, National University of Defense Technology, Changsha 410073, China
2 JD Explore Academy, Beijing 102628, China
3 Meituan Inc., Beijing 100102, China
4 Institute for Quantum & State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha 410073, China
5 Laboratory of Digitizing Software for Frontier Equipment, National University of Defense Technology, Changsha 410073, China
* Correspondence: caoqiong1@jd.com (Q.C.); long.lan@nudt.edu.cn (L.L.)

Abstract: Multi-object tracking in unmanned aerial vehicle (UAV) videos is a critical visual perception task with numerous applications. However, existing multi-object tracking methods, when directly applied to UAV scenarios, face significant challenges in maintaining robust tracking due to factors such as motion blur and small object sizes. Additionally, existing UAV methods tend to underutilize crucial information from the temporal and spatial dimensions. To address these issues, on the one hand, we propose a temporal feature aggregation module (TFAM), which effectively combines temporal contexts to obtain rich feature response maps in dynamic motion scenes to enhance the detection capability of the proposed tracker. On the other hand, we introduce a topology-integrated embedding module (TIEM) that captures the topological relationships between objects and their surrounding environment globally and sparsely, thereby integrating spatial layout information. The proposed TIEM significantly enhances the discriminative power of object embedding features, resulting in more precise data association. By integrating these two carefully designed modules into a one-stage online MOT system, we construct a robust UAV tracker. Compared to the baseline approach, the proposed model demonstrates significant improvements in MOTA on two UAV multi-object tracking benchmarks, namely VisDrone2019 and UAVDT. Specifically, the proposed model achieves a 2.2% improvement in MOTA on the VisDrone2019 benchmark and a 2.5% improvement on the UAVDT benchmark.

Keywords: multiple object tracking; unmanned aerial vehicle videos; feature aggregation; deformable attention; topological relationships

check for updates

Citation: Xiao, C.; Cao, Q.; Zhong, Y.; Lan, L.; Zhang, X.; Cai, H.; Luo, Z. Enhancing Online UAV Multi-Object Tracking with Temporal Context and Spatial Topological Relationships. *Drones* **2023**, *7*, 389. https://doi.org/10.3390/drones7060389

Academic Editor: Diego González-Aguilera

Received: 22 May 2023
Revised: 8 June 2023
Accepted: 8 June 2023
Published: 10 June 2023

1. Introduction

Visual multi-object tracking is a fundamental computer vision task that aims to determine the trajectories of objects of interest in video sequences. With the advancement of deep learning, multiple-object tracking (MOT) techniques have rapidly evolved over the past decade [1–5]. However, the majority of research has focused on tracking objects in fixed or horizontally moving camera settings, such as handheld or vehicle-mounted cameras, with limited perceptual range. In recent years, unmanned aerial vehicles (UAVs) have gained widespread popularity in various domains, such as search and rescue, agriculture, sports analysis, and geographical surveying [6–10]. Multi-object tracking in airborne camera views faces more complex challenges than with fixed or horizontally moving cameras, such as small target sizes and fast camera movements [11,12]. As a result, there is a growing need for innovative techniques that can handle the unique complexities present in UAV scenarios.

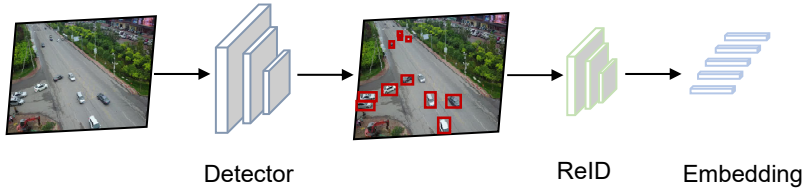
In the past, MOT mainly used the paradigm of detection followed by tracking [13–15], which consists of two steps: detection and data association. These algorithms extract the object representation features from the bounding boxes and combine the trajectory location information for data association, namely separate detection and embedding (SDE) methods. This paradigm mainly focuses on extracting features for target re-identification and optimizing the data association step. With the development of object detection [16–19] and re-identification techniques [20], this type of approach has achieved rapid development. However, this cascading architecture is not efficient enough and cannot be jointly optimized.

A more practical approach is to directly extend the one-stage object detector [21,22] to jointly perform object localization and ReID feature extraction, i.e., joint detection and embedding (JDE) approaches. Nonetheless, the high observation altitude and fast flight speed of UAVs present challenges in object perception. These factors in combination with video blurring further make the detection of small objects more difficult. As a consequence, the single-frame-based methods are prone to generating temporally inconsistent detection results, such as false negatives. This, in turn, negatively impacts the overall tracking performance. In addition, the topological relationships between objects and contextual information are disregarded during the extraction of object re-identification features. In contrast, recent studies [23–25] have demonstrated that taking into account the relationships between objects can enhance the discriminative power of the ReID embedding features, which subsequently improves the accuracy of data association. Consequently, a one-stage tracker that does not utilize contextual information is sub-optimal.

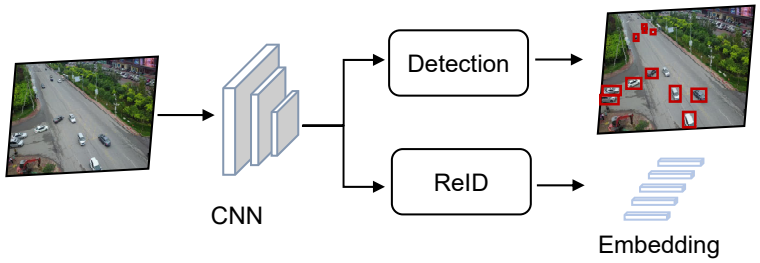
To address the aforementioned challenges, we explore the utilization of temporal consistency information within the video, along with the incorporation of topological relationships pertaining to the objects and background. The difference between the SDE methods, the JDE methods, and the proposed method is briefly illustrated in Figure 1. In earlier frames, objects that are currently occluded and blurred might be identifiable. Therefore, by leveraging the features from previous frames, we can enhance the current features and recover potentially overlooked objects. First, we propose a novel temporal feature aggregation module (TFAM) that leverages features from previous frames to supplement the information of missing objects in the current frame. Specifically, this module leverages the spatial correlation between the current frame and the previous frames to obtain offsets, which are subsequently used in a deformable convolution [26] layer to warp the features of the previous frames. Next, the propagated features and the features of the current frame are fused to improve the coherence of object detection. Further, we design an object topology-integrated embedding module (TIEM) to model the topological relationships between the objects and the environment. More specifically, we exploit a global and sparse approach based on a deformable attention mechanism for effective global relationship modeling.

In summary, the main contributions of this paper are as follows:

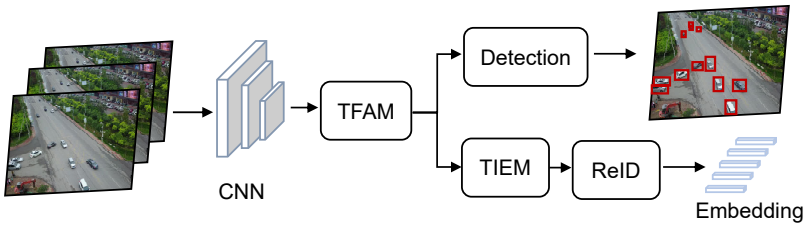
- We design a novel temporal feature aggregation module to enhance the temporal consistency of network perception. By adaptively fusing multiple frame features, we improve the robustness of perception in UAV scenarios, including occlusion, small objects, and motion blur.
- We propose a topology-integrated embedding module to model the long-range dependencies of the entire image in a global and sparse manner. By incorporating global contextual information through a deformable attention mechanism, the discriminative power of the object embedding is enhanced, leading to improved accuracy in data association.
- Combining the two proposed modules together, our approach shows advanced performance on two existing UAV multi-object tracking benchmarks: VisDrone [11] and UAVDT [12].



(**a**) Separate Detection and Embedding



(**b**) Joint Detection and Embedding



(**c**) The proposed tracker

Figure 1. Comparison between (**a**) the separate detection and embedding method, (**b**) the joint detection and embedding method, and (**c**) the proposed tracker. The TFAM module aggregates multi-frame features, and TIEM establishes topological relationships between the object and other objects as well as the environment.

This paper is organized as follows. Section 2 reviews the related work. Section 3 describes our proposed model. Section 4 provides the experimental results and ablation analysis. Section 5 concludes the paper.

2. Related Work

2.1. Multi-Object Tracking (MOT)

MOT is a technology with diverse applications that has garnered significant attention from scholars. In its early stages, researchers primarily focused on leveraging optimization algorithms to derive object trajectories [27,28]. To make the multi-object tracking algorithm more practical, Bochinski et al. proposed the simplest IOUTracker [29], which relies solely on the intersection over union of bounding boxes to achieve tracking. Based on this, researchers added the motion model [13] and the Kalman filter [30] to predict the position of the target in the next frame. Despite exhibiting fast running speeds and significantly improved performance, this model not perform well in challenging scenarios, such as those with occlusion and object loss. To this end, researchers [14,15] introduced re-identification (ReID) features as appearance models to improve the accuracy of the association between trajectories and detection results. MOANA [31] is an adaptive appearance model that learns online the long-term changes in the appearance of objects. Its main focus is on solving the problem of object appearance changes to improve tracking performance. However, the additional introduction of the ReID model with individual object image patches as inputs

makes it a computational bottleneck in the overall tracking system. Furthermore, it is imperative to note that the ReID feature extraction approach, which relies on object patches, may manifest itself as being unreliable, consequently leading to sub-optimal tracking outcomes. Wang et al. [32] proposed a tracklet booster algorithm that can be embedded into existing trackers.

2.2. Joint Detection and Tracking

To improve the speed of operation of the entire tracking system, researchers [22,33,34] integrated the object detection and extraction of ReID features into the same neural network to share most of the computations. JDE [33] was the first work to do so, with the innovative addition of a ReID branch to the one-stage detector, YOLOv3 [35]. FairMOT [22] achieved more balanced detection and recognition tasks by reducing anchor ambiguity using an anchor-free detector, CenterNet [18]. In addition to these joint detection and embedding methods, some other one-stage trackers have emerged. Tracktor [21] implements the inter-frame association of object trajectories directly using the detector's regression head. CenterTrack [36] and ChainedTracker [37] employ a multi-frame approach to predict the bounding boxes for consecutive frames, which facilitates efficient short-term associations, ultimately obtaining long-term trajectories. Nevertheless, these techniques tend to generate numerous identity switches owing to their inability to capture long-term dependencies.

2.3. Regression-Based Tracking Method

Feichtenhofer et al. [38] proposed a multi-task network to perform joint detection and tracking. This network makes use of a convolutional neural network to extract features from input frames and generate corresponding feature response maps. A correlation layer is then employed to compute the local similarity between the feature maps of two consecutive frames. Finally, position-sensitive region of interest (RoI) pooling is performed on the feature response maps of each frame to obtain detection results on a single frame. The same process is executed on the correlated features to compute the objects' offsets in adjacent frames, facilitating the association of objects across neighboring frames. Tracktor [21] utilizes a two-stage object detector, Faster R-CNN [39], where region of interest (RoI) layers in the current frame are generated based on the bounding boxes of objects in the previous frame. The bounding box in the current frame is considered as the same object if its intersection over union (IoU) with the object's bounding box in the previous frame is deemed sufficiently high, thus establishing the association. CTrack [37] integrates the regression results of two paired bounding boxes generated for overlapping nodes that cover two adjacent frames. Inter-frame regression is accomplished via target attention and identity attention, which are introduced by the detection module. This method boasts a simple structure and operates at a high speed. Building upon Tracktor, Guo et al. [40] proposed a method to improve tracking performance by leveraging the synergistic effect between position prediction and embedded association. These two tasks were correlated through time-aware object attention and interference attention, as well as identity-aware memory aggregation. Specifically, the attention module directs predictions to focus more on the target and reduce interference, thus extracting more reliable embedding for associations. On the other hand, these reliable embeddings enhance identity awareness through memory aggregation, strengthening the attention module and suppressing drift. This synergistic effect between position prediction and embedded association improves robustness to occlusions and yields better performance in occluded scenarios.

2.4. Attention-Based Methods

In recent years, the success of transformer-based models in computer vision, particularly in the field of object detection, has led to the emergence of several transformer-based approaches for MOT. Notable among them are TransTrack [41], TrackFormer [42] and MOTR [43], all of which are online trackers based on DETR [44] and its variants. TrackFormer utilizes track queries to maintain object identities and suppress duplicate tracks.

TransTrack, on the other hand, employs previous object features as track queries to acquire tracking boxes and associates detection boxes based on IoU matching. Additionally, MOTR performs end-to-end object tracking by iteratively updating the track query without requiring post-processing. MeMOT [45] is an end-to-end tracking method similar to MOTR based on attention, which enables the prediction of object states through the attention mechanism. GTR is an offline transformer-based tracker that employs queries to divide detected boxes into trajectories all at once instead of generating tracking boxes. Although these methods explore new tracking paradigms, their performance is still sub-optimal to that of advanced tracking algorithms.

2.5. Graph-Based Methods

The use of graph optimization [27,46] for MOT is a technique that was commonly used in the past, which uses a single object obtained by cropping as a node. However, recent advancements have shown that graph-neural-network (GNN)-based methods [23,47,48] are a promising alternative. STRN [47] uses a graph convolutional network (GCN) to propagate features across spatial–temporal space. Another approach, MPN [23], utilizes a message-passing network (MPN) to parse the information and associate detections through edge classification. GM-Tracker [48] models the relationships between tracklets and the intra-frame detections as a general undirected graph. These methods model the relationships between objects at the instance level only. In contrast to alternative approaches, our methodology operates at the frame feature level. This technique confers a notable advantage by enabling the extraction of information from both foreground and background elements while minimizing the loss of contextual information.

3. Method

Given a video sequence $\{I_t \in \mathbb{R}^{(H\times W\times 3)}\}_{t=1}^{T}$ obtained by a moving UAV, we aim to give the categories, bounding boxes, and object identities for all objects of interest. Figure 2 shows the overall pipeline of the proposed method. Our approach consists of five main components, namely, generic feature extraction, multi-frame feature aggregation, detection, object embedding enhancement, and association. At first, the backbone $\mathbf{\Phi}$ (DLA-34 [49]) extracts the feature map $f \in \mathbb{R}^{(H'\times W'\times C)}$ of each frame, where $H' = \frac{H}{4}$ and $W' = \frac{W}{4}$. Our main contributions lie in the proposal of the temporal feature aggregation module (TFAM) and the topology-integrated embedding module (TIEM). In the association stage, we utilize the Hungarian [50] matching algorithm to allocate the detected objects to the corresponding trajectories based on the acquired detection results and object embedding features.

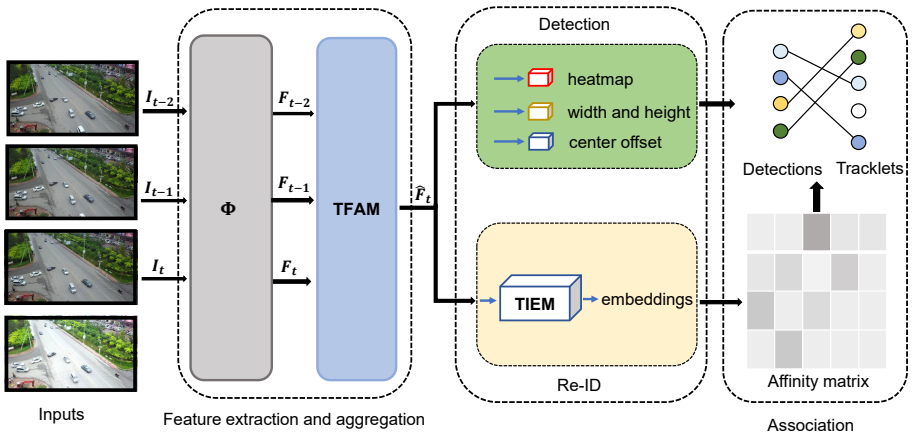


Figure 2. An overview of the proposed method.

3.1. Temporal Feature Aggregation

Addressing motion blur and occlusion is a critical aspect in UAV scenarios, and a key strategy involves supplementing the missing object cues. In the context of videos,

where a temporal dimension is present, it is often assumed that absent object cues in the current frame may have been visible in previous frames. Thus, utilizing previous frames to enhance feature representation is a natural approach. However, due to the motion of objects and the UAV, the features of different frames are not spatially aligned, which further exacerbates the problem. To solve this problem, we design a novel temporal feature fusion module, shown in Figure 3. It consists of two steps, feature propagation and adaptive feature aggregation.

3.1.1. Feature Propagation

Denote by $\mathbf{F_q} \in \mathbb{R}^{H' \times W' \times C}$ and $\mathbf{F_r} \in \mathbb{R}^{H' \times W' \times C}$ the feature map of the current frame (or query frame) and a reference frame (namely, a previous frame), respectively. To achieve the propagation of the reference feature maps, we exploit a single layer of deformable convolution (DCN) [26]. DCN takes a reference feature map $\mathbf{F}_r$ and a spatial offset $\mathbf{D}$ as input and outputs a calibrated feature $\widetilde{\mathbf{F}}_r$. We first calculate the spatial offset $\mathbf{D}$ based on the spatial correlation [51] between $\mathbf{F}_q$ and $\mathbf{F}_r$. Formally, the 'correlation' of two patches centered at x_q in $\mathbf{F}_q$ and x_r in $\mathbf{F}_r$ can be formulated as

$$c(x_q, x_r) = \sum_{o \in [-k,k] \times [-k,k]} \mathbf{F}_q(x_q + o)\mathbf{F}_r(x_r + o)^T. \tag{1}$$

The above operation will cover a spatial area of $d \times d$ centered on x_q. The larger the d, the larger the range of position offsets that can be handled. The output of the correlation is in four dimensions, where each pair of 2D positions produces a corresponding correlation value. In practice, we reshape the relative displacements in channels and obtain $\mathbf{S} \in \mathbb{R}^{H' \times W' \times d^2}$. Then, the spatial offset $\mathbf{D} \in \mathbb{R}^{H' \times W' \times 2K^2}$ is calculated by three convolutional layers, where K is the size of the DCN convolutional kernel. Then, given the spatial offset $\mathbf{D}$ and a previous feature map, the propagated feature is obtained via a DCN as

$$\widetilde{\mathbf{F}}_r = DCN(\mathbf{D}, \mathbf{F}_r). \tag{2}$$

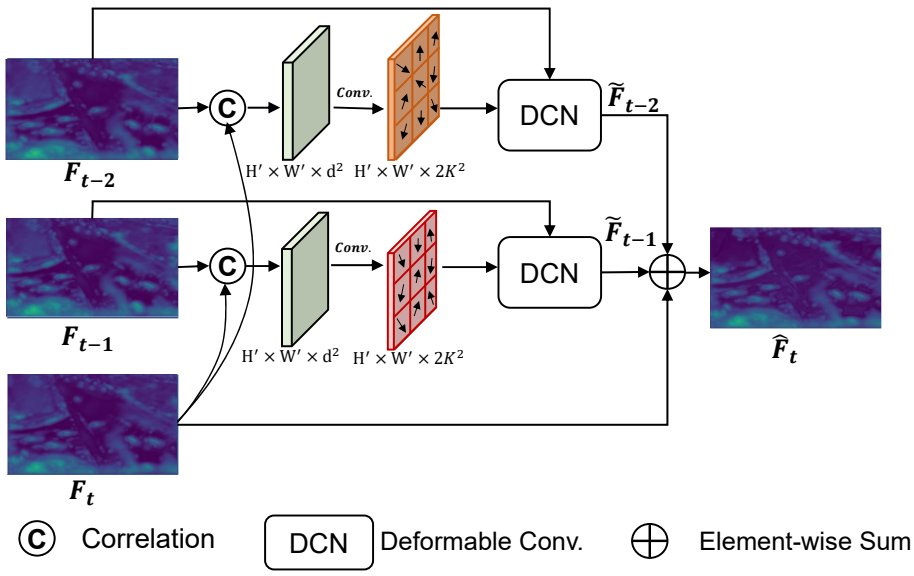


Figure 3. The architecture of the proposed temporal feature aggregation module. For the sake of simplicity, the adaptive feature aggregation step is not shown in the above figure.

3.1.2. Adaptive Feature Aggregation

We propose to improve the perception of the tracker for occlusion, blur, and small objects by aggregating the propagated features of L previous frames and the feature map of the current frame $\mathbf{F}_t$. Specifically, we first obtain the global and local information of each feature map by global average pooling and a point-wise convolution operation. Then, the

adaptive weights for each frame are obtained from a convolutional layer and subsequent softmax calculations. The adaptive weight indicates the importance of the buffer feature maps $\{\widetilde{\mathbf{F}}_L, \widetilde{\mathbf{F}}_{L-1}, \cdots, \mathbf{F}_t\}$ at each spatial location. The reinforced feature map is denoted as $\hat{\mathbf{F}}_t$, which can be calculated by the following equation:

$$\hat{\mathbf{F}}_t = \mathbf{w}^t \otimes \mathbf{F}_t + \sum_{j=1}^{L} \mathbf{w}^{t-j} \otimes \widetilde{\mathbf{F}}_{t-j}, \tag{3}$$

where $\mathbf{w} \in \mathbb{R}^{H' \times W' \times 1}$ and $\sum_{j=0}^{L} \mathbf{w}^{t-j} = 1$. $\otimes$ denotes the element-wise multiplication. The reinforced feature map $\hat{\mathbf{F}}_t$ is fed into the subsequent head network to obtain the bounding boxes and object embedding features. This has the potential to detect missed objects and improve the consistency of the object ID features, resulting in more complete trajectories.

3.2. Topology-Integrated Embedding Module

It is an intuitive approach [22] to extract the object embeddings based on the object center location directly from the feature map obtained from the ReID branch. Nevertheless, utilizing extracted point-wise embeddings directly, without taking contextual information into account, results in weak discriminative capabilities. As a consequence, it leads to inaccurate matching and sub-optimal tracking outcomes. Exploiting an attention mechanism [52] for long-range modeling is a straightforward strategy, but the quadratic complexity of its global operations limits fast training and applications to larger-resolution feature maps.

Given the ReID feature map $\mathbf{F}_{id} \in \mathbb{R}^{H' \times W' \times C}$, we reshape it as a sequence $\mathbf{x} \in \mathbb{R}^{N \times C}$, where $N = H'W'$. With $\mathbf{x}$ as input, the vanilla multi-head self-attention mechanism (MHSA) can be formulated as follows:

$$\begin{aligned} \mathbf{q} = \mathbf{x}\mathbf{W}_q, \mathbf{k} = \mathbf{x}\mathbf{W}_k, \mathbf{v} = \mathbf{x}\mathbf{W}_v, \\ \mathbf{z}^{(m)} = \phi\left(\frac{\mathbf{q}^{(m)}\mathbf{k}^{(m)\top}}{\sqrt{d}}\right)\mathbf{v}^{(m)}, m = 1, \ldots, M, \\ \mathbf{z} = \mathrm{Concat}\left(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(M)}\right)\mathbf{W}_o, \end{aligned} \tag{4}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$, and $\mathbf{W}_o \in \mathbb{R}^{C \times C}$ are learnable projection matrices. $\phi(\cdot)$ denotes the softmax function, m denotes the number of heads, and $d = C/M$ is the dimension of each head. The computational complexity is $O(N^2C)$, which grows quadratically with respect to the size of the input feature map.

To maintain the long-range dependence of the attention mechanism and reduce computational expenses, we design an object-topology-integrated embedding module, illustrated in Figure 4, which incorporates deformable attention [53]. This module enables the establishment of a global sparse relationship with both other objects and the surrounding environment. Specifically, a uniform grid $\mathbf{p} \in \mathbb{R}^{(H_G \times W_G)}$ is generated as reference for the input feature map $\mathbf{F}_{id}$, where $H_G = H'/r, W_G = W'/r$, and r is the down sampling factor. A lightweight sub-network $\theta(\cdot)$ takes $\mathbf{q}$ as input and predicts the offset $\Delta\mathbf{p} = \theta(\mathbf{q})$ for it. For training stability, the size of $\Delta\mathbf{p}$ is constrained by a preset coefficient s to avoid too large an offset, i.e., $\Delta\mathbf{p} \leftarrow s\tanh(\Delta\mathbf{p})$. Then, the features can be sampled based on the reference grid $\mathbf{p}$ and offsets $\Delta\mathbf{p}$ to obtain values and keys. We thus have

$$\begin{aligned} \mathbf{q} = \mathbf{x}\mathbf{W}_q, \widetilde{\mathbf{k}} = \widetilde{\mathbf{x}}\mathbf{W}_k, \widetilde{\mathbf{v}} = \widetilde{\mathbf{x}}\mathbf{W}_v, \\ \text{With} \quad \Delta\mathbf{p} = \theta(\mathbf{q}), \widetilde{\mathbf{x}} = G(\mathbf{x}, \mathbf{p} + \Delta\mathbf{p}). \end{aligned} \tag{5}$$

where $\widetilde{\mathbf{k}}$ and $\widetilde{\mathbf{v}}$ denote the deformed keys and values, respectively. Since $\Delta\mathbf{p}$ is usually fractional, the sampling function $G(\cdot, \cdot)$ is implemented by bilinear interpolation to achieve a differentiable equation

$$G(I;(p_x,p_y)) = \sum_{(i_x,i_y)} g(i_x,p_x)g(i_y,p_y)I[i_x,i_y,:], \tag{6}$$

where $g(a,b) = \max(0, 1-|a-b|)$ and (i_x, i_y) indexes all locations on $\mathbf{I} \in \mathbb{R}^{H' \times W' \times C}$. In addition to a deformable multi-head attention(DMHA) block, the TIEM also contains a feed-forward network (FFN) and a layer normalization layer (LN) layer [54]. The TIEM module can be formulated as

$$\begin{aligned} \hat{\mathbf{z}} &= \mathrm{DMHA}(\mathbf{z}) + \mathbf{z} \\ \mathbf{z} &= \mathrm{LN}(\mathrm{FFN}(\hat{\mathbf{z}})) + \hat{\mathbf{z}}, \end{aligned} \tag{7}$$

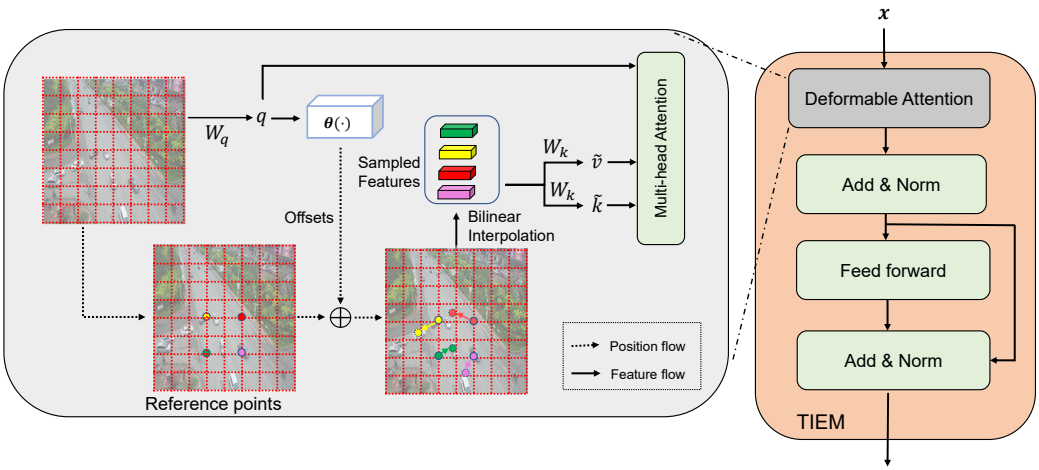


Figure 4. Diagram of the proposed feature enhancement module.

3.3. Optimization Objectives

The proposed network contains detection and re-identification branches and multiple optimization objects. Here, we present the optimization objects of the task separately.

Detection branch. First, the central location of the objects of interest is obtained by the heatmap branch. The expected response of a location within a heatmap is one when it coincides with the center of the ground-truth object. The response value decreases exponentially as the distance between the heatmap location and the object center increases. For each bounding box annotation $\mathbf{b}_m^i = (x_1^i, y_1^i, x_2^i, y_2^i)$ of class m in the image, we can obtain its central location (c_{mx}^i, c_{my}^i) as $c_{mx}^i = \frac{x_1^i + x_2^i}{2}$ and $c_{my}^i = \frac{y_1^i + y_2^i}{2}$, respectively. Next, by dividing down the sampling stride, its position on the feature map becomes $(\tilde{c}_{mx}^i, \tilde{c}_{my}^i) = (\lfloor \frac{c_{mx}^i}{4} \rfloor, \lfloor \frac{c_{my}^i}{4} \rfloor)$. Suppose the image has N object bounding boxes. The heatmap ground truth of the class m can be obtained by the following equation:

$$Y_{mxy} = \sum_{i=1}^{N} \exp\left(-\frac{(x-c_{mx}^i)^2 + (y-c_{my}^i)^2}{2(\sigma_p)^2}\right), \tag{8}$$

where Y_{mxy} is the pixel value at coordinate (x, y) in the rendered heatmap, and σ_p represents the standard deviation of the object size. $\hat{Y}_{mxy}$ denotes the predicted heatmap pixel at (x, y). The loss for this predicted value can be calculated by the following equation:

$$L_{mxy}^h = \begin{cases} (1-\hat{Y}_{mxy})^\alpha \log \hat{Y}_{mxy}, & \text{if } Y_{mxy} = 1; \\ (1-Y_{mxy})^\beta (\hat{Y}_{mxy})^\alpha \log(1-\hat{Y}_{mxy}), & \text{otherwise}, \end{cases} \tag{9}$$

where α and β are hyper-parameters [18]. Further, we can formalize the loss function of the heatmap for M categories, as follows:

$$L_{heat} = -\frac{1}{N} \sum_{c=1}^{M} \sum_{y=1}^{H} \sum_{x=1}^{W} L_{xy}^h. \tag{10}$$

Then, the object is located more precisely by offset and size branches. In this case, the offset branch serves to eliminate the error of up to four pixel values introduced by the down-sampling process. The optimization objectives of these two branches are given in the following equation:

$$L_{box} = \sum_{i=1}^{N} \|o^i - \hat{o}^i\|_1 + \lambda_s \|s^i - \hat{s}^i\|_1, \tag{11}$$

where o^i and s^i are the predicted values of the offset and size branches, respectively. The weighting parameter λ_s is set to 0.1, as in the original CenterNet [18].

ReID branch. We formalize the ReID task as a classification problem, where objects of the same identity are treated as the same category. Based on the center position $(\tilde{c}_x^i, \tilde{c}_y^i)$ of the GT annotation in the heatmap, we extract the ReID feature vector $\mathbf{e}_{xy}$ of the object at the corresponding position on the feature map output from the ReID branch. Additionally, a fully connected layer and a softmax operation are exploited to transform it to the distribution vector $P = \{p_i\}_{i=1}^{K}$, where K is the total number of categories. Suppose the one-hot annotation of the ReID task is $Q = \{q_i\}_{i=1}^{K}$; then, the loss function can be formalized as follows:

$$L_{id} = \sum_{i=1}^{N} \sum_{j=1}^{K} q_j \log(p_i). \tag{12}$$

Overall Losses. By adding up the above losses, we can train both the detection and ReID branches. Specifically, we adopt uncertainty loss [55] to automatically balance the detection and re-identification tasks.

$$L_{det} = L_{heat} + L_{box}, \tag{13}$$

$$L = \frac{1}{2}\left(\frac{1}{e^{w_1}} L_{det} + \frac{1}{e^{w_2}} L_{id} + w_1 + w_2\right), \tag{14}$$

where w_1 and w_2 are learnable coefficients that balance the two tasks.

3.4. Online Tracking

In this subsection, we describe the online tracking process of the proposed method in detail. We adopt the cascade association strategy used in previous works [15,22,33]. In the first round of association, we calculate the object embedding features similarity A_e between candidate objects and the existing trajectories. Furthermore, we use a Kalman filter [30] to predict the spatial coordinates of tracklets in the current frame. Subsequently, the Mahalanobis similarity, denoted as A_m, is calculated between the predicted tracklet positions and the corresponding detected bounding boxes as in DeepSORT [14]. Next, we fuse A_e and A_m by the following equation to obtain the final affinity matrix A:

$$A = \alpha A_e + (1 - \alpha) A_m, \tag{15}$$

where α is a weighting coefficient and is set to be 0.98. Finally, we obtain optimal bipartite matching results using the Hungarian algorithm [50]. It is worth noting that in cases where the tracklet and the candidate target are located at a significant spatial distance, the corresponding match will be deemed as unreasonable and rejected.

In the second association process, we rely solely on the intersection over union (IOU) value of the bounding box between unmatched objects and trajectories to make associations. Unmatched candidates are initialized as new tracks, and unmatched tracklets are kept for up to 30 frames in case of reappearance. To address the variations of objects in appearance, we update the identity embedding of tracklets that have been successfully matched in each time step t. This update is performed according to the following equation:

$$\tilde{e}_t = \beta \tilde{e}_{t-1} + (1 - \beta) e, \tag{16}$$

where e represents the object feature embedding of the assigned object and $\tilde{e}_t$ indicates the embedding of a tracklet at time step t. β is a momentum coefficient for smoothing and is set to be 0.9.

4. Experiments

4.1. Datasets and Metrics

4.1.1. Dataset

To verify the effectiveness of the proposed method, research experiments were conducted on two existing UAV video multi-object tracking datasets, namely VisDrone2019 [11] and UAVDT [12]. VisDrone2019 contains 56 training video sequences, 7 validation sequences, and 17 test-dev sequences. The videos includes a diverse range of scenarios, including sports fields, commercial streets, highways, and suburbs. In evaluating the multi-object tracking task, five categories of objects were considered: cars, buses, trucks, pedestrians, and vans. In contrast, only a single category of objects, cars, is tracked in the UAVDT dataset. This dataset consists of 50 videos (30 for training and 20 for testing), mainly taken in plazas, intersections, and highways under different lighting conditions (e.g., day, night, fog, etc.). Videos captured by UAVs face more complex challenges than other multi-object tracking benchmarks [2,3]. These challenges include a higher proportion of small objects (whose pixel values are less than 32×32) and motion blur caused by the motion overlap between the UAV and the objects. These challenges may cause object tracking to fail.

4.1.2. Metrics

We used the official evaluation toolbox provided by these two benchmarks to evaluate the performance of our algorithms. The evaluation metrics mainly include multiple object tracking accuracy (MOTA) [56], the number of false negatives (FN), the number of false positives (FP), and the number of identity switches (IDs). The MOTA is defined as follows:

$$\mathrm{MOTA} = 1 - \frac{\mathrm{FN} + \mathrm{FP} + \mathrm{IDs}}{\mathrm{GT}}, \tag{17}$$

where GT is the number of ground truth bounding boxes. The identification F1-score (IDF1) [57] matches the ground truth and the predicted trajectories and calculates the corresponding F1-score on the trajectory level. It is defined as:

$$\mathrm{IDF1} = \frac{|\mathrm{IDTP}|}{|\mathrm{IDTP}| + 0.5|\mathrm{IDFN}| + 0.5|\mathrm{IDFP}|}, \tag{18}$$

where IDTP, IDFN, and IDFP are the true-positive, false-negative, and false-positive trajectories. IDF1 mainly focuses on measuring the association performance of trackers. MOTA and IDF1 are the main metrics for measuring the models' tracking performances.

4.2. Implementation Details

We trained the proposed network with a backbone of a variant DLA-34 pre-traind on the COCO dataset [58]. The parameters of the proposed model were updated by exploiting the Adam optimizer [59] with an initial learning rate of 7×10^{-5} for 30 epochs. The learning rate was decreased by a factor of 10 at the twentieth epoch. We used common data augmentation techniques, including rotation, scaling, and color jittering, and the input image was resized to 1088×608. During the training phase, a set of L reference frames was randomly selected from a range of 5 frames centered around the current frame. In the inference phase, the L consecutive previous frames were utilized. We conducted the experiments using two NVIDIA A100 GPUs with a batch size of 12.

4.3. State-of-the-Art Comparison

As can be seen from Table 1, our method achieves the best performance regarding MOTA, IDF1, and MT. Our method achieves 30.9% MOTA and 42.7% IDF1, which represent significant improvements over the baseline method, FairMOT [22], by 2.2% and 2.9%, respectively. Meanwhile, compared with the transformer-based end-to-end approach, Trackformer [42], the proposed tracker has more significant advantages, reaching values of 5.9% and 12.2% for the MOTA and IDF1 metrics, respectively. In addition, our method outperforms all other methods in terms of MT metrics, which indicates that our method maintains a more complete trajectory.

UAVDT. We further evaluated the proposed method on the UAVDT benchmark. As can be seen from Table 2, our method outperforms the previous methods in most metrics. Notably, our proposed method outperforms FairMOT [22] by 2.5% and 1.5% in the evaluation metrics of MOTA and IDF1, respectively. Furthermore, our proposed tracker achieves the highest scores in terms of the MT, FN, and IDs compared to the existing methods. These findings collectively demonstrate the outstanding performance of our proposed model, which can be attributed to the TFAM and TIEM modules that are designed to improve the consistency of the detection and obtain discriminative object embedding.

Table 1. Results on VisDrone2019 test-dev dataset. The best results are shown in **bold**.

Method	MOTA↑	IDF1↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓
MOTDT [15]	−0.8	21.6	68.5	87	1196	44,548	185,453	1437	3609
SORT [13]	14.0	38.0	73.2	506	**545**	80,845	**112,954**	3629	4838
IOUT [29]	28.1	38.9	74.7	467	670	36,158	126,549	2393	3829
GOG [60]	28.7	36.4	**76.1**	346	836	17,706	144,657	1387	**2237**
MOTR [43]	22.8	41.4	72.8	272	825	28,407	147,937	**959**	3980
TrackFormer [42]	25.0	30.5	73.9	385	770	25,856	141,526	4840	4855
FairMOT [22]	28.7	39.8	75.1	449	758	**22,771**	137,215	3611	6162
Ours	**30.9**	**42.7**	74.4	**491**	668	27,732	126,811	3998	7061

Table 2. Results on UAVDT test dataset. The best results are shown in **bold**.

Method	MOTA↑	IDF1↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓
CEM [61]	−6.8	10.1	70.4	94	1062	64,373	298,090	1530	2835
SMOT [62]	33.9	45.0	72.2	524	367	57,112	166,528	1752	9577
GOG [60]	35.7	0.3	72	627	374	62,929	153,336	3104	5130
IOUT [29]	36.6	23.7	72.1	534	357	42,245	163,881	9938	10,463
CMOT [63]	36.9	57.5	**74.7**	**664**	351	69,109	144,760	1111	3656
SORT [13]	39.0	43.7	74.3	484	400	**33,037**	172,628	2350	5787
DeepSORT [14]	40.7	58.2	73.2	595	338	44,868	155,290	2061	6432
MDP [63]	43.0	61.5	73.5	647	324	46,151	147,735	541	**4299**
FairMOT [22]	44.5	66.3	72.2	640	**193**	71,922	116,510	664	6326
Ours	**47.0**	**67.8**	72.9	652	**193**	68,282	**111,959**	**506**	5884

4.4. Ablation Analysis

We trained the model on the training set of VisDrone 2019 and validated the effectiveness of the proposed method on the VisDrone validation set. In this section, we adopted FairMOT with a backbone of variant DLA-34 [18] as the baseline. We conducted a series of ablation analyses on the critical components and related hyper-parameters of the proposed method.

4.4.1. Component-Wise Analysis

As shown in Table 3, the proposed temporal feature aggregation module brings a 2.4% gain in MOTA, verifying that fusing multi-frame features can improve detection consistency and reduce false positives. The topology-integrated embedding module improves IDF1

from 43.2% to 45.7%, indicating its effectiveness in enhancing the discriminative nature of target embedding. Finally, by combining the two, our method achieves a boost compared to the baseline method, with MOTA and IDF1 improving by 3.1% and 2.9%, respectively.

Table 3. Component-wise analysis of the proposed method. The best results are shown in **bold**.

TFAM	TIEM	MOTA↑	IDF1↑	FP↓	FN↓	IDs↓
		26.0	43.2	10,648	41,605	1019
✓		28.4	44.0	9074	41,398	948
	✓	27.3	45.7	11,645	**39,601**	971
✓	✓	**29.1**	**46.1**	**9589**	40,380	**938**

4.4.2. Effect of Different Feature Fusion Strategies

As shown in Table 4, we compared different strategies for fusing video frame features. Among them, the first strategy directly added $\widetilde{\mathbf{F}}_{L-1}$ with $\mathbf{F}_t$ in an element-wise manner. The second strategy cascaded them directly, resulting in a 0.5% MOTA improvement compared to the former. The adaptive convergence strategy further improves the MOTA by 1.3% compared to the cascade strategy, indicating that adaptive fusion based on the feature map is optimal.

Table 4. The ablation study of feature aggregation strategy. The best results are shown in **bold**.

Fusion Strategies	MOTA↑	IDF1↑	FP↓	FN↓	IDs↓
Addition	26.6	45.5	11,109	**40,635**	1001
Concatenation	27.1	**45.8**	10,815	40,708	**851**
Adaptive feature aggregation	**28.4**	44.0	**9074**	41,398	948

4.4.3. Effect of Different Sizes of Local Regions

Table 5 showcases the impact of different local sizes, d. As described in Section 3.1, a larger value of d covers a broader spatial area and enables the handling of larger motion offsets. The results indicate that increasing the local size can enhance the tracking performance by expanding the spatial coverage and accommodating larger motion changes. Therefore, an appropriate local size d of 16 was adapted to optimize the tracking performance.

Table 5. Effect of spatial local size d. The best results are shown in **bold**.

d	MOTA↑	IDF1↑	FP↓	FN↓	IDs↓
4	26.6	44.4	11,532	41,187	**905**
8	27.4	**46.4**	11,112	**40,122**	912
16	**28.4**	44.0	**9074**	41,398	948
20	26.3	45.6	11,279	40,710	937
24	25.1	44.7	12,369	40,549	906

4.4.4. Effect of Number of Previous Features

In addition to the aforementioned analysis, we investigated the impact of fusing different numbers of previous features (as defined in Equation (3)) and present the results in Table 6. It was found that the optimal value of MOTA (28.4%) is achieved when the number L of preceding features is chosen as 2. However, continuing to use more previous frame features does not result in more gain, so L was set to 2. As expected, aggregating multi-frame features helped to improve the perception of the tracker.

Table 6. Ablation studies on number of previous features, L. The best results are shown in **bold**.

L	MOTA↑	IDF1↑	FP↓	FN↓	IDs↓
1	28.1	44.4	9679	**41,077**	900
2	**28.4**	44.0	9074	41,398	948
3	27.5	**45.4**	9958	41,253	850
4	27.4	43.2	9034	42,429	**735**
5	27.6	44.2	**8759**	42,501	791

4.4.5. Effect of Different Sizes of the Coefficient s

As demonstrated in Table 7, we explored the impact of different sizes of s in the topology-integrated embedding module on the tracking performance, particularly focusing on the IDF1 metric, which measures trajectory consistency. A noteworthy observation is the significant improvement in IDF1 (from 44.3% to 45.7%) when s is increased from 2 to 4. However, when s is further increased beyond this point, there is a decline in performance. To strike a balance between MOTA and IDF1, we set the value of s to 4.

Table 7. Ablation on the sizes of the coefficient s. The best results are shown in **bold**.

s	IDF1↑	MOTA↑	FP↓	FN↓	IDs↓
2	44.3	25.4	11,434	41,155	976
3	44.7	**27.7**	**8880**	42,300	**796**
4	**45.7**	27.3	11,645	**39,601**	971
5	43.4	25.4	10,175	42,422	968
6	44.5	25.7	12,027	40,307	1034

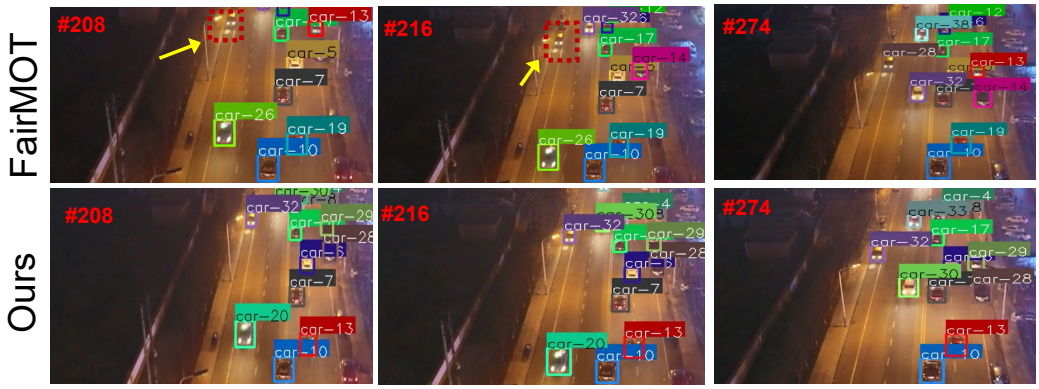
4.5. Qualitative Results

In this section, we first give the visualization results of the proposed method and analyze its robustness. We compare the proposed method with the baseline method FairMOT in Figure 5. As shown in Figure 5a,b, FairMOT failed to detect and track a large number of objects in the areas marked by red dashed boxes, while our proposed method successfully located and tracked these objects in these challenging low-light scenes, which we attribute to the proposed multi-frame feature fusion module and target embedding feature enhancement module. This validates the idea that multi-frame features can provide more temporal contextual cues and that the spatial–topological relationships can improve the discriminative power of object feature embeddings. In addition to showing two random cases in the VisDrone2019 dataset, we also present another case in the UAVDT dataset in Figure 5c. Overall, our experimental results demonstrate the superiority and effectiveness of our proposed method compared to FairMOT.

We also present more qualitative results on the VisDrone2019 [11] test set (refer to Figure 6) and the UAVDT [12] test set (refer to Figure 7) in this section. It can be observed that our approach is effective in detecting objects at different scales (even with small objects) and maintaining their identity correctly. Moreover, it can be noticed that the proposed tracker performs robustly in a variety of scenarios (during the daytime and nighttime and over commercial streets and intersections) and performs well even in crowded scenarios.

(**a**) uav0000119_02301_v

(**b**) uav0000073_00600_v



(**c**) M1007

Figure 5. Robustness analysis of the proposed method compared with FairMOT. The yellow arrows and red dashed boxes mark the missed objects.

Figure 6. Qualitative results of the proposed method on VisDrone2019 test-dev set. The different colored bounding boxes represent different identities, and the frame number is displayed in the upper-left corner of each frame. Best viewed in color and zoomed in.

Figure 7. Qualitative results of the proposed method on UAVDT test set. The different colored bounding boxes represent different identities, and the frame number is displayed in the upper-left corner of each frame. Best viewed in color and zoom in.

5. Conclusions

In this paper, we propose a novel method for tracking multiple objects in UAV videos that fully utilizes both temporal and spatial information. Our approach incorporates

a novel temporal feature aggregation module (TFAM), which effectively incorporates temporal context to improve the accuracy and consistency of the tracker's perception ability. Additionally, we introduce a topology-integrated embedding module (TIEM), which captures topological relationships between objects and their environments, resulting in the enhanced discriminative power of the object embedding features. Through extensive experiments on the VisDrone2019 and UAVDT benchmarks, we demonstrate that our approach achieves state-of-the-art performance.

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

Funding: This work was supported by the National Key Research and Development Program of China under Grant No. 2021YFB0300101.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available through the open access datasets VisDrone2019 [11] and UAVDT [12].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Luo, W.; Xing, J.; Milan, A.; Zhang, X.; Liu, W.; Kim, T.K. Multiple object tracking: A literature review. *Artif. Intell.* **2021**, *293*, 103448. [CrossRef]
2. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A benchmark for multi-object tracking. *arXiv* **2016**, arXiv:1603.00831.
3. Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; Leal-Taixé, L. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv* **2020**, arXiv:2003.09003.
4. Dendorfer, P.; Osep, A.; Milan, A.; Schindler, K.; Cremers, D.; Reid, I.; Roth, S.; Leal-Taixé, L. Motchallenge: A benchmark for single-camera multiple target tracking. *Int. J. Comput. Vis.* **2021**, *129*, 845–881. [CrossRef]
5. Wang, F.; Luo, L.; Zhu, E. Two-stage real-time multi-object tracking with candidate selection. In *MMM 2021: MultiMedia Modeling, Proceedings of the International Conference on Multimedia Modeling, Prague, Czech Republic, 22–24 June 2021*; Springer: Cham, Switzerland, 2021; pp. 49–61.
6. Filkin, T.; Sliusar, N.; Ritzkowski, M.; Huber-Humer, M. Unmanned aerial vehicles for operational monitoring of landfills. *Drones* **2021**, *5*, 125. [CrossRef]
7. Fan, J.; Yang, X.; Lu, R.; Xie, X.; Li, W. Design and implementation of intelligent inspection and alarm flight system for epidemic prevention. *Drones* **2021**, *5*, 68. [CrossRef]
8. Svanström, F.; Alonso-Fernandez, F.; Englund, C. Drone Detection and Tracking in Real-Time by Fusion of Different Sensing Modalities. *Drones* **2022**, *6*, 317. [CrossRef]
9. Dewangan, V.; Saxena, A.; Thakur, R.; Tripathi, S. Application of Image Processing Techniques for UAV Detection Using Deep Learning and Distance-Wise Analysis. *Drones* **2023**, *7*, 174. [CrossRef]
10. Sun, L.; Zhang, J.; Yang, Z.; Fan, B. A Motion-Aware Siamese Framework for Unmanned Aerial Vehicle Tracking. *Drones* **2023**, *7*, 153. [CrossRef]
11. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Hu, Q.; Ling, H. Vision meets drones: Past, present and future. *arXiv* **2020**, arXiv:2001.06303 .
12. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–386.
13. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.
14. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.
15. Long, C.; Haizhou, A.; Zijie, Z.; Chong, S. Real-time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-identification. In Proceedings of the ICME, San Diego, CA, USA, 23–27 July 2018.
16. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
17. Yang, K.; Li, D.; Dou, Y. Towards precise end-to-end weakly supervised object detection network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8372–8381.

18. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.
19. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
20. Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; Gu, J. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Trans. Multimed.* **2019**, *22*, 2597–2609. [CrossRef]
21. Bergmann, P.; Meinhardt, T.; Leal-Taixe, L. Tracking without bells and whistles. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 941–951.
22. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3069–3087. [CrossRef]
23. Brasó, G.; Leal-Taixé, L. Learning a neural solver for multiple object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6247–6257.
24. Weng, X.; Wang, Y.; Man, Y.; Kitani, K.M. Gnn3dmot: Graph neural network for 3d multi-object tracking with 2d-3d multi-feature learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6499–6508.
25. Wang, Y.; Kitani, K.; Weng, X. Joint object detection and multi-object tracking with graph neural networks. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13708–13715.
26. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
27. Zhang, L.; Li, Y.; Nevatia, R. Global data association for multi-object tracking using network flows. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
28. Lan, L.; Tao, D.; Gong, C.; Guan, N.; Luo, Z. Online Multi-Object Tracking by Quadratic Pseudo-Boolean Optimization. In Proceedings of the IJCAI, New York, NY, USA, 9–15 July 2016; pp. 3396–3402.
29. Bochinski, E.; Eiselein, V.; Sikora, T. High-speed tracking-by-detection without using image information. In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–6.
30. Kalman, R.E. Contributions to the theory of optimal control. *Bol. Soc. Mat. Mex.* **1960**, *5*, 102–119.
31. Tang, Z.; Hwang, J.N. Moana: An online learned adaptive appearance model for robust multiple object tracking in 3d. *IEEE Access* **2019**, *7*, 31934–31945. [CrossRef]
32. Wang, G.; Wang, Y.; Gu, R.; Hu, W.; Hwang, J.N. Split and connect: A universal tracklet booster for multi-object tracking. *IEEE Trans. Multimed.* **2023,** *25*, 1256–1268. [CrossRef]
33. Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; Wang, S. Towards real-time multi-object tracking. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 107–122.
34. Wang, Q.; Zheng, Y.; Pan, P.; Xu, Y. Multiple object tracking with correlation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3876–3886.
35. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
36. Zhou, X.; Koltun, V.; Krähenbühl, P. Tracking objects as points. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 474–490.
37. Peng, J.; Wang, C.; Wan, F.; Wu, Y.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Fu, Y. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 145–161.
38. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Detect to track and track to detect. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3038–3046.
39. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
40. Guo, S.; Wang, J.; Wang, X.; Tao, D. Online Multiple Object Tracking with Cross-Task Synergy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
41. Sun, P.; Cao, J.; Jiang, Y.; Zhang, R.; Xie, E.; Yuan, Z.; Wang, C.; Luo, P. Transtrack: Multiple object tracking with transformer. *arXiv* **2020**, arXiv:2012.15460.
42. Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; Feichtenhofer, C. TrackFormer: Multi-Object Tracking with Transformers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
43. Zeng, F.; Dong, B.; Zhang, Y.; Wang, T.; Zhang, X.; Wei, Y. MOTR: End-to-End Multiple-Object Tracking with TRansformer. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2022.
44. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
45. Cai, J.; Xu, M.; Li, W.; Xiong, Y.; Xia, W.; Tu, Z.; Soatto, S. MeMOT: Multi-object tracking with memory. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8090–8100.
46. Hornakova, A.; Henschel, R.; Rosenhahn, B.; Swoboda, P. Lifted disjoint paths with application in multiple object tracking. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual, 13–18 July 2020; pp. 4364–4375.
47. Xu, J.; Cao, Y.; Zhang, Z.; Hu, H. Spatial-temporal relation networks for multi-object tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3988–3998.

48. He, J.; Huang, Z.; Wang, N.; Zhang, Z. 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, Nashville, TN, USA, 20–25 June 2021; pp. 5299–5309.
49. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep layer aggregation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2403–2412.
50. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [CrossRef]
51. Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; Brox, T. Flownet: Learning optical flow with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2758–2766.
52. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
53. Xia, Z.; Pan, X.; Song, S.; Li, L.E.; Huang, G. Vision transformer with deformable attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4794–4803.
54. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
55. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7482–7491.
56. Bernardin, K.; Stiefelhagen, R. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 246309. [CrossRef]
57. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 17–35.
58. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
59. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
60. Pirsiavash, H.; Ramanan, D.; Fowlkes, C.C. Globally-optimal greedy algorithms for tracking a variable number of objects. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1201–1208.
61. Milan, A.; Roth, S.; Schindler, K. Continuous energy minimization for multitarget tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 58–72. [CrossRef] [PubMed]
62. Dicle, C.; Camps, O.I.; Sznaier, M. The way they move: Tracking multiple targets with similar appearance. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2304–2311.
63. Bae, S.H.; Yoon, K.J. 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, Columbus, OH, USA, 23–28 June 2014; pp. 1218–1225.