remote sensing

MDPI

Article

An Asymmetric Feature Enhancement Network for Multiple Object Tracking of Unmanned Aerial Vehicle

Jianbo Ma [1,2,3,4], Dongxu Liu [1,2,3], Senlin Qin [1,2,3,4], Ge Jia [1,2,3], Jianlin Zhang [1,2,3] and Zhiyong Xu [1,2,3,*]

1 National Key Laboratory of Optical Field Manipulation Science and Technology, Chinese Academy of Sciences, Chengdu 610209, China; majianbo22@mails.ucas.ac.cn (J.M.); liudongxu18@mails.ucas.ac.cn (D.L.); qinsenlin22@mails.ucas.ac.cn (S.Q.); jiage@ioe.ac.cn (G.J.); jlin@ioe.ac.cn (J.Z.)
2 Key Laboratory of Optical Engineering, Chinese Academy of Sciences, Chengdu 610209, China
3 Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China
4 University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: xuzhiyong@ioe.ac.cn

Abstract: Multiple object tracking (MOT) in videos captured by unmanned aerial vehicle (UAV) is a fundamental aspect of computer vision. Recently, the one-shot tracking paradigm integrates the detection and re-identification (ReID) tasks, striking a balance between tracking accuracy and inference speed. This paradigm alleviates task conflicts and achieves remarkable results through various feature decoupling methods. However, in challenging scenarios like drone movements, lighting changes and object occlusion, it still encounters issues with detection failures and identity switches. In addition, traditional feature decoupling methods directly employ channel-based attention to decompose the detection and ReID branches, without a meticulous consideration of the specific requirements of each branch. To address the above problems, we introduce an asymmetric feature enhancement network with a global coordinate-aware enhancement (GCAE) module and an embedding feature aggregation (EFA) module, aiming to optimize the two branches independently. On the one hand, we develop the GCAE module for the detection branch, which effectively merges rich semantic information within the feature space to improve detection accuracy. On the other hand, we introduce the EFA module for the ReID branch, which highlights the significance of pixel-level features and acquires discriminative identity embedding through a local feature aggregation strategy. By efficiently incorporating the GCAE and EFA modules into the one-shot tracking pipeline, we present a novel MOT framework, named AsyUAV. Extensive experiments have demonstrated the effectiveness of our proposed AsyUAV. In particular, it achieves a MOTA of 38.3% and IDF1 of 51.7% on VisDrone2019, and a MOTA of 48.0% and IDF1 of 67.5% on UAVDT, outperforming existing state-of-the-art trackers.

Keywords: multiple object tracking; data association; feature enhancement; unmanned aerial vehicle

Citation: Ma, J.; Liu, D.; Qin, S.; Jia, G.; Zhang, J.; Xu, Z. An Asymmetric Feature Enhancement Network for Multiple Object Tracking of Unmanned Aerial Vehicle. *Remote Sens.* **2024**, *16*, 70. https://doi.org/10.3390/rs16010070

Academic Editor: Pedro Melo-Pinto

Received: 30 October 2023
Revised: 18 December 2023
Accepted: 21 December 2023
Published: 23 December 2023

1. Introduction

Multiple Object Tracking (MOT) focuses on the identification and tracking of multiple targets belonging to different classes in a video sequence. With the widespread utilization of MOT task and advances in navigation technology, the MOT of unmanned aerial vehicle (UAV) has emerged as a vital research area in remote sensing [1–3]. This technology is increasingly employed for applications in agriculture, monitoring and emergency rescue. Nevertheless, the MOT task, especially from the perspective of UAV, presents a multitude of challenges. The UAV-based MOT encounters issues such as image deterioration, alterations in scale, ambiguous object brightness and immediate tracking [4–6].

In recent years, deep learning methods have been widely adopted to address the obstacles encountered in MOT. Due to the specific requirements of the MOT task, it is necessary for the relevant work to establish a unique trajectory for each tracked target in the video sequence. Concretely, the tracking methods can be broadly categorized into two paradigms, the two-stage tracker [7–9] and the one-shot tracker [10–12]. Although the two-stage tracker has exhibited significant improvements in detection performance owing to the superior quality of detection results as reported in previous studies [13–15], it still requires a distinct training phase for a feature extraction module that can handle re-identification (ReID) information. In other words, the two-stage tracker divides detection and feature extraction into two independent steps, which is a complex and time-consuming tracking framework.

To eliminate the independence of the two processes, an effective solution is to construct a federated framework that combines the detection and ReID tasks into a cohesive model, providing simultaneous output from the detection and ReID head. In particular, since the introduction of JDE [10], which pioneered the concept of the joint detection and embedding paradigm, a range of one-shot trackers have emerged to enhance performance based on this framework. Many researchers recognize the training conflict between the two tasks and propose various efficient feature deconstruction techniques to address the prevalent feature misalignment issue in the one-shot tracking algorithms [16–18]. Specifically, the detection branch is essential for enhancing inter-class variability, whereas the ReID branch plays a crucial role in improving intra-class discrimination. Utilizing the shared feature map for both tasks invariably leads to sub-optimal network optimization. When the UAV is hovering at a considerable altitude from the ground, the tracker's task is to identify targets with small pixel areas and accurately locate them to extract specific characteristics. As a result, the one-shot tracking approach continues to suffer from performance degradation in UAV videos when it solely relies on channel-based decoupling to address feature conflicts between the two branches, without taking into account the task-specific demands.

Based on the above analysis, we introduce a novel asymmetric feature enhancement framework for the MOT task. Similar to the previous methodologies, we maintain the belief that using specific feature maps to represent the detection and ReID branches will facilitate the learning of task-independent representations. Nevertheless, instead of relying on channel-based attention for feature decoupling, we concentrate on feature optimization according to the diverse demands of each branch. The detection branch necessitates outstanding recognition capabilities to accurately identify specific categories in complex and diverse scenarios. To address this, we introduce the custom-designed global coordinate-aware enhancement (GCAE) module specifically for this branch. The GCAE module hierarchically encodes the shared feature map in the feature space along different directions, thus improving the representation of the object of interest in a more effective manner.

For the ReID branch, it is essential to extract the identity embedding of detected targets based on their central position. Especially in situations involving dense or small objects, having more precise position information proves invaluable for embedding extraction. We contend that local information related to potential targets is more conducive to aggregating ReID features. Consequently, we adopt an embedding feature aggregation (EFA) module to emphasize the significance of feature information at the pixel level. The EFA module employs the pre-existing detection heatmap and pseudo-Gaussian heatmap as prior conditions to guide the generation of the more distinctive task-related feature map.

By deliberately treating detection and embedding features as two distinct task optimizations, AsyUAV improves detection accuracy and extracts dependable identity embedding from the UAV viewpoint. Experiments indicate that the proposed modules achieve remarkable tracking and association performance, demonstrating their superiority on two public datasets. To provide a comprehensive summary, we outline the key contributions of this paper as follows:

- We propose a global coordinate-aware enhancement (GCAE) module within the detection branch to enable interactions over long distances and improve recognition capability.
- We introduce an embedding feature aggregation (EFA) module that applies prior spatial attention to the ReID branch. By fusing pre-existing feature information, the generated ReID-specified feature map effectively mitigates background interference and enhances robustness against changing views observed from the UAV.

- An asymmetric feature enhancement network, comprising the GCAE and EFA modules, has been seamlessly integrated into the one-shot MOT framework for UAV-based tracking, referred to as AsyUAV. All experiments show that compared with the current leading MOT techniques on VisDrone2019 and UAVDT datasets, AsyUAV obtains competitive performance in terms of detection accuracy and the acquisition of discriminative identity features, and achieves the best results on MOTA and IDF1 metrics (Figure 1).

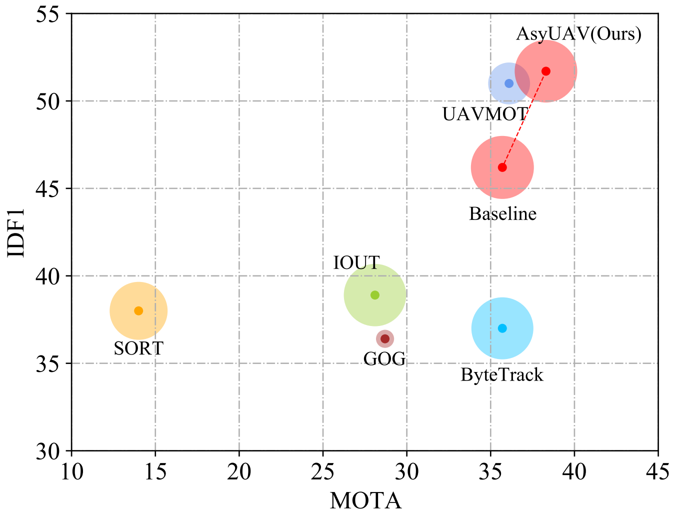

Figure 1. Comparisons of preceding state-of-the-art trackers with our proposed AsyUAV on the VisDrone test set using IDF1-MOTA-FPS measurement metrics. The x-axis represents MOTA, while the y-axis represents IDF1, and the size of the circle denotes FPS. Our AsyUAV demonstrates superior performance in terms of MOTA, IDF1, and competitive tracking speed (FPS).

2. Related Work

In this section, we describe the research on common paradigms for multiple object tracking and techniques to improve tracking performance. For further elaboration, Section 2.1 discusses the multiple object tracking task, Section 2.2 introduces the feature enhancement method and Section 2.3 covers the data association strategy.

2.1. Multiple Object Tracking

In recent years, the convenience and flexibility of drone platform cameras have sparked growing interest among researchers in the field of MOT for UAV video. Different from the video detection task, the MOT task demands the continuous tracking of specific object categories throughout a video sequence, with each tracked object maintaining a unique identity number across numerous frames. Leveraging the efficient feature representation capabilities of convolutional neural networks, the MOT algorithms have made remarkable advancements. One prevalent approach for MOT involves the tracking by detection paradigm, where targets are first detected and then associated based on their appearance and motion cues.

For example, SORT [19] uses a deep learning-based detector (Fast R-CNN [20]) to identify potential targets in each frame. Building on the foundation laid by SORT, DeepSORT [21] incorporates a pre-trained feature extraction model for creating discriminative appearance embedding and employs a deliberate cascade matching method to elevate the tracking performance of MOT. BOT-SORT [13] introduces a highly reliable tracking system by integrating camera motion-compensated features and employing a suitable Kalman filter [22] state vector for precise box localization. However, despite its achievement in reaching state-of-the-art performance, the inherent complexity of the tracking process inevitably leads to reduced speed.

Fortunately, the advent of the one-shot paradigm has struck a balance between accuracy and speed. This paradigm combines object detection with the extraction of corresponding embeddings into a unified framework, offering an alternative solution for the MOT task. For instance, JDE [10] accomplishes the training of ReID and detection tasks within a shared YOLOv3 model. This approach substantially reduces computational overhead and serves as a straightforward and efficient baseline for designing real-time MOT frameworks. FairMOT [23] addresses the issue of an individual anchor being assigned to multiple targets or multiple anchors being assigned to one target, which is a problem arising from the use of anchor-based methods. As a solution, an anchor-free tracker based on the CenterNet [24] detector is proposed.

Recently, following the triumph of the Transformer architecture in computer vision [25], researchers are exploring the integration of identity information from previous frames to develop a Transformer-based MOT framework. For example, TransCenter [26] presents a Transformer-based architecture for MOT which is centered around the objects. TransMOT [27], on the other hand, employs a spatial-temporal graph Transformer to capture the spatial-temporal relationships among objects. TrackFormer [28] concurrently handles object detection and track formation by employing an encoder-decoder Transformer structure. MOTR [29] introduces the concept of a "track query" and uses each track query to represent a unique trajectory throughout the entire video. Importantly, it eliminates the need for post-processing in the association, making it the first end-to-end MOT tracker.

Although there are various multiple object tracking paradigms, the tracking by detection paradigm involves a complex tracking process and Transformer-based MOT methods require a large number of learning parameters. In contrast, the one-shot tracker offers a more balanced trade-off between model complexity and inference time, making it more practical for real-world applications. Consequently, our method adheres to the one-shot tracker paradigm.

2.2. Feature Enhancement

Feature enhancement plays a critical role in MOT, particularly within the one-shot tracking paradigm. A primary issue often encountered is the optimization conflicts arising from distinct feature demands between the detection and ReID branches. In particular, CSTrack [30] spotlight the inherent competition between object detection and ReID. It introduces a reciprocal network with self-attention and cross-attention mechanisms, enabling both tasks to effectively fulfil their unique feature requirements. Taking inspiration from the excellent work of CStrack and aiming to address the contradiction and disentangle the learning of detection and ReID features, various methods have been introduced. For instance, RelationTrack [31] develops a global context decoupling module, MOTFR [29] proposes a locally shared information decoupling module, FPUAV [16] introduces a novel feature decoupling network, and FDTrack [32] designs a mutual inhibition decoupling module. These approaches collectively contribute to resolving the feature optimization conflict between detection and ReID tasks.

Simultaneously, researchers have shown a growing interest in enhancing the feature representation of the ReID branch. They argue that harnessing distinctive identity features can lead to a more robust association process [31,33]. DcMOT [34] presents a multi-attention feature learning module, known as recurrent across-channel attention with spatial attention, to improve the discriminative power of the ReID task in the one-shot MOT framework. MOTFR [29] and RelationTrack [31] both utilize an enhanced attention mechanism to capture global contextual relationships on the ReID feature map. UAVMOT [35] incorporates the embedding feature from two consecutive frames to update and improve the representation of object identity, effectively adapting to changes in the UAV perspective. In addition, OMC [36] amplifies the role of the encoding feature, which uses the identity information existing in the previous frame and the feature map of the current frame to perform a cross-correlation operator to rectify misclassified targets.

The above analysis highlights the critical importance of enhancing relevant feature representation. Previous scholars have made notable advancements in the one-shot tracking paradigm, resulting in significant performance improvements. However, their improvement strategies primarily rely on symmetric network structures to address feature conflicts or incorporate spatio-temporal attention into the ReID branch, often without careful consideration of the specific requirements of each branch. This oversight can lead to suboptimal results as it neglects the distinct demands of both branches. Unlike the previously mentioned methods, we take into account that the detection branch needs to handle diverse multi-category and multi-scale scenes. Simultaneously, the ReID branch requires precise feature extraction positions and robust embedding to represent the unique information of each associated target. Therefore, an asymmetric network enhancement strategy is further proposed to systematically enhance feature representation in both the detection and ReID branches. By carefully considering the specific requirements of each task and optimizing them proactively, our tracker is better equipped to handle complex and dynamic scenarios.

2.3. Data Association

In contrast to single-image detection, in the case of MOT, the objective is to track objects across a video sequence. Therefore, data association becomes a critical step in establishing connections between frames.

Current methods rely on both motion and appearance cues for association. Initially, SORT [19] makes use of motion information by predicting the location of matched tracks in the current frame with the help of the Kalman filter [22]. Subsequently, Intersection over Union (IoU) is employed as the similarity metric to complete data association between the bounding boxes of detected targets and the matched tracks. ByteTrack [37] is a straightforward yet highly effective tracker that employs a low-scoring detection strategy and depends on the IoU association to deliver outstanding performance in pedestrian datasets. MAT [38] focuses on high frame rate motion modelling and provides a range of solutions for trajectory prediction, reconnection, and matching over extended time intervals. Additionally, researchers have developed deep networks as replacements for various filter algorithms in learning object motion, thereby achieving high-performance tracking results [39,40].

However, the motion model is notably reliant on a high frame rate and tends to be sensitive when it comes to long-range association. In scenarios involving camera motion and low frame rates, the motion cue offers no clear advantage over the appearance cue. Specifically, the appearance information bears similarity to the ReID task in pedestrian re-identification [41]. When the viewpoint of the drone alters, the ReID-based approach can globally search for optimal matching results, thus enhancing the data association in MOT. For the training of the ReID task, there are two different feature extraction methods. One of them uses an additional feature extraction network to extract the identity embedding from the bounding box of the detected object, renowned as the tracking by detection paradigm [21]. Conversely, the alternative approach involves the joint learning of detection and ReID branches, known as the one-shot tracking paradigm [10].

These two association cues possess their individual benefits. We argue that the motion-based strategy is better suited for local-scale matching, whereas the appearance-based strategy is more appropriate for global-scale matching. Both motion-based and appearance-based matching are important for tracking in the UAV video. In situations where the camera undergoes sudden movement, appearance information proves to be a more reliable association cue. When encountering targets that are blurred or obstructed, motion information becomes the more dependable cue. Therefore, it is unquestionably important to leverage both cues to address various scenarios.

3. Methodology

This section provides a detailed explanation of our methodology. The overall framework of our tracker is described in Section 3.1. Next, we proceed to introduce the global coordinate-aware enhancement (GCAE) module and embedding feature aggregation (EFA) module in Section 3.2 and Section 3.3, respectively. Lastly, Section 3.4 provides a brief description of the online matching strategy.

3.1. Overall Framework

When UAV-captured images are fed into the MOT pipeline, our AsyUAV is designed to accurately detect and robustly track designated object categories across successive frames. The comprehensive structure of AsyUAV is depicted in Figure 2. First, individual frames from the video sequence are transmitted in sequence to AsyUAV, which then generates the shared feature map through the backbone. Subsequently, we employ an asymmetric feature enhancement network to selectively optimize both the detection branch and the ReID branch. This asymmetric feature enhancement network comprises the GCAE module for the detection branch and the EFA module for the ReID branch. GCAE is designed to enhance detection performance, while EFA is aimed at extracting more discriminative identity features. After obtaining the detection-specified feature map and ReID-specified feature map using the two proposed modules, respectively, we further derive the object bounding box, heatmap, and identity embedding through a 1×1 convolutional layer. Finally, these outputs are matched online with the previous trajectory.

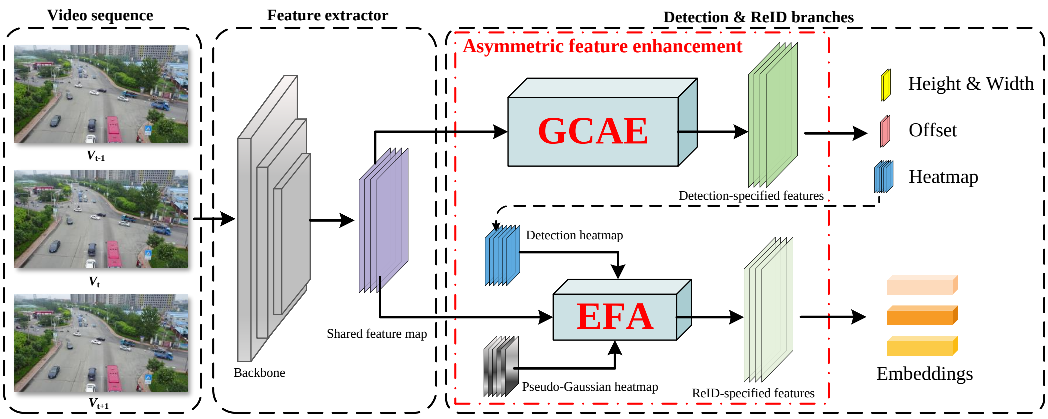

Figure 2. The overall framework of AsyUAV. After generating shared features through the backbone, we utilize an asymmetric feature enhancement network to produce detection-specific and ReID-specific features. The proposed asymmetric feature enhancement consists of a GCAE module and an EFA module. The GCAE module is implemented in the detection branch to improve detection accuracy, while the EFA module is employed in the ReID branch to enhance embedding discriminability.

3.2. Global Coordinate-Aware Enhancement

The effectiveness of the attention mechanism for enhancing network learning is well-established. However, attention constrained to specific regions poses difficulties in capturing interactive information across a global context. It is worth noting that global information helps the network comprehend the context and extract useful features. Especially in object detection task, it is crucial to understand the global positional relationship between targets [42]. Therefore, with the goal of strengthening the representation power of the detection feature map and enhancing detection accuracy, we introduce the GCAE module, a pivotal component of the asymmetric feature enhancement network.

The specific structure of the GCAE module is illustrated in Figure 3. We designate the shared feature map output from the feature extractor part as $\mathbf{F^s} \in \mathbb{R}^{C \times H \times W}$. The GCAE module takes it as input and generates the detection-specified feature map $\mathbf{F^d} \in \mathbb{R}^{C \times H \times W}$ with amplified depiction. In particular, we employ global max pooling and global average pooling along both the row direction and column direction of $\mathbf{F^s}$ to obtain channel-based global coordinate attention. The output of the c-th channel through global max pooling can be defined as:

$$GMP_c^{row}(h) = max(\mathbf{F}_c^{\mathbf{s}}[h, 0 : W]) \tag{1}$$

$$GMP_c^{col}(w) = max(\mathbf{F}_c^{\mathbf{s}}[0 : H, w]) \tag{2}$$

where GMP_c^{row} and GMP_c^{col} represent the global max pooling at the row (row) direction and column (col) direction, respectively. Similarly, the output of the c-th channel through global average pooling can be defined as

$$GAP_c^{row}(h) = \frac{1}{W} \sum_{i=0}^{W-1} \mathbf{F}_c^{\mathbf{s}}(h, i) \tag{3}$$

$$GAP_c^{col}(w) = \frac{1}{H} \sum_{j=0}^{H-1} \mathbf{F}_c^{\mathbf{s}}(j, w) \tag{4}$$

where GAP_c^{row} and GAP_c^{col} represent the global average pooling at the row (row) direction and column (col) direction, respectively.

The dimensions of the four tensors obtained through the different pooling operations are $GMP^{row} \in \mathbb{R}^{C\times H}$, $GMP^{col} \in \mathbb{R}^{C\times W}$, $GAP^{row} \in \mathbb{R}^{C\times H}$, and $GAP^{col} \in \mathbb{R}^{C\times W}$, respectively. Then, we adopt the following equation to produce the global coordinate-aware channel information $\mathbf{Z} \in \mathbb{R}^{C\times[2*(H+W)]}$.

$$\mathbf{Z} = [GMP^{row}, GMP^{col}, GAP^{row}, GAP^{col}] \tag{5}$$

Recall that $[\cdot,\cdot,\cdot,\cdot]$ denotes the concatenation operation along the channel dimension. Fusing the four tensors with distinct spatial orientations and contextual details enables $\mathbf{Z}$ to capture comprehensive interdependencies in the spatial dimension while maintaining sensitivity to the region of interest.

After that, we obtain the refined global coordinate-aware channel information, denoted as $\hat{\mathbf{Z}} \in \mathbb{R}^{C\times(H+W)}$. This is accomplished through a tailored transformation, which adaptively reweights position information with diverse characteristics while reducing dimensionality to align with the original feature dimension. The procedure is outlined as follows

$$\hat{\mathbf{Z}} = ReLU(\Psi_{bn}(Conv(W_z\mathbf{Z}))) \tag{6}$$

Here, W_z is a learnable matrix, and $Conv(\cdot)$ denotes a standard 1×1 convolutional layer. Additionally, $\Psi_{bn}(\cdot)$ and $ReLU(\cdot)$ correspond to the rectified linear unit and batch normalization operator, respectively. Later, we split $\hat{\mathbf{Z}}$ into two distinct vectors $\hat{\mathbf{Z}}^H \in \mathbb{R}^{C\times H}$ and $\hat{\mathbf{Z}}^W \in \mathbb{R}^{C\times W}$ along the height and width directions. Two additional non-shared 1×1 convolutional layers with non-linear activation function $\sigma(\cdot)$ are implemented to $\mathbf{Z}^H$ and $\mathbf{Z}^W$, respectively. Considering that

$$\mathbf{g}^H = \sigma(Conv(\hat{\mathbf{Z}}^H)) \tag{7}$$

$$\mathbf{g}^W = \sigma(Conv(\hat{\mathbf{Z}}^W)) \tag{8}$$

Lastly, $\mathbf{g}^H$ is expanded horizontally and $\mathbf{g}^W$ is expanded vertically. The resulting detection-specific feature map, denoted as $\mathbf{F}^{\mathbf{d}}$, is obtained through the following numerical operations

$$\mathbf{F}_c^{\mathbf{d}}(i, j) = \mathbf{F}_c^{\mathbf{s}}(i, j) \times \mathbf{g}_c^H(i) \times \mathbf{g}_c^W(j) + \mathbf{F}_c^{\mathbf{s}}(i, j) \tag{9}$$

The proposed GCAE module is capable of allowing the detection branch to prioritize regions of interest for objects located on the feature map. In contrast to the standard global pooling process, which merely condenses global spatial information into a channel vector, our approach takes into account both the direction and type of pooling operation. This allows the detection-specific feature map to establish interactions over long distances, thereby assisting the detection head in recognizing and locating targets.

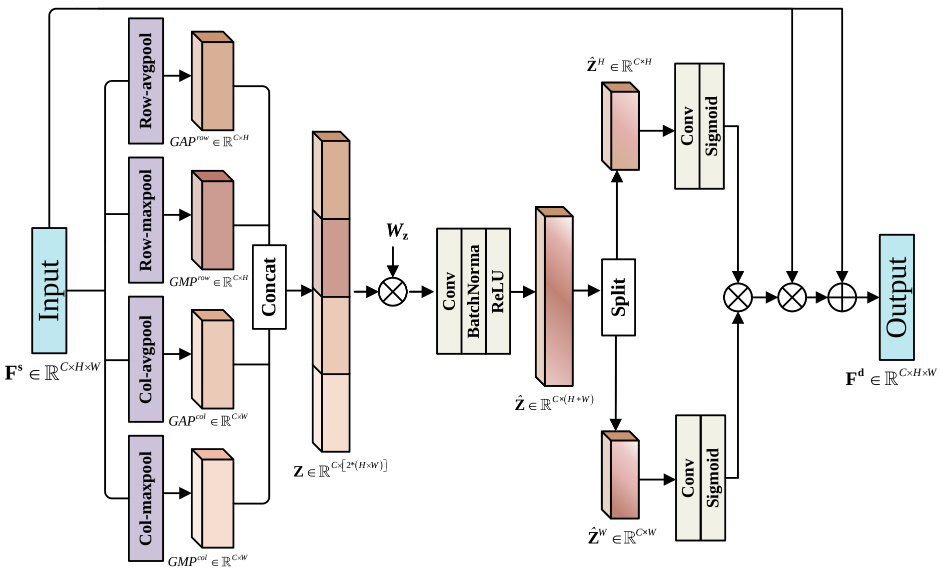

Figure 3. The specific structure of the Global Coordinate-Aware Enhancement (GCAE) module. W_z is a learnable matrix. "Col-maxpool", "Col-avgpool", "Row-maxpool" and "Row-avgpool" represent global max pooling at the column direction, global average pooling at the column direction, global max pooling at the row direction and global average pooling at the row direction. "$\otimes$" means element-wise multiplication and "$\oplus$" means element-wise summation.

3.3. Embedding Feature Aggregation

In this section, we introduce a novel EFA module, specifically designed to enhance high-level features for the ReID branch. This module is aimed at creating the more distinctive task-related feature map, and it serves as a crucial component of the asymmetric feature enhancement network.

As depicted in Figure 4, the EFA module takes three inputs, namely detection heatmap $\mathbf{HM^s} \in \mathbb{R}^{K\times H\times W}$, pseudo-Gaussian heatmap $\mathbf{HM^g} \in \mathbb{R}^{K\times H\times W}$ and shared feature map $\mathbf{F^s} \in \mathbb{R}^{C\times H\times W}$. Then, generating the ReID-specific feature map $\mathbf{F^r} \in \mathbb{R}^{C\times H\times W}$. Generally, $\mathbf{HM^s}$ is generated by the detection head and contains the predicted center position of the object. The target category is represented by a specified feature layer of $\mathbf{HM^s}$, with the number of channels K indicating the number of categories. As shown in the first stage of Figure 4, the non-maximum suppression (NMS) algorithm is utilized to eliminate redundant or overlapping targets over the detection heatmap. It is implemented via a standard 3×3 max pooling layer for improved efficiency [24].

Following the NMS operation, which extracts the peak keypoints from each layer of the detection heatmap, we manually proceed to generate the pseudo-Gaussian heatmap $\mathbf{HM^g}$ according to the bounding box annotation. The $\mathbf{HM^g}$ is produced as follows

$$\mathbf{HM}_i^g = exp(-\frac{(x-c_x)^2+(y-c_y)^2}{2(\sigma_p)^2}) \qquad i = \{1,2,3...,K\} \tag{10}$$

where (x, y) represents the pixel coordinates on the $\mathbf{HM^g}$ and (c_x, c_y) denotes the center point of the bounding box annotation. σ_p is the standard deviation value that determines the Gaussian radius. The feature map layers in $\mathbf{HM^g}$ correspond to those in $\mathbf{HM^s}$ and each layer represents an object category.

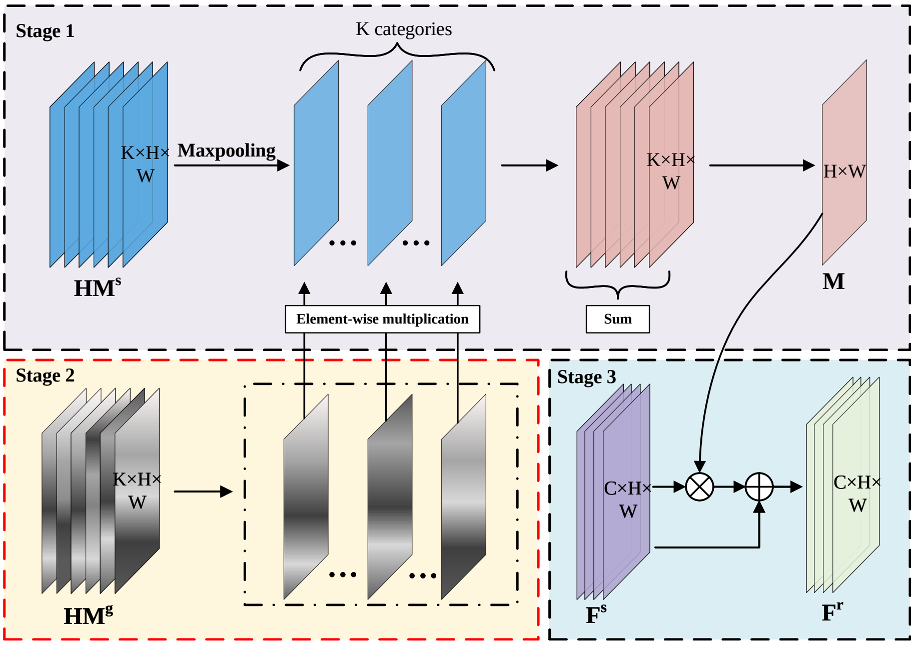

Figure 4. The specific structure of the Embedding Feature Aggregation (EFA) module. The EFA module combines the detection heatmap $\mathbf{HM}^{\mathbf{s}}$ and pseudo-Gaussian heatmap $\mathbf{HM}^{\mathbf{g}}$ to produce aggregated feature map $\mathbf{M}$, which is subsequently employed as an attention weight to generate the ReID-specified feature map $\mathbf{F}^{\mathbf{r}}$. Note that the Stage 2 is only used for training phase.

As shown in the second stage of Figure 4, we utilize the pseudo-Gaussian heatmap to limit the feature representation areas for each layer of the detection heatmap. This way can effectively highlight the region of interest and mitigate background interference. We perform element-wise multiplication and summation along the channel dimension to create a unified aggregated feature map, denoted as $\mathbf{M}$, yielding

$$\mathbf{M} = \sum_{i=1}^{K} NMS(\mathbf{HM}_i^s) \times \mathbf{HM}_i^g \tag{11}$$

We focus on the importance of fine-grained semantic information within $\mathbf{M}$, employing it as prior knowledge to guide the enhancement of feature representation in the ReID branch. The incorporation of pixel-level priors enables the ReID phase to concentrate on potential target areas during the training stage, thus enhancing the discriminative power and robustness of the extracted identity embeddings during the inference stage. Therefore, in the third stage, as depicted in Figure 4, by utilizing $\mathbf{M}$ as an attentional weight, the output of our ReID-specific feature map $\mathbf{F}^{\mathbf{r}}$ can be expressed as

$$\mathbf{F}_c^{\mathbf{r}}(i,j) = \mathbf{M}(i,j) \times \mathbf{F}_c^{\mathbf{s}}(i,j) + \mathbf{F}_c^{\mathbf{s}}(i,j) \tag{12}$$

All in all, we propose the integration of the EFA module during the ReID phase to enhance the reliability of the target identity. In contrast to adaptive optimization methods reliant on deep convolutional networks, we employ a task-guided learning strategy to facilitate pixel-level learning at the ReID branch. During the training phase, both the pseudo-Gaussian and detection heatmaps are simultaneously used as attentional weights in the third stage to update the embedding feature. However, during the inference phase, the pseudo-Gaussian heatmap is not available. This encoding process allows the EFA module to strike a balance between network generalization and its superior performance.

3.4. Online Matching Strategy

In this section, we provide a comprehensive explanation of our matching strategy to associate detected objects across consecutive frames. Our approach is based on the cascade matching strategy introduced in MOTDT [43] and incorporates method from ByteTrack [37] to maximize the use of low-scoring detection results. The pseudocode of the online matching strategy is shown in Algorithm 1.

The association process requires processing a multi-frame video sequence $\boldsymbol{V}$, where each frame contains the detection results $\boldsymbol{Det}_N$ and their corresponding identity embeddings $\boldsymbol{ID}_N$, with N representing the number of detected targets. Additionally, we define two detection thresholds, d_{high} and d_{low}, to categorize $\boldsymbol{Det}_N$. The result of the association process is the set of tracks $\boldsymbol{T}$.

For each frame, we employ the Kalman filter to predict the positional status of the tracks in the current frame (lines 3 to 4). We initialize both the low-scoring detection results $\boldsymbol{Det}_{low}$ and the high-scoring detection results $\boldsymbol{Det}_{high}$ along with their corresponding identity information $\boldsymbol{ID}_{high}$, and subsequently categorize the detection results based on the detection threshold (lines 5 to 11).

During the cascade matching stage (lines 12 to 26), we initially utilize appearance-based information to establish associations between $\boldsymbol{T}$ and $\boldsymbol{Det}_{high}$. This entails computing the Mahalanobis distance, denoted as $\mathcal{D}_m$, between the predicted tracks $\boldsymbol{T}$ and the high-scoring detected bounding boxes $\boldsymbol{Det}_{high}$. Then, we combine the Mahalanobis distance with the cosine distance computed from the identity embedding $\boldsymbol{ID}_{high}$, generating the composite distance metric $\mathcal{D}$, which can be expressed as

$$\mathcal{D} = \lambda \mathcal{D}_r + (1 - \lambda)\mathcal{D}_m \tag{13}$$

where λ serves as a weighting parameter and is set to 0.98 in our experiment following the default setting. The Hungarian algorithm with a matching threshold τ is used to determine the matching targets. If $\mathcal{D}$ exceeds τ, the i_{th} detection is deemed successfully associated with the corresponding tracks. Otherwise, we keep the unmatched tracks $\boldsymbol{T}'_{remain}$ and detections $\boldsymbol{Det}'_{remain}$.

Secondly, for the remaining tracks $\boldsymbol{T}'_{remain}$ and detections $\boldsymbol{Det}'_{remain}$, we associate them using Intersection over Union (IOU) distance based on motion information. The second remaining detections from $\boldsymbol{Det}'_{remain}$ is put into $\boldsymbol{Det}''_{remain}$ and the second remaining tracks from $\boldsymbol{T}'_{remain}$ is put into $\boldsymbol{T}''_{remain}$. Last but not least, we update the appearance features of the identified targets in each frame to accommodate appearance variations, which can be described as

$$\boldsymbol{ID}^t_{track} = \epsilon \boldsymbol{ID}^{t-1}_{track} + (1 - \epsilon)\boldsymbol{ID}^t_{detection} \tag{14}$$

where, $\boldsymbol{ID}^t_{track}$ denotes the identity embedding of matched targets in the current frame, $\boldsymbol{ID}^{t-1}_{track}$ denotes the identity embedding of tracks from the previous frame and $\boldsymbol{ID}^t_{detection}$ denotes the identity embedding of matched tracks in the current frame. Additionally, we initialize new tracks for any detections that fail to correspond with previously identified targets (lines 25–26).

For the low-scoring detection results $\boldsymbol{Det}_{low}$, the IoU distance is used between $\boldsymbol{Det}_{low}$ and $\boldsymbol{T}''_{remain}$ to preserve detections that may be subject to severe occlusion or motion blur. These detections are considered background during the cascade matching stage (lines 27–28).

Finally, we store the last remaining tracks $\boldsymbol{T}^{last}_{remain}$ after the entire matching process for 30 frames in case they reappear again (lines 29–30).

Algorithm 1: Pseudo-code of online matching strategy

Input:

- A Video sequence: $\boldsymbol{V}$;
- Detection results: $\boldsymbol{Det}_N$ and corresponding identity embedding: $\boldsymbol{ID}_N$;
- Detection thresholds: d_{high}, d_{low};

Output:

- Tracks: $\boldsymbol{T}$

Initialization: $\boldsymbol{T} \leftarrow \emptyset$;
for *frame* f_k *in* $\boldsymbol{V}$ **do**
    `// Predict location of tracks at` f_{k-1}
    for t *in* $\boldsymbol{T}$ **do**
        $t \leftarrow \textit{Kalman filter}(t)$;
    `// Classify the detection results`
    $\boldsymbol{Det}_{high} \leftarrow \emptyset$; $\boldsymbol{Det}_{low} \leftarrow \emptyset$; $\boldsymbol{ID}_{high} \leftarrow \emptyset$;
    for $\boldsymbol{Det}_i$ *in* $\boldsymbol{Det}_N$ **do**
        if $\boldsymbol{Det}_i.score > d_{high}$ **then**
            $\boldsymbol{Det}_{high} \leftarrow \boldsymbol{Det}_i$;
            $\boldsymbol{ID}_{high} \leftarrow \boldsymbol{ID}_i$;
        else if $\boldsymbol{Det}_i.score > d_{low}$ **then**
            $\boldsymbol{Det}_{low} \leftarrow \boldsymbol{Det}_i$;
    `// Cascade matching tracks and detection results`
    First associate $\boldsymbol{T}$ and $\boldsymbol{Det}_{high}$ using appearance cues;
    $\mathcal{D}_m \leftarrow$ Mahalanobis distance from $\boldsymbol{T}$ and $\boldsymbol{Det}_{high}$;
    $\mathcal{D}_r \leftarrow$ cosine distance from $\boldsymbol{ID}_{high}$;
    $\mathcal{D} \leftarrow \lambda \mathcal{D}_r + (1-\lambda)\mathcal{D}_m$;
    for $\boldsymbol{Det}_i$ *in* $\boldsymbol{Det}_{high}$ **do**
        if $\mathcal{D} > \tau$ **then**
            $\boldsymbol{T} \leftarrow \boldsymbol{T} \cup \boldsymbol{Det}_i$
        else
            $\boldsymbol{Det}'_{remain} \leftarrow$ first remaining results from $\boldsymbol{Det}_{high}$;
            $\boldsymbol{T}'_{remain} \leftarrow$ first remaining tracks from $\boldsymbol{T}$;
    second associate $\boldsymbol{T}'_{remain}$ and $\boldsymbol{Det}'_{remain}$ using IoU distance;
    $\boldsymbol{Det}''_{remain} \leftarrow$ second remaining results from $\boldsymbol{Det}'_{remain}$;
    $\boldsymbol{T}''_{remain} \leftarrow$ second remaining tracks from $\boldsymbol{T}'_{remain}$;
    initialize new tracks ;
    $\boldsymbol{T} \leftarrow \boldsymbol{T} \cup \{\boldsymbol{Det}''_{remain}\}$;
    `// preserve low-score results like BytetTrack`
    associate $\boldsymbol{T}''_{remain}$ and $\boldsymbol{Det}_{low}$ using IoU distance;
    $\boldsymbol{T}^{last}_{remain} \leftarrow$ last remaining tracks from $\boldsymbol{T}''_{remain}$;
    `// delete unmatched tracks`
    $\boldsymbol{T} \leftarrow \boldsymbol{T} \setminus \boldsymbol{T}^{last}_{remain}$;
Return: $\boldsymbol{T}$

4. Experiments

In this section, we provide an overview of the experiments. Sections 4.1 and 4.2 describe the implementation details, datasets and metrics. Section 4.3 compares the tracking performance of AsyUAV with preceding benchmarks. Section 4.4 proves the effectiveness of asymmetric feature enhancement network through ablation study. Section 4.5 and Section 4.6 present analyses and visualizations of various scenarios using the VisDrone2019 and UAVDT datasets.

4.1. Implementation Datails

For training, we choose a variant of DLA-34 pre-trained on the COCO dataset [44] as our backbone and then equipped with proposed modules. Following the common setting [35], we adopt the random crop and multi-scaling strategy as data augmentation. All experiments are conducted on a single GeForce RTX 3090 GPU with a batch size of 12. The network is optimized using the Adam optimizer [45] and is trained for 30 epochs with an initial learning rate of 7×10^{-5}. The learning rate decays by a factor of 10 at the 10th and 20th epochs. Since AsyUAV encompasses both detection and ReID tasks, we employ an uncertainty loss [46] to cater to the requirements of multi-task learning. Specifically, the target heatmap and bounding box size in the detection branch are supervised by focal loss [47] and L1 loss, respectively. The ReID branch is treated as a classification task and is trained using cross-entropy loss.

4.2. Datasets and Metrics

We use MOT datasets captured from a UAV perspective to evaluate the effectiveness of our tracking framework. The primary datasets used are VisDrone2019 [48] and UAVDT [49], both of which are shot in open environments, presenting challenges such as small targets and perspective changes.

The VisDrone2019 dataset consists of 80 annotated video sequences, divided into a training set (56 sequences), a validation set (7 sequences) and a test set (17 sequences). There are 10 categories: pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle. The category and number of objects in each video are randomly distributed. The UAVDT dataset comprises mainly traffic vehicle video under an aerial view. In the MOT task, the dataset is divided into a training set (30 sequences) and a test set (20 sequences). It focuses on three categories: car, truck, and bus.

For a fair comparison, we use the official evaluation toolkits to assess the tracking performance of our algorithm. In the experiment of VisDrone2019, we use all ten categories provided for training and evaluate the tracking performance for five categories, specifically car, bus, truck, pedestrian, and van. In the case of the UAVDT experiment, our analysis is primarily focused on the car category. Furthermore, a series of ablation experiments are performed on the VisDrone2019 validation set to verify the individual modules within AsyUAV, and the FairMOT [23] is adopted as the baseline.

The CLEAR Metrics [50] are widely used to assess the quantitative performance of the MOT algorithm. The common indicators we used are summarized in Table 1. Specifically, we select the MOTA, which focuses on detection performance, and the IDF1, which focuses on tracking performance, as our main evaluation criteria. The MOTA is calculated as:

$$\mathrm{MOTA} = 1 - \frac{\mathrm{FP} + \mathrm{FN} + \mathrm{IDS}}{\mathrm{GT}} \tag{15}$$

where FP is the number of false positives, FN is the number of false negatives, IDS is the number of ID switches and GT denotes the number of ground-truth objects. The IDF1 is calculated as:

$$\mathrm{IDF1} = \frac{2\mathrm{IDFTP}}{2\mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}} \tag{16}$$

where IDTP, IDFP and IDFN are the number of true positives, false positives and false negatives that take into account identity information.

Table 1. Summary of evaluation indicators, where ↑ or ↓ represent better performance for each metric.

Metric	Better	Perfect	Description
MOTA	↑	100%	Multiple object tracking accuracy, see Equation (15).
IDF1	↑	100%	ID F1 Score of the predicted identities, see Equation (16).
MT	↑	100%	Mostly tracked targets, see [51] for details.
ML	↓	0	Mostly lost targets, see [51] for details.
FP	↓	0	The total number of false positives.
FN	↓	0	The total number of false negatives
IDS	↓	0	The number of Identity Switches
IDP	↑	100%	Ratio of IDTP/(IDTP + IDFP).
IDR	↑	100%	Ratio of IDTP/(IDTP + IDFN).
Precision	↑	100%	Ratio of TP/(TP + FP).
Recall	↑	100%	Ratio of TP/(TP + FN).
FPS	↑	-	Processing speed on the benchmark.

4.3. Performance Comparison with Preceding Trackers

To provide a comprehensive comparison, we conducted experiments with AsyUAV on the VisDrone2019 and UAVDT test sets, comparing it with other established algorithms. The results are presented in Tables 2 and 3. For a clearer depiction of the results, we bold the top performance for each indicator.

Table 2. Quantitative comparisons with preceding state-of-the-art methods on the VisDrone2019 test set.

Dataset	Tracker	MOTA	IDF1	MT	ML	FP	FN	IDS	FPS
VisDrone2019	GOG [52]	28.7	36.4	346	836	**17,706**	144657	1387	2.0
	SORT [19]	14.0	38.0	506	545	80,845	112,954	3629	23.5
	IOUT [53]	28.1	38.9	467	670	36,158	126,549	2393	27.3
	MOTR [54]	22.8	41.4	272	825	28,407	147,937	**959**	-
	TrackFormer [28]	25.0	30.5	385	770	25,856	141,526	4840	-
	ByteTrak [37]	35.7	37.0	-	-	21,434	124,042	2168	27.0
	UAVMOT [35]	36.1	51.0	520	574	27,983	115,925	2775	12.0
	AsyUAV(ours)	**38.3**	**51.7**	**671**	**413**	46,392	**93,681**	3954	**27.5**

Table 3. Quantitative comparisons with preceding state-of-the-art methods on the UAVDT test set.

Dataset	Tracker	MOTA	IDF1	MT	ML	FP	FN	IDS	FPS
UAVDT	GOG [52]	35.7	0.3	627	374	62,929	153,336	3104	2.0
	SORT [19]	39.0	43.7	484	400	33,037	172,628	2350	23.5
	IOUT [53]	36.6	23.7	534	357	42,245	163,881	9938	27.3
	DSORT [21]	40.7	58.2	595	358	44,868	155,290	2061	-
	SMOT [55]	33.9	45.0	524	367	57,112	166,528	1752	-
	ByteTrack [37]	41.6	59.1	-	-	**28,819**	189,197	**296**	27.0
	UAVMOT [35]	46.4	67.3	**624**	**221**	66,352	**115,940**	456	12.0
	AsyUAV(ours)	**48.0**	**67.5**	600	310	46,571	130,121	349	**27.5**

(1) Results on VisDrone2019.

As demonstrated in Table 2, the AsyUAV attains the highest performance on various MOT metrics, with a score of 38.3% on MOTA and 51.7% on IDF1.

This highlights that our innovative asymmetric feature enhancement approach markedly boosts both detection and association capabilities. Specifically, the UAVMOT employs an embedding feature update module to enhance the target feature association, although it achieves state-of-the-art performance, the inference speed is lower. In comparison, our AsyUAV demonstrates competitive IDF1 metrics and twice the FPS of UAVMOT. ByteTrack is a two-stage tracker that relies on motion-based matching. Compared with ByteTrack, our AsyUAV outperforms it with a 2.6% (35.7% → 38.3%) increase in MOTA and a 14.7% (37.0% → 51.7%) increase in IDF1. Furthermore, AsyUAV demonstrates the stability and continuity of tracks with the highest MT and the lowest ML values.

(2) Results on UAVDT.

To further evaluate the feasibility of our algorithm, we conducted experiments on the UAVDT test set, and the results are presented in Table 3. The UAVDT dataset encompasses various challenging scenarios, including fast-moving targets, complex backgrounds and viewpoint changes, which impose significant demands on MOT algorithms. The experiment demonstrates that our AsyUAV achieves impressive performance on the UAVDT dataset, with MOTA and IDF1 scores of 48.0% and 67.5%. Notably, it surpasses previously established state-of-the-art methods. For instance, AsyUAV significantly outperforms ByteTrack, resulting in a substantial increase in MOTA from 41.6% to 48.0% and in IDF1 from 59.1% to 67.5%. While UAVMOT is a recently proposed one-shot tracker, our approach surpasses it by 1.6% (46.4% → 48.0%) on MOTA and attains competitive IDF1 scores (67.3% vs. 67.5%).

4.4. Ablation Study

To determine the viability of the AsyUAV framework, we executed ablation experiments on two key modules, namely, the GCAE module and the EFA module. To elaborate, the GCAE module is primarily designed to enhance detection performance whereas the EFA module is geared towards boosting association performance.

(1) Effectiveness of GCAE module.

The GCAE module is responsible for refining the detection-specified feature map generated from the shared feature map produced by the backbone. As observed in the first and second rows of Table 4, when the baseline is equipped with the GCAE module, there is a significant increase in MOTA, from 29.7% to 31.8%, and an IDF1 improvement of 0.8% (47.0% → 47.8%). These enhancements demonstrate the effectiveness of our GCAE module in improving detection performance and facilitating the matching process. This module not only introduces coordinate-sensitive attention to the detection branch but also makes full use of valuable information obtained through different pooling forms.

Table 4. Ablation study for the effectiveness of different proposed modules in AsyUAV. Where "✓" represents the baseline equipped with this module.

Baseline	GCAE	EFA	MOTA	IDF1	IDS
✓			29.7	47.0	1148
✓	✓		31.8	47.8	1149
✓		✓	31.2	49.0	1114
✓	✓	✓	**32.1**	**49.6**	**1111**

To be specific, we apply global average pooling (GAP) and global max pooling (GMP) operations along both the row and column directions of the shared feature map. GAP is employed to gather general information, while GMP concentrates on capturing prominent features. As depicted in Table 5, it is evident that employing GMP and GAP individually leads to excellent strong performance in the Recall and Precision indicators, respectively. Due to their different concerns, we use a concatenation operation and learnable weight to allow the module to select meaningful features adaptively. Different from directly using the summation operation, the concatenation operation yields the best overall performance.

(2) Effectiveness of EFA module.

The EFA module effectively complements and extends the embedding capability of the ReID branch. As shown in the first and third rows of Table 4, the proposed EFA module enhances the tracking performance of the baseline, resulting in a 2.0% (47.0% →

49.0%) increase in IDF1 and a 1.5% (29.7% $\rightarrow$ 31.2%) increase in MOTA. Furthermore, the combined application of EFA and GCAD achieves even more outstanding tracking performance. As shown in the fourth row of Table 4, the AsyUAV (baseline with GCAD and EFA modules) has a MOTA of 32.1%, an IDF1 of 49.6%, and reduces the IDS from 1148 to 1111.

Table 5. Ablation study for the fusion method of pooling operation in GCAE module. Where "Max" means the global max pooling and "Avg" means the global average pooling. "$\oplus$" is the summation operation and "©" is the concatenation operation.

Module	Fusion Method	MOTA	Precision	Recall	IDF1↑
GCAE	Max	30.5	74.8	**48.3**	47.7
	Avg	31.5	76.5	47.7	46.5
	Max$\oplus$Avg	31.2	76.4	47.3	47.5
	Max©Avg	**31.8**	**77.1**	47.6	**47.8**

During the design of the EFA module, optional pseudo-Gaussian prior information is considered during the training phase. To evaluate the feasibility of the Gaussian prior, we conducted ablation experiments on the EFA module and the AsyUAV model. The results are presented in Table 6. The EFA module with Gaussian prior improves the IDF1 by 1.3% (47.8% $\rightarrow$ 49.1%), the IDP by 1.0% (58.4% $\rightarrow$ 59.4%) and the IDR by 1.3% (40.5% $\rightarrow$ 41.8%). Besides, the AsyUAV model with Gaussian prior sees an increase in IDF1 from 48.2% to 49.6%, IDP from 62.2% to 62.9% and IDR from 39.3% to 40.9%. Such results demonstrate that the use of Gaussian prior is beneficial for enhancing the learning of the ReID branch and positively impacts data association.

Table 6. Ablation study for the feasibility of the pseudo-gaussian prior information in the EFA module. Where "✓" indicates that the Gaussian prior is used.

Module	Gaussion Prior	IDF1	IDP	IDR	MOTA
EFA		47.8	58.4	40.5	30.5
	✓	**49.1**	**59.4**	**41.8**	**30.7**
AsyUAV		48.2	62.2	39.3	31.4
	✓	**49.6**	**62.9**	**40.9**	**32.1**

4.5. Case Analysis

(1) Analyse different cases of UAV movement.

To better illustrate the advantages of our tracker in UAV-captured videos, we compare the visualization results of AsyUAV and the baseline under various UAV movements, as shown in Figure 5. When the UAV is hovering in the sky, the size of the objective on the ground will alter. As depicted in Figure 5a, the baseline performs inadequately in such scenarios, leading to many missed detections. Conversely, AsyUAV displays notable adaptability to changes in target size while accurately identifying and tracking targets. Another scenario occurs when the UAV moves forward along the ground, as depicted in Figure 5b. Compared with the baseline, which struggles to detect some small targets, AsyUAV efficiently tracks them. Contrary to good visibility during daylight, tracking at night presents a more challenging task. In Figure 5c, the baseline encounters identity switches due to a sudden change in the UAV's shooting angle, but AsyUAV successfully tracks these vehicles. Moreover, when the UAV and the vehicle are in relatively fast motion, the vehicle in the picture appears blurred. As shown in Figure 5d, the baseline has failure cases where two ID numbers correspond to a single target. On the contrary, these targets can be accurately detected and tracked in AsyUAV.

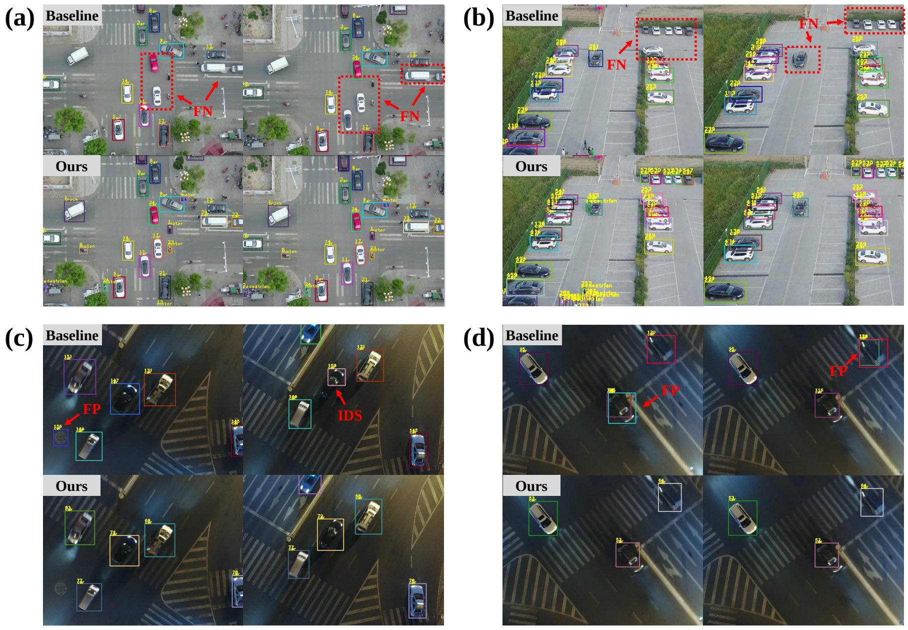

Figure 5. Case analysis of different movements. We list four special cases, including (**a**): UAV hovers up and down, (**b**): UAV moves forward, (**c**): UAV alters orientation and modifies viewpoint, and (**d**): Relatively fast motion between UAV and ground vehicles.

(2) Discuss occlusion by traffic signals.

When the UAV hovers over the intersection, the primary challenge is the temporary occlusion caused by the traffic signals that obstruct all passing vehicles. As shown in Figure 6, both during the daylight and at night, due to the occlusion of traffic signals, the baseline appears to ID switches and false negatives. This observation demonstrates that the baseline is vulnerable to partial occlusion. However, benefiting from the synergy of the GCAE and EFA modules, the AsyUAV can effectively achieve trajectory continuity. The visualization results offer excellent detection performance and robust identity embedding of our tracker in UAV videos.

Figure 6. Visualization results of the impact of traffic signals. Different color bounding boxes represent the identity number of different targets and the dashed lines indicate trajectory continuity.

4.6. Visualization

To show the effectiveness of our method more intuitively, we present visual tracking results of AsyUAV on the VisDrone2019 dataset (Figure 7) as well as the UAVDT (Figure 8) dataset. The AsyUAV performs remarkably in dynamic UAV environments, accurately tracking small and moving targets under varying lighting conditions. The obtained visualisation results exhibit that AsyUAV executes the MOT task proficiently, even when encountering intricate and varied UAV videos.

Figure 7. Qualitative results of AsyUAV on part of VisDrone2019 dataset.

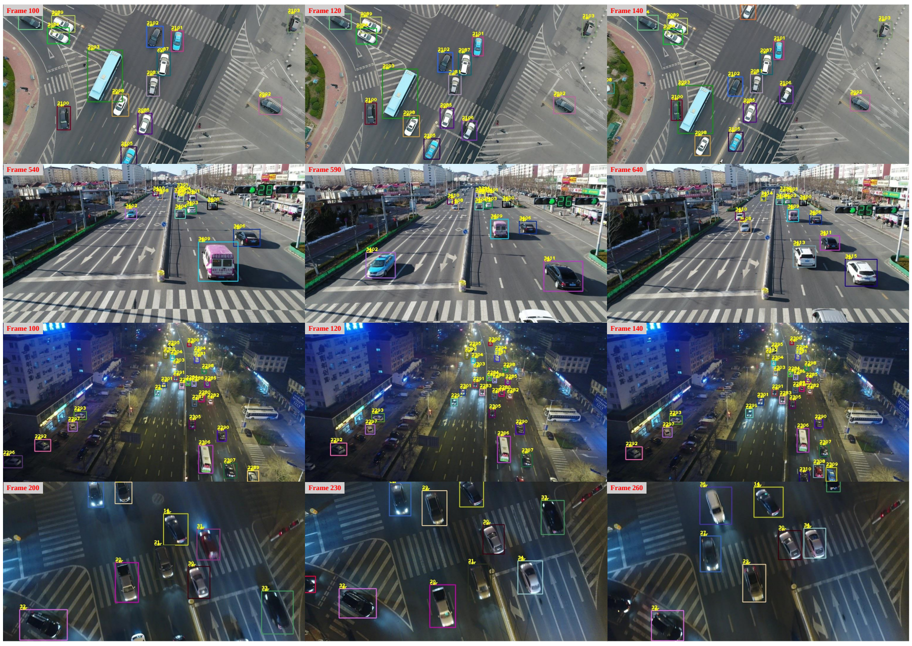

Figure 8. Qualitative results of AsyUAV on part of UAVDT dataset.

5. Conclusions

In this paper, we introduce AsyUAV, a novel one-shot tracker with an asymmetric feature enhancement network for the multiple object tracking task in unmanned aerial vehicle view. We incorporate global coordinate-aware enhancement (GCAE) and embedding feature aggregation (EFA) modules to purposefully reduce competition and promote collaboration between the detection and ReID branches. Additionally, GCAE is dedicated to enhancing detection performance, while EFA is specifically designed to improve association performance. Due to the combination of two components, AsyUAV excels in detecting targets and preserving the continuity of their trajectories, even in challenging scenarios like target occlusion and size changes induced by the unpredictable movement of UAV platform. In comparison with other trackers on the public MOT benchmarks based on the UAV video, our model surpasses them and attains the best performance in both the MOTA and IDF1 metrics. In the future, we intend to delve into leveraging temporal information in videos to enhance discriminative representation learning.

Author Contributions: Conceptualization, J.M.; software, J.M.; validation, J.Z. and Z.X.; formal analysis, J.M. and D.L.; investigation, J.M. and S.Q.; original draft preparation, J.M.; review and editing, J.M., D.L., S.Q., G.J., J.Z. and Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant number 62101529.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

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

References

1. Wu, Z.; Liu, Q.; Zhou, S.; Qiu, S.; Zhang, Z.; Zeng, Y. Learning Template-Constraint Real-Time Siamese Tracker for Drone AI Devices via Concatenation. *Drones* **2023**, *7*, 592. [CrossRef]
2. Avola, D.; Cinque, L.; Diko, A.; Fagioli, A.; Foresti, G.L.; Mecca, A.; Pannone, D.; Piciarelli, C. MS-Faster R-CNN: Multi-Stream Backbone for Improved Faster R-CNN Object Detection and Aerial Tracking from UAV Images. *Remote Sens.* **2021**, *13*, 1670. [CrossRef]
3. Li, X.; Wu, J. Extracting High-Precision Vehicle Motion Data from Unmanned Aerial Vehicle Video Captured under Various Weather Conditions. *Remote Sens.* **2022**, *14*, 5513. [CrossRef]
4. Wang, G.; Song, M.; Hwang, J.N. Recent advances in embedding methods for multi-object tracking: A survey. *arXiv* **2022**, arXiv:2205.10766.
5. Varga, L.A.; Koch, S.; Zell, A. Comprehensive Analysis of the Object Detection Pipeline on UAVs. *Remote Sens.* **2022**, *14*. [CrossRef]
6. Liu, Z.; Shang, Y.; Li, T.; Chen, G.; Wang, Y.; Hu, Q.; Zhu, P. Robust Multi-Drone Multi-Target Tracking to Resolve Target Occlusion: A Benchmark. *IEEE Trans. Multimed.* **2023**, *25*, 1462–1476. [CrossRef]
7. Sun, S.; Akhtar, N.; Song, H.; Mian, A.; Shah, M. Deep affinity network for multiple object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 104–119. [CrossRef]
8. 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.
9. Yuan, Y.; Wu, Y.; Zhao, L.; Chen, J.; Zhao, Q. DB-Tracker: Multi-Object Tracking for Drone Aerial Video Based on Box-MeMBer and MB-OSNet. *Drones* **2023**, *7*, 607. [CrossRef]
10. 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.
11. Tsai, C.Y.; Shen, G.Y.; Nisar, H. Swin-JDE: Joint detection and embedding multi-object tracking in crowded scenes based on swin-transformer. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105770. [CrossRef]
12. Lu, Z.; Rathod, V.; Votel, R.; Huang, J. Retinatrack: Online single stage joint detection and tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14668–14678.
13. Aharon, N.; Orfaig, R.; Bobrovsky, B.Z. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv* **2022**, arXiv:2206.14651.
14. Ren, H.; Han, S.; Ding, H.; Zhang, Z.; Wang, H.; Wang, F. Focus On Details: Online Multi-object Tracking with Diverse Fine-grained Representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 11289–11298.
15. Maggiolino, G.; Ahmad, A.; Cao, J.; Kitani, K. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. *arXiv* **2023**, arXiv:2302.11813.
16. Wu, H.; Nie, J.; He, Z.; Zhu, Z.; Gao, M. One-Shot Multiple Object Tracking in UAV Videos Using Task-Specific Fine-Grained Features. *Remote Sens.* **2022**, *14*, 3853. [CrossRef]

17. Yang, P.; Luo, X.; Sun, J. A simple but effective method for balancing detection and re-identification in multi-object tracking. *IEEE Trans. Multimed.* **2022**, *25*, 7456–7468. [CrossRef]
18. Lin, Y.; Wang, M.; Chen, W.; Gao, W.; Li, L.; Liu, Y. Multiple Object Tracking of Drone Videos by a Temporal-Association Network with Separated-Tasks Structure. *Remote Sens.* **2022**, *14*, 3862. [CrossRef]
19. 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.
20. 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* . [CrossRef] [PubMed]
21. 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.
22. Welch, G.; Bishop, G. An Introduction to the Kalman Filter. *Proc. SIGGRAPH Course* **2001**, *8*, 41.
23. 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]
24. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
26. Xu, Y.; Ban, Y.; Delorme, G.; Gan, C.; Rus, D.; Alameda-Pineda, X. Transcenter: Transformers with Dense Queries for Multiple-Object Tracking. *arXiv* **2021**, arXiv:2103.15145 .
27. Chu, P.; Wang, J.; You, Q.; Ling, H.; Liu, Z. Transmot: Spatial-temporal graph transformer for multiple object tracking. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 4870–4880.
28. Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; Feichtenhofer, C. Trackformer: Multi-object tracking with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8844–8854.
29. Kong, J.; Mo, E.; Jiang, M.; Liu, T. MOTFR: Multiple Object Tracking Based on Feature Recoding. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 7746–7757. [CrossRef]
30. Liang, C.; Zhang, Z.; Zhou, X.; Li, B.; Zhu, S.; Hu, W. Rethinking the competition between detection and reid in multiobject tracking. *IEEE Trans. Image Process.* **2022**, *31*, 3182–3196. [CrossRef] [PubMed]
31. Yu, E.; Li, Z.; Han, S.; Wang, H. Relationtrack: Relation-aware multiple object tracking with decoupled representation. *IEEE Trans. Multimed.* **2022**, *25*, 2686–2697. [CrossRef]
32. Jin, Y.; Gao, F.; Yu, J.; Wang, J.; Shuang, F. Multi-object Tracking: Decoupling Features to Solve the Contradictory Dilemma of Feature Requirements. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 5117–5132. [CrossRef]
33. 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. [CrossRef]
34. Deng, K.; Zhang, C.; Chen, Z.; Hu, W.; Li, B.; Lu, F. Jointing Recurrent Across-Channel and Spatial Attention for Multi-Object Tracking with Block-Erasing Data Augmentation. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 4054–4069. [CrossRef]
35. Liu, S.; Li, X.; Lu, H.; He, Y. Multi-object tracking meets moving UAV. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8876–8885.
36. Liang, C.; Zhang, Z.; Zhou, X.; Li, B.; Hu, W. One more check: Making “fake background” be tracked again. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 28 February and 1 March 2022; Volume 36; pp. 1546–1554.
37. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. Bytetrack: Multi-object tracking by associating every detection box. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 1–21.
38. Han, S.; Huang, P.; Wang, H.; Yu, E.; Liu, D.; Pan, X. Mat: Motion-aware multi-object tracking. *Neurocomputing* **2022**, *476*, 75–86. [CrossRef]
39. Qin, Z.; Zhou, S.; Wang, L.; Duan, J.; Hua, G.; Tang, W. 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, Vancouver, BC, Canada, 17–24 June 2023; pp. 17939–17948.
40. You, S.; Yao, H.; Bao, B.K.; Xu, C. UTM: A Unified Multiple Object Tracking Model With Identity-Aware Feature Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 21876–21886.
41. Xiao, T.; Li, S.; Wang, B.; Lin, L.; Wang, X. Joint detection and identification feature learning for person search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 3415–3424.
42. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
43. Chen, L.; Ai, H.; Zhuang, Z.; Shang, C. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23-27 July 2018; pp. 1–6.

44. 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.
45. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
46. 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.
47. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
48. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019 .
49. 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.
50. Bernardin, K.; Stiefelhagen, R. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 246309. [CrossRef]
51. Milan, A.; Schindler, K.; Roth, S. Challenges of Ground Truth Evaluation of Multi-target Tracking. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 13–28 June 2013; pp. 735–742. [CrossRef]
52. 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. [CrossRef]
53. 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.
54. 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, Tel Aviv, Israel, 23–27 October 2022; pp. 659–675.
55. 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.