remote sensing

MDPI

Article

MFACNet: A Multi-Frame Feature Aggregating and Inter-Feature Correlation Framework for Multi-Object Tracking in Satellite Videos

Hu Zhao [1,2], Yanyun Shen [1,2], Zhipan Wang [1,2] and Qingling Zhang [1,2,*]

[1] School of Aeronautics and Astronautics, Shenzhen Campus of Sun Yat-sen University, No. 66, Gongchang Road, Guangming District, Shenzhen 518107, China; zhaoh68@mail2.sysu.edu.cn (H.Z.); shenyy28@mail2.sysu.edu.cn (Y.S.); wangzp@mail2.sysu.edu.cn (Z.W.)

[2] Shenzhen Key Laboratory of Intelligent Microsatellite Constellation, Shenzhen Campus of Sun Yat-sen University, No. 66, Gongchang Road, Guangming District, Shenzhen 518107, China

* Correspondence: zhangqling@mail.sysu.edu.cn; Tel.: +86-18127095863

Abstract: Efficient multi-object tracking (MOT) in satellite videos is crucial for numerous applications, ranging from surveillance to environmental monitoring. Existing methods often struggle with effectively exploring the correlation and contextual cues inherent in the consecutive features of video sequences, resulting in redundant feature inference and unreliable motion estimation for tracking. To address these challenges, we propose the MFACNet, a novel multi-frame features aggregating and inter-feature correlation framework for enhancing MOT in satellite videos with the idea of utilizing the features of consecutive frames. The MFACNet integrates multi-frame feature aggregation techniques with inter-feature correlation mechanisms to improve tracking accuracy and robustness. Specifically, our framework leverages temporal information across the features of consecutive frames to capture contextual cues and refine object representations over time. Moreover, we introduce a mechanism to explicitly model the correlations between adjacent features in video sequences, facilitating a more accurate motion estimation and trajectory associations. We evaluated the MFACNet using benchmark datasets for satellite-based video MOT tasks and demonstrated its superiority in terms of tracking accuracy and robustness over state-of-the-art performance by 2.0% in MOTA and 1.6% in IDF1. Our experimental results highlight the potential of precisely utilizing deep features from video sequences.

Keywords: multi-object tracking; satellite video; inter-feature mapping; feature aggregation

Citation: Zhao, H.; Shen, Y.; Wang, Z.; Zhang, Q. MFACNet: A Multi-Frame Feature Aggregating and Inter-Feature Correlation Framework for Multi-Object Tracking in Satellite Videos. *Remote Sens.* **2024**, *16*, 1604. https://doi.org/10.3390/rs16091604

Received: 29 February 2024
Revised: 21 April 2024
Accepted: 29 April 2024
Published: 30 April 2024

1. Introduction

Multi-object tracking (MOT) aims to detect and associate multiple objects of interest in a video sequence, generating corresponding trajectories across consecutive frames. Recently, advanced remote-sensing video satellites have significantly enhanced spatial observation capabilities and data quality, providing information services and insight support in critical domains such as economic development, government administration, and national security. Monitoring multiple man-made objects (e.g., vehicles, vessels, and aircraft) with video satellites offers intuitive real-time dynamics information, which effectively supports the perception of high-dynamic scenes. MOT trackers enable the effective localization of multiple objects and generate individual trajectories that provide crucial information for further analysis. Currently, MOT tasks based on video satellites exhibit the following characteristics:

- Remote-sensing video satellites provide wide-field imaging, resulting in large frame sizes with a high proportion of background pixels. Additionally, video sequences are affected by satellite platform motion and atmospheric refraction, leading to image jitter. Cloud clusters, ground reflections, and the preprocessing of standard video sequence products also contribute to background noise. A sample of objects in remote-sensing images and video satellite sequences is shown in Figure 1;

- Video sequences contain numerous objects with dense spatial distributions and complex dynamics. Furthermore, due to the limited resolution of video sequences, individual objects are represented by dozens of pixels and lack fine texture details, which results in significant challenges in object detection and low discriminability between different objects. A sample of objects in video satellite sequences, along with their pixel statistics, is shown in Figure 2;
- When video satellites capture videos with a tiny off-nadir angle, the resulting video sequences can be approximated as being captured in a near-vertical manner. Compared with nature scenarios, the object sizes are relatively consistent, and fewer occlusions occur between objects. The motion patterns of objects are similar, and their trajectories tend to be linear. Additionally, unlike conventional remote-sensing imagery, remote-sensing video satellites capture imagery with high frame rates, which results in high scene overlap in the captured landscapes and significant overlap of the same object between adjacent frames.

Figure 1. Instances of different objects in high-resolution remote-sensing imagery and video satellite sequence, along with their pixel statistics. (**a**–**c**) Instances of vehicles, vessels, and aircraft in video satellite sequence. (**d**–**f**) Instances of vehicles, vessels, and aircraft in high-resolution remote-sensing imagery from the DOTA dataset.

Figure 2. Pixel statistics for vehicles, vessels, and aircraft in video satellite sequences with a ground sample distance (GSD) of 1.1 m.

Recently, a trend of rapid iteration exists in algorithms for MOT in video satellites. Early research primarily relied on background modeling-based object detection algorithms [1–7], such as background subtraction-based [8], optical flow-based [9], and frame difference-based [10] methods. These trackers initially extract the regions of interest that may contain objects, then exclude the background noise and reduce false alarms using prior knowledge, noise models, and motion models. etc. Similarly, background modeling algorithms in machine learning [11,12] have also been applied in the research of object tracking in video satellite imagery, such as the Gaussian mixture model (GMM) and ViBE [13]. For instance, Yang, et al. [12] used a visual background extractor (ViBE) to perform foreground separation on roughly segmented motion-salient regions. Ao, et al. [6] modeled the motion information of the original frame using the frame-difference method [10]. Traditional methods utilize logic algorithms based on a predetermined set of fixed parameters to distinguish the targets from the background. These fixed parameters are often determined through manual configuration, guided by human expertise, and tailored to specific scenes. However, in scenarios with substantial variations, the reliance on fixed parameters often leads to inconsistent tracking performance. Additionally, both slight camera platform jitter and rapidly changing ground scenes may introduce subtle yet widespread background noise. Traditional methods, including background modeling algorithms like ViBE, often exhibit limited robustness when dealing with such background noise and require additional post-processing.

In recent years, the emergence of large video satellite tracking datasets such as VISO [14], AIR-MOT [15], and SAT-MTB [16] has greatly advanced the development of deep learning-based (DL) MOT methods. Various tracking methods proposed new insights in terms of feature enhancement, association logic, trajectory recovery, and multi-task training strategies. Advanced online MOT methods can currently be classified into two-stage tracking-by-detection (TBD) and one-stage joint detection and tracking (JDT) [17]. Figure 3 illustrates the general architecture sketch of both the TBD and JDT methods. The TBD paradigm associates the detection results from an independent detector with trajectories using various similarity metrics. In recent years, deep learning-based object detectors [18,19] have been rapidly developed, leading to significant improvements in the TBD paradigm. Maher, et al. [20] used YOLO v2 [21] to obtain aircraft heads, tails, and full bodies and generate heading state parameter updates of the classic SORT [22] algorithm using the Karman filter [23]. Similarly, Wang, et al. [24] achieved the real-time tracking of multiple moving ships using three consecutive frames as an input to the YOLO v3 [25] network. Xiao, et al. [26] proposed the DSFNet, which uses lightweight 3D convolution to extract dynamic cues from multiple frames. Li, et al. [27] designed the D-RGB network based on the two-stage F-RCNN [28] model. The network subtracts the features from adjacent frames to extract the ROIs for second-stage detection in the RCNN.

The joint detection and tracking (JDT) paradigm enables the integration of detection and tracking tasks within a single model. For example, TGraM [15] utilizes graph convolution to update graphs based on extracted features and employs a simplified DANN [29] to balance detection and reidentification (ReID) tasks. Additionally, Feng, et al. [30] proposed the CKDNet, a tracking network based on CornerNet [31], to predict the correlation between tracking objects and detection results through a dual-branch LSTM [32] network, and generate matching trajectories employing the Hungarian algorithm [33].

The aforementioned research mainly focuses on enhancing the detection performance by incorporating contextual cues at the feature level. However, this often leads to the redundant extraction of deep features during inference. Additionally, motion-based tracking methods (e.g., SORT [22] and ByteTrack [34]) usually lack the utilization of appearance information. In contrast, some JDT models implicitly extract correlations between objects using graph convolution and LSTM for object association. We suggest enhancing feature reusability to address these challenges, which involves extracting and inheriting deep features from consecutive features to enhance the representation of current features, predicting inter-frame displacement through pixel mapping relationships between the features of

two adjacent frames, and leveraging ReID methods for the association between objects and trajectories.

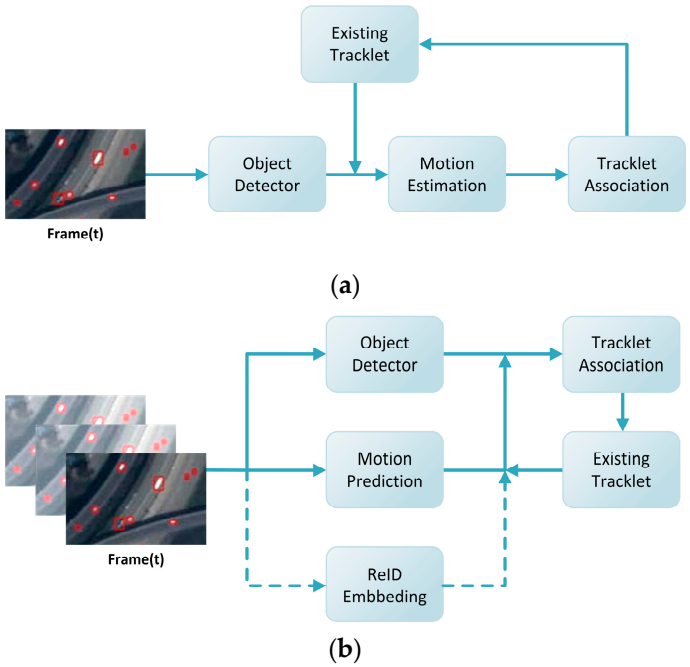


Figure 3. General architecture of online DL methods for MOT in video satellites. (**a**) Architecture of the TBD methods. (**b**) Architecture of the JDT methods; the dashed line indicates that the portion can be omitted; the red boxes represent the visualized object tracking results.

MOT based on natural scenes has also been a trending research topic. Despite the differences in application domains, some MOT algorithms based on natural scenes show remarkable potential in remote-sensing video sequences as well. Different from video satellites, MOT research based on natural scenes mainly focuses on common application scenes, such as public security, city traffic, autopilot, and sports broadcasting [35]. Several approaches have been proposed to address these issues. ByteTrack [34] introduces a two-stage cascaded matching logic that utilizes low-confidence detections to explore potential trajectory matches. OC-SORT [36] incorporates motion inertia into Kalman filter observations. GSTD [37] employs a graph model to match detections and recover low-confidence objects. FairMOT [38] employs an anchor-free detector with parallel ReID feature encoding heads to predict object detections and pixel-level ReID features simultaneously. CenterTrack [39] integrates information from two consecutive frames and directly predicts the relative displacements of object centers. TraDeS [40] supervises ReID feature encoding through cost volume and extracts similarity information to infer the object displacement between frames.

To efficiently incorporate motion cues from video sequences and fully utilize appearance information and motion prediction for robust tracking, we propose an online joint detection and tracking (JDT) framework called the MFACNet (multi-frame feature aggregating and correlation network). The MFACNet leverages the feature aggregation wrapper (FAW) module to aggregate consecutive feature information and enhance the current feature representation in channel dimensions. The feature-mapping motion estimation (FMME) module computes a correlation volume between the features of two adjacent frames, thereby constructing an inter-frame correlation map based on deep features, which enables the rapid estimation of object displacements. The IoU-prior cascaded matching method sequentially associates object detections with trajectories based on both spatial localization and appearance information. Furthermore, we developed an end-to-end tracking frame that adheres to the JDT paradigm as well as online inference.

The experimental results demonstrated significant improvements in key MOT metrics on video satellite tracking datasets using the MFACNet. The performance of our framework is highly competitive and achieves optimality. The contributions can be summarized as follows.

1. We proposed an end-to-end JDT framework called the MFACNet, which extracts the static features of man-made objects in video sequences while effectively integrating dynamic cues from consecutive frames. This framework significantly enhances the detection and tracking performance of the network;
2. For detection, we devised a lightweight feature aggregation wrapper (FAW) module, which utilizes sets of deformable convolutions (DCN) [41] to extract correlated information in different channel dimensions from the feature groups of multiple frames. All of the information is then employed to enhance the feature representation of the current frame. For tracking, we employed an end-to-end learnable feature-mapping motion estimation (FMME) module to estimate the displacement of individual objects. Regarding object-trajectory matching, we designed an IoU-prioritized cascaded matching scheme that effectively utilizes both localization and appearance information to generate and manage object trajectories;
3. We conducted the training and validation of our model on a video satellite object tracking dataset, which was constructed from the video sequences captured by the Jilin-1 video satellite and GF-3 high-resolution video satellite. The experimental results demonstrated that the proposed model achieves state-of-the-art performance on the experimental dataset methods in terms of tracking accuracy and robustness, which highlight the potential of precisely utilizing deep features from video sequences.

2. Materials and Methods

The MFACNet is based on the anchor-free detector CenterNet [42], and its overall architecture is illustrated in Figure 4. We adopted the same Dla-34 [43] from CenterNet with extra up-sampling connections and deformable convolutions as the backbone. The input image $I^{(t)} \in R^{H \times W \times 3}$ produces a down-sampled feature $f^{(t)} \in R^{Hf \times Wf \times C}$, where $Hf = \frac{H}{4}$, $Wf = \frac{W}{4}$.

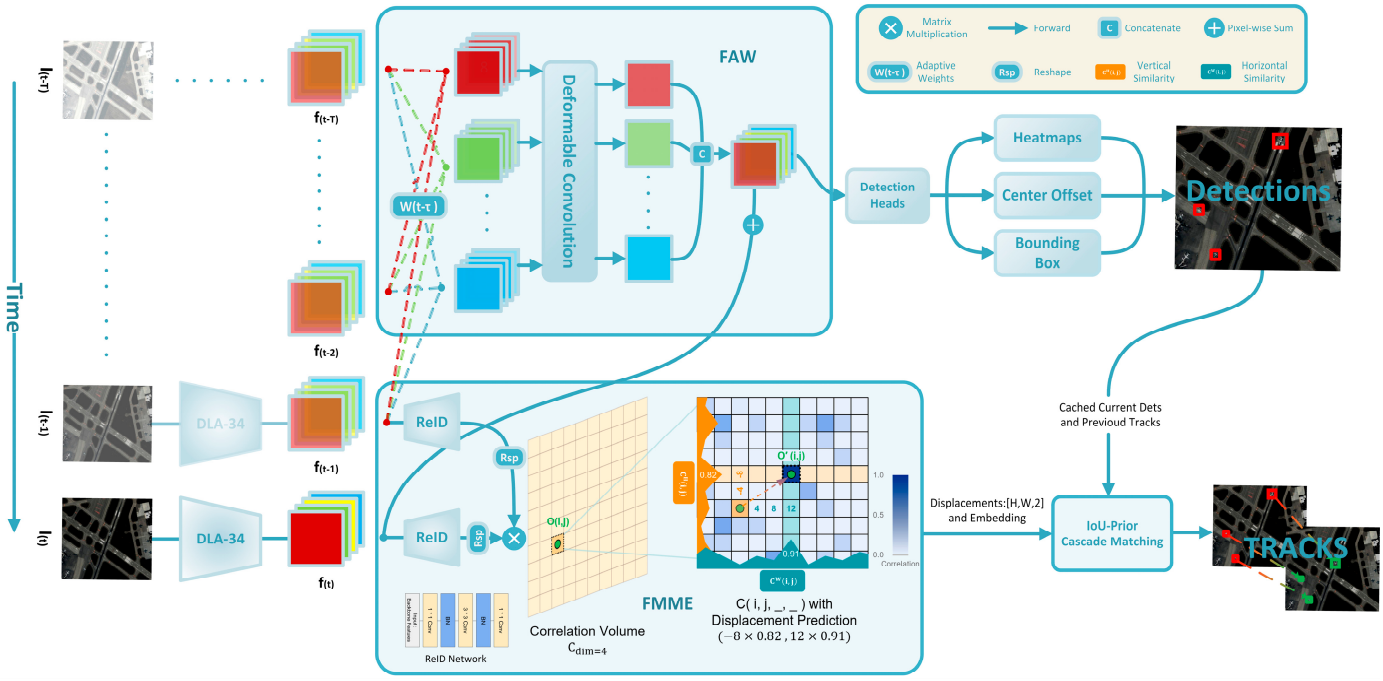


Figure 4. Overview of the MFACNet. The MFACNet may propagate features from multiple previous frames for object feature enhancement (i.e., T > 1), which is not shown in the above figure for simplicity.

In addition to predicting object class-wise heatmaps, $H^{(t)} \in [0,1]^{Hf \times Wf \times C}$, and box sizes, $S^{(t)} \in R^{Hf \times Wf \times 2}$, which is similar to CenterNet, the features are also fed into an additional convolutional head to predict object displacements: $M^{(t)} \in R^{Hc \times Wc \times 2}$. These displacements, $M^{(t)}$, are utilized in the subsequent IoU-prior cascaded matching to associate the object trajectories with the detection results.

2.1. Feature Aggregation Wrapper Module

Small-sized objects in remote-sensing video sequences pose challenges for effectively capturing their deep features using a backbone network. To address this issue, we proposed an online multi-frame features aggregation module inspired by [44,45]. The feature aggregation wrapper (FAW) module efficiently aggregates multi-level semantic information, making it suitable for modifying the features of tiny objects with small displacements in remote-sensing video imagery. Additionally, it provides a concise and fast implementation while preserving online processing ability.

The FAW module takes the deep features of previous T frames, $f^{(t-\tau)}$, as inputs, where $\tau = 1, 2, \ldots, T$. Then, a channel-wise feature collection, f_{chan}, is extracted from T-consecutive features in all channel dimensions. The feature information from different channel dimensions is aggregated using the deformable convolutional network (DCN). The output's aggregated information is then utilized to enhance the representation of the object features in the current frame. The overall architecture of the FAW is illustrated in Figure 5.

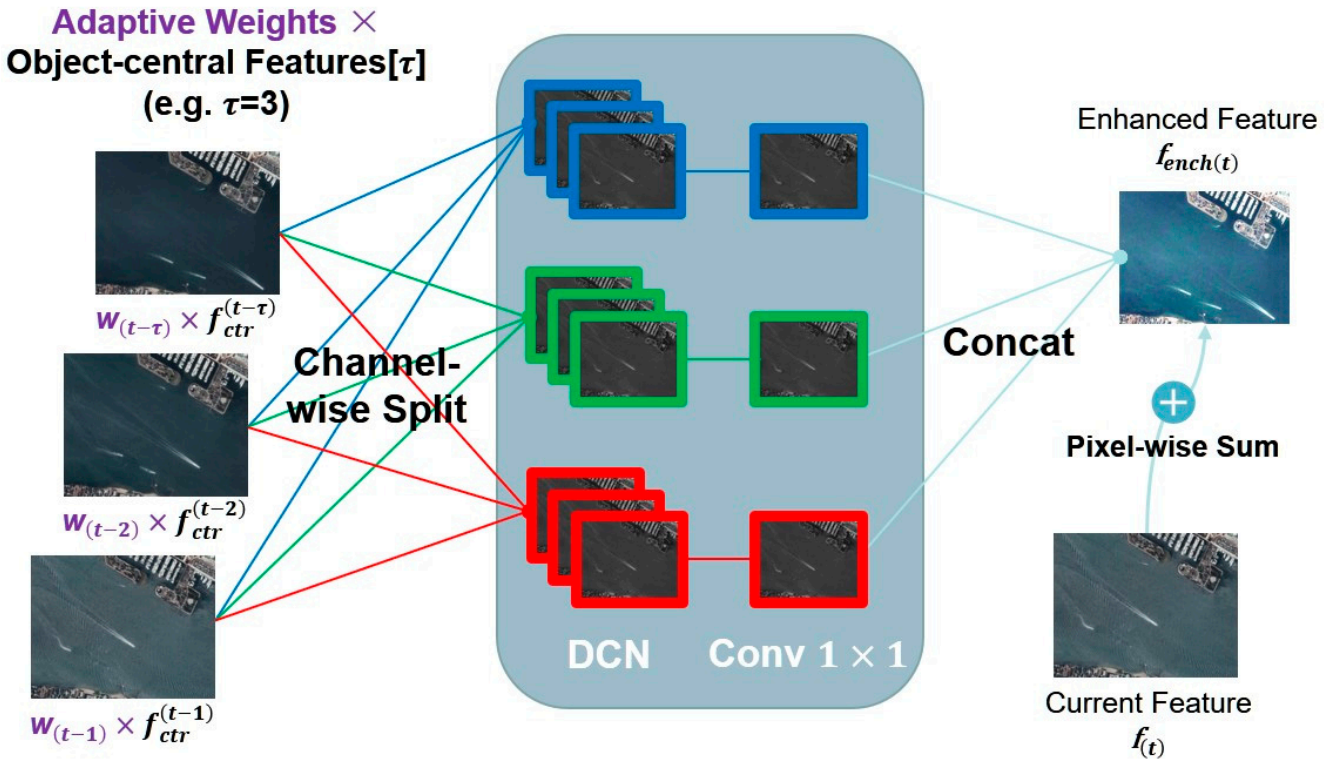


Figure 5. Overview of the FAW module. The FAW module may adaptively propagate contextual cues from previous T features for object feature enhancement (i.e., T > 1) through a DCN.

2.1.1. Feature Propagation

For previous T frames of the imagery $I^{(t-\tau)}$, where $\tau = 1, 2, \ldots, T$, we utilized their features, $f^{(t-\tau)} \in R^{Hf \times Wf \times 64}$, from the backbone and the corresponding center heatmaps, $H^{(t-\tau)} \in R^{Hf \times Wf \times C}$, as the input. Similar to the characteristics of anchor-free models like CenterNet, we first added Gaussian-shaped peaks to all the center points of all the cate-

gories, generating the aggregated center heatmap $H_{ctr}^{(t-\tau)} \in R^{Hf \times Wf \times C}$. We then generated central attention features, $f^{(t-T)}$, based on $H_{ctr}^{(t-\tau)}$ using the following equation:

$$f_{ctr}^{(t-\tau)} = H_{ctr}^{(t-\tau)} \circledcirc f^{(t-T)},\ \tau = 1, 2, \ldots, T \quad (1)$$

where $\circledcirc$ denotes the Hadamard product, $f_{ctr}^{(t-\tau)}$ contains the feature information of the center points in $H_{ctr}^{(t-\tau)}$, which can be utilized for the inheritance of object-central features. Simultaneously, to ensure the effectiveness of feature propagation, we employed adaptive fusion weights, $w^{(t-\tau)} \in R^{Hf \times Wf \times 1}$, to measure the significance of the features in each frame, resulting in the weighted central attention feature $\overline{f_{ctr}}$ in the following equation:

$$\overline{f_{ctr}} = \text{concat}\left(\left(w^{(t-\tau)} \circledcirc f_{ctr}^{(t-\tau)}\right)\right) \in R^{Hf \times Wf \times C},\ \tau = 1, 2, \ldots, T \quad (2)$$

where $w^{(t-\tau)}$ is a normalized attention map predicted using a set of convolutions in the FAW module.

2.1.2. Feature Enhancement

The FAW module performs a channel-wise segmentation on $\overline{f_{ctr}}$. For the segmented features, the FAW module collects and concatenates the channel features belonging to the same dimension across the temporal dimension T, to obtain sets of channel features $f_{chan} \in R^{Hf \times Wf \times T}$ for C channels. Subsequently, a DCN modified the f_{chan} of different channels, and then a 1×1 convolution was utilized to aggregate all information modified by DCN, generating an aggregated feature with a size of $Hf \times Wf \times 1$ for each channel. Finally, the inherited features, $f_{aggr} \in R^{Hf \times Wf \times C}$, were obtained by concatenating the channel-wise tensor. The features f_{aggr} were then summed pixel-wise with the existing output features, f_t, to obtain the enhanced features of the imagery in the following equation:

$$f_{aggr} = \text{concat}(\text{Conv}_{1\times1}(\text{DCN}(\, f_{chan}[i]\,))\,),\ i = 0, 1, 2, \ldots, C \quad (3)$$

$$f_{ench} = f^{(t)} + f_{aggr} \quad (4)$$

where the DCN is set with a kernel size of 3. The enhanced features, $f_{ench} \in R^{Hf \times Wf \times 64}$, incorporated motion cues from the previous T frames, and were subsequently fed into the detection heads for object center heatmaps and center offset predictions, while the bounding box predictions were based on the original $f^{(t)}$.

2.2. Feature-Mapping Motion Estimation Module

Man-made objects often exhibit tiny sizes, small displacements, and low feature expressiveness in video satellite imagery. These characteristics result in a high correlation between the features of consecutive frames. The correlation volume, commonly used in stereo-matching [46,47], depth estimation [48,49], and optical flow estimation networks, can compute the matching similarity between two consecutive features, reflecting their mapping relationship. Laga, et al. [50] summarized different methods for constructing correlation volumes in depth-estimation research.

When predicting displacements with an anchor-free model, we were particularly interested in the global mapping relationship between the center features of multiple objects in different frames. JDE models in [39,40] learned relative displacement information by analyzing the corresponding identity features between two frames. FMME constructs an all-pairs pixel correlation [51] between the features of the current frame and the preceding one. It extracts the minimum-cost matching of correlated elements within object-central pixels of different features, thereby obtaining the mapping relationship between multiple object centers across features. Finally, the FMME model outputs the displacement prediction based on the constructed inter-frame mapping relationship. The overall architecture of the FMME model is illustrated in Figure 6.

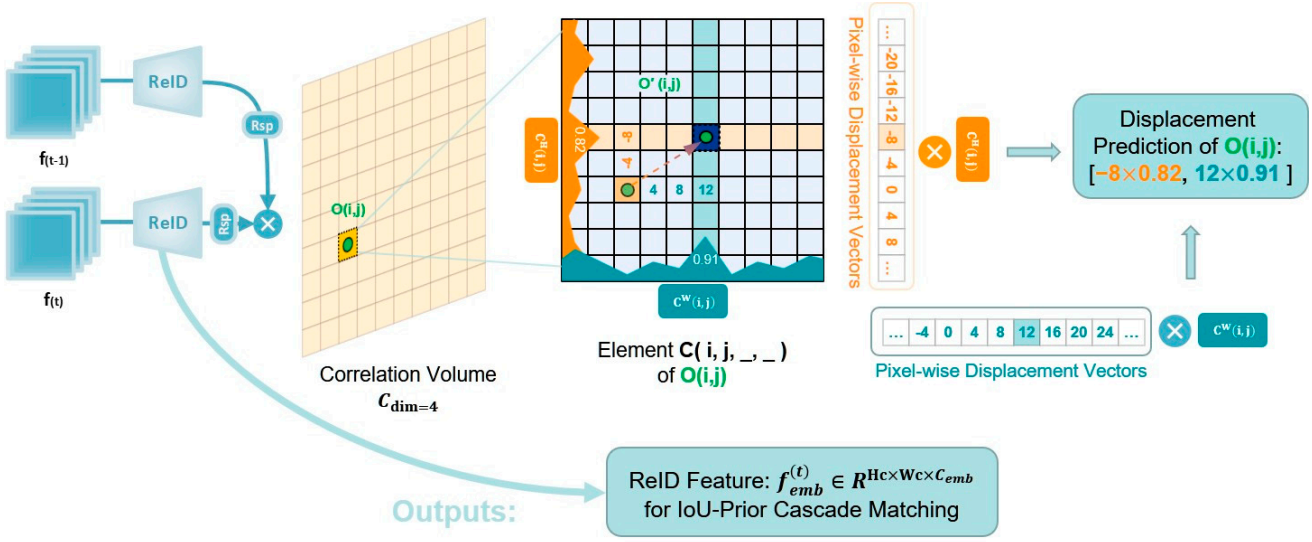


Figure 6. Overview of the FMME module. The FMME module may predict the object displacement based on the inter-frame mapping constructed using a correlation volume and generate ReID features for trajectory management.

2.2.1. ReID Sub-Network

Directly using the Backbone features $f^{(t)}$ is not conducive to ReID tasks or network optimization, resulting in a significant computational burden while calculating the correlation volume. Therefore, we first employed a ReID sub-network, $\epsilon(\bullet)$, to adjust the original features of $f^{(t)} \in R^{Hc \times Wc \times 64}$. The sub-network produced ReID features with reduced channel dimensionality, denoted as $f_{emb}^{(t)} = \epsilon\left(f^{(t)}\right) \in R^{Hc \times Wc \times C_{emb}}$, where $Wc = Wf$, $Hc = Hf$, and $C_{emb} = 32$. Features $f_{emb}^{(t)}$ as well as $f_{emb}^{(t-1)}$ from the previous frame were used for the computation of the correlation volume.

2.2.2. Displacement Prediction

For optical flow estimation networks, the correlation volume indicates the pixel-wise similarity between $f_{emb}^{(t)}$ and $f_{emb}^{(t-\tau)}$. Treating the object-tracking task as a matching of center point feature pixels, the displacement matrix $M \in R^{Hc \times Wc}$ could be predicted using the correlation volume. First, we calculated the four-dimensional all-pairs correlation volume C by performing a matrix multiplication between the extracted features $f_{emb}^{(t)}$ and the preceding feature $f_{emb}^{(t-\tau)}$ using the following equation:

$$C = \left\langle f_{emb}^{(t)}, f_{emb}^{(t-1)} \right\rangle \in R^{Hc \times Wc \times Hc \times Wc}, C(i,j,k,l) = f_{emb}^{(t)}(i,j) f_{emb}^{(t-1)}(k,l)^{T}, \quad (5)$$

where $\langle \_, \_ \rangle$ denotes the matrix multiplication operation and ^{T} represents the transpose operation.

Based on the matching similarity from the correlation volume, global matches of the target centers could be generated, which in turn allowed for the prediction of displacements, M. For a specific object $O_{i,j}$ at time t with its correlation volume $C(i,j,k,l)$, we utilized pooling convolutions with kernel sizes of $Hc \times 1$ and $1 \times Wc$ to individually generate the normalized feature mapping $C^{W}(i,j) \in [0,1]^{1 \times Wc}$ and $C^{H}(i,j) \in [0,1]^{Hc \times 1}$ in the vertical and horizontal dimensions, respectively. Then, we reflected these mappings to obtain the actual displacements in the original image. The process is as follows: we first defined the pixel-wise displacement vectors, $U(i,j) \in R^{1 \times Wc}$ and $V(i,j) \in R^{Hc \times 1}$. The displacement prediction, $M^{t} \in R^{Hc \times Wc \times 2}$, could be obtained by taking the element-wise product of

the normalized feature mappings and corresponding displacement vectors of $O_{i,j}$ in the following equation:

$$\begin{cases} U(i,j,k,*) = (k-i) \times s, \ \ 1 \le k \le Wc \\ V(i,j,*,l) = (l-j) \times s, \ \ 1 \le l \le Hc \end{cases}, \tag{6}$$

$$M(i,j) = \left[C^H(i,j)^T V(i,j), C^W(i,j) U(i,j)^T \right]^T, \tag{7}$$

where s = 4 and is the factor of recovery. For a specific object $O_{i,j}$ at time t, the displacement prediction $M(i,j)$ reflects the real pixel displacement between the central feature $f_{emb}^{(t)}(i,j)$ of $O_{i,j}$ and $f_{emb}^{(t-\tau)}(k,l)$ of the $O_{k,l}$ at time $t-\tau$, where $O_{k,l}$ represents, with the maximum similarity, to $O_{i,j}$. This displacement is calculated in the original image size.

To obtain the ReID loss under sparse labels, when the target $O_{i,j}$ at time t correctly matches the target $O_{k,l}$ of the same identity at the time $t-\tau$ in a certain dimension, we set $Y_{ijkl} = 1$, and 0 otherwise. We employed a focal loss for FMME training and only considered positive samples with their normalized similarity to define the displacement prediction loss, which increases the margin between the true identity targets and other different identity targets. This improvement supervises the training of the ReID sub-network and subsequently supervises the prediction process of the FMME module.

2.3. IoU-Prior Cascade Matching

Generally, man-made objects exhibit small displacements and almost no changes in size between adjacent frames. Therefore, based on the idea of matching targets in order of overlap, we designed an IoU-prior cascade, matching for matching trajectories, T, with the current frame detections $D^{(t)}$. In the matching step, we computed the intersection over union (IoU) between the trajectory $T^{(t-1)}$ and current frame detection $D^{(t)}$, and prioritized the association based on a predefined IoU threshold when IoU ≥ 0.6.

For rapidly moving targets, we introduced a second step, where we associated unmatched detection, $D_{um\_i}^{(t)}$, with the nearest unmatched trajectory, $T_{um\_i}^{(t-1)}$, within a radius of r using the greedy matching algorithm. Finally, for the remaining unmatched detection, $D_{um\_ii}^{(t)}$, and trajectories, $T_{um\_ii}^{(t-1)}$, we performed the association based on the cosine similarity between their ReID features. The corresponding pseudo-code for the IoU-prior cascade matching is shown in Table 1.

2.4. Tracklet Management and Training Strategy

Once a detection with a high confidence of $conf(d_{um_{ii}}) \geq 0.4$ cannot be matched with any existing trajectories, a new tracklet is to be included in the trajectory management. If a trajectory fails to match any detections for 15 frames, it is regarded as disappeared and will no longer be updated.

For detection, we employed the same 2D detection heads and loss functions as CenterNet [42]. The feature fusion in the FAW module was supervised by the loss function L_{det}. For the displacement prediction, we trained the ReID sub-network with ID labels as supervision and subsequently fine-tuned the FMME module. The overall model loss is computed as the sum of all the losses in the following equation:

$$L_{total} = L_{heatsmaps} + L_{reg} + 0.1 \times L_{boxes} + L_{FMME}, \tag{8}$$

Table 1. Pseudo-code of IoU-prior cascade matching in the MFACNet, which effectively utilizes displacement prediction, localization, and appearance information to manage object trajectories.

Pseudo-code of IoU-prior cascade matching
Input: A video sequence V; tracking model *Net*; IoU matching threshold ζ; similarity threshold σ; new track threshold θ.
Output: Tracklets T of the video.
Initialization: $T \leftarrow \varnothing$
For frame $f^{(t)}$ in V **do:**
$D^{(t)}, M^{(t)} \leftarrow Net(f^{(t)})$
/* First association */
/* Associate $D^{(t)}$ and $T^{(t-1)}$ using IoU threshold $\zeta = 0.60$ */
For d , $t \in D^{(t)}$,$T^{(t-1)}$ **do**:
If IoU$(t\|d) \geq \zeta$ **then**
adding detection d to track t
end if
end for
$D^{(t)}_{um\_i} \leftarrow$ remaining object boxes from $D^{(t)}$
$T^{(t-1)}_{um\_i} \leftarrow$ remaining tracks from $T^{(t-1)}$
/* Second association */
/* Associate $D^{(t)}_{um\_i}$ and $T^{(t-1)}_{um\_i}$ using the Greedy Order Algorithm */
$Cost^{(t)} \leftarrow$ where *l2 distance*$\left(D^{(t)}_{um\_i} + M^{(t)}_{um\_i} \middle| T^{(t-1)}_{um\_i}\right) \leq \sqrt{width^2_{T^{(t-1)}_{um_i}} + height^2_{T^{(t-1)}_{um_i}}}$
Linear assignment using the greedy algorithm with $Cost^{(t)}$
$D^{(t)}_{um\_ii} \leftarrow$ remaining object boxes from $D^{(t)}_{um\_i}$
$T^{(t-1)}_{um\_ii} \leftarrow$ remaining tracks from $T^{(t-1)}_{um\_i}$
/* Third association */
/* Associate $D^{(t)}_{um\_ii}$ & $T^{(t-1)}_{um\_ii}$ using ReID similarity threshold $\sigma = 0.3$ */
For d_{um_ii} , $t_{um\_ii} \in D^{(t)}_{um\_ii}$,$T^{(t-1)}_{um\_ii}$ **do:**
If (cosine similarity$\left(emb[d_{um_ii}] \middle| emb[t_{um_ii}]\right) \geq \sigma$ and t_{um_ii} still alive) **then**
adding detection d_{um_ii} to track t_{um_ii}
end
Else if conf$\left(\mathrm{d_{um_{ii}}}\right) \geq \theta$ where $\theta = 0.4$ **then**
Initial new track based on d_{um_ii}
end
end
end

3. Experiments

3.1. Experimental Data

We constructed a dataset comprising 76 video satellite sequences captured using Jilin-1 and GF-03 satellites. All sequences were manually annotated with localization and identity information following the MOT label [52]. The resulting dataset included three major categories: vehicles, ships, and aircraft, with a resolution of 1600 × 1200 and a ground sampling distance of 0.91 ∼ 0.95 m. The frame rate of these sequences ranged from 9 to 13 fps. Additionally, we added the VISO dataset [14] as additional training samples. In total, we collected 123 annotated sequences and 6770 objects with unique identities for model training and validation. Among them, 100 sequences were used for training, while 37 sequences (including 7 VISO dataset testing sequences) were used for testing. For the ablation study, we split our training sequences into two halves and used 50% of the

sequences for training and the remaining 50% for performance validation. Part of the test samples of our dataset is shown in Figure 7.

Figure 7. Our tested dataset included three categories: vehicles, vessels, and aircraft. The red boxes represent the ground truth bounding boxes of objects. Subset of videos was utilized for the model comparison experiments.

3.2. Implementation Details

For the MFACNet, we used a sequence of three consecutive frames (including the current frame) as the input, which is discussed in Section 3.4. Considering the inclusion of additional training data and different model scaling strategies, we set the training resolution to 1280 × 1024 with a batch size of 8, and we employed the Adam optimizer [53] for 60 epochs of training. The initial learning rate was 1.5×10^{-4} and dropped by a factor of 10 at 30 epochs. The remaining parameters were the same as CenterTrack. All the inferences were carried out on an RTX 3090 GPU.

3.3. Evaluation Metrics

Our experiment used the Clear Matrix [54] and IDF1 [55] for the evaluation of multi-object tracking performance. The overall performance assessment was based on the multiple-object tracking accuracy (MOTA) and IDF1 score (IDF1), as depicted in the following equations:

$$\mathrm{MOTA} = 1 - \frac{\sum_t (\mathrm{FN}_t + \mathrm{FP}_t + \mathrm{IDSW}_t)}{\sum_t \mathrm{GT}_t}, \tag{9}$$

$$\mathrm{IDF1} = \frac{2 \times \mathrm{IDTP}}{2 \times \mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}} \tag{10}$$

MOTA is a comprehensive metric that evaluates the overall tracking performance by considering multiple aspects, such as false positives, false negatives, identity switches, and fragmentation. It provides a unified measure of tracking accuracy, where a higher value signifies better tracking performance.

IDF1 is a single-value metric that combines the precision and recall of the tracking system. Precision measures the ratio of correctly identified objects to the total number of identified objects, while recall measures the ratio of correctly identified objects to the total number of ground truth objects. IDF1 combines these two metrics to provide a balanced evaluation of tracking accuracy, where a higher IDF1 score indicates better precision and recall performance.

Additionally, the following quantitative metrics were employed as supplementary performance measures: false negatives (FNs), which refer to the number of missed detections, false positives (FPs), which represent the number of incorrect detections, identity

switches (IDSWs), which quantify the number of times the tracker incorrectly switches the identities of objects, mostly tracked trajectories (MTs), and mostly lost trajectories (MLs). Specifically, MTs and MLs are utilized to assess the tracking performance of the tracking model in capturing the complete target trajectories. The MT is defined as the percentage of trajectories for which the tracking model's output covers more than 80% of its full trajectory, while the ML is defined as the percentage of trajectories for which the output only covers less than 20% of its full trajectory. The statistical results of the MTs and MLs reflect the consistency and robustness of the tracking model in tracking full-target trajectories. These performance metrics provide valuable insights into the accuracy, robustness, and consistency of multi-object tracking methods, enabling the effective evaluation and comparison of different tracking approaches.

3.4. Number of Preceding Features Input

The FAW module adaptively propagates contextual cues from previous features for object feature enhancement. In this section, we investigate the impact of different numbers of input features (T frames) on the MOT performance. It is worth emphasizing that we used 50% of videos from the training set for training and the remaining 50% for performance validation in the experiments.

We experimented with different numbers of input features selected from the preceding T frames as the inputs of our FAW module, where T was set to 1, 2, 3, and 4, respectively. The results are summarized in Table 2.

Table 2. MOT performance with different numbers of input features (T frames) of the FAW module. ↑ indicates that higher is better and ↓ indicates that lower is better.

T Frames	MOTA ↑	IDF1 ↑	FP ↓	FN ↓	IDS ↓
1	27.2	45.1	119,084	405,698	5683
2	28.4	43.3	75,495	438,956	6809
3	27.6	41.0	69,564	450,244	7408
4	26.5	38.8	67,934	458,443	8863

As shown in Table 2, using only one feature of the previous frame as the input results in the highest IDF1 performance. When using the features of the previous two frames as the input, the model achieves the highest MOTA performance, albeit with a slight decrease in the IDF1. This phenomenon may be attributed to the FAW module significantly reducing the false detection rate while introducing a small number of missed detections, resulting in a certain number of identity switches in the tracking algorithm. Furthermore, as the number of input features (T frames) increases, there is a noticeable performance decline in both the MOTA and IDF1. To ensure the accuracy of tiny object detection, we selected T = 2 as our FAW module input. For the MFACNet, aggregating two input features in the FAW module indicates utilizing three consecutive frames (including the current frame) as the input.

3.5. Comparison with Existing Methods

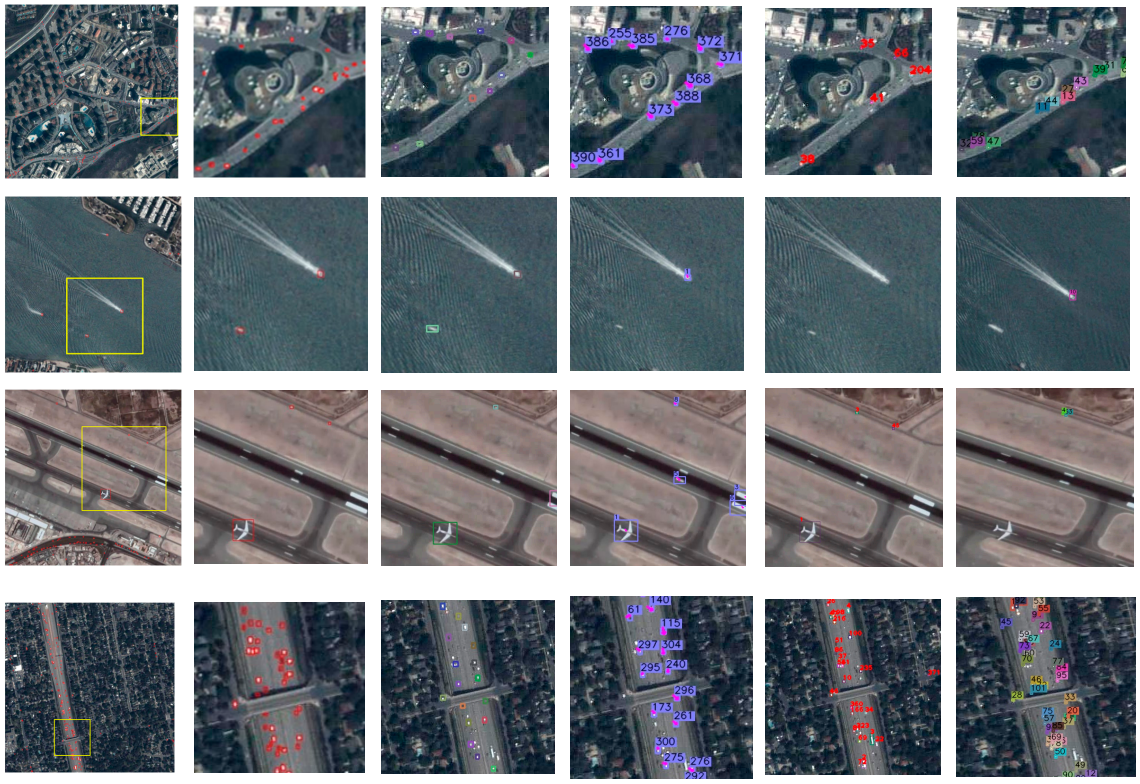
We conducted a comparative analysis between the MFACNet and representative online trackers, including CenterTrack, FairMOT, TraDes, SORT, OC-SORT, and ByteTrack. CenterTrack, FairMOT, and TraDes belong to the JDT framework models, while SORT, OC-SORT, and ByteTrack can be classified as TBD frameworks. For TBD models, we utilized YoloX [18] to provide the benchmark detection results. For the models included in the comparative experiments, we trained them by treating all the class categories of the targets as one category and tried our best to minimize the tracking performance loss of multi-class tracking models. Table 3 presents the tracking performance of the comparative models on a test dataset consisting of 37 videos.

Table 3. Quantitative results on our test set of 37 videos. ↑ indicates that higher is better and ↓ indicates that lower is better.

Method	Year	Joint	MOTA ↑	IDF1 ↑	MT ↑	ML ↓	FP ↓	FN ↓	IDS ↓
SORT	2016		30.6	43.6	22.9	36.7	**46,459**	212,073	3461
ByteTrack	2021		**31.1**	51.2	**30.9**	31.9	64,455	190,420	3360
OC-SORT	2022		26.8	43.6	14.0	52.8	**14,334**	258,921	**1093**
FairMOT	2020	✔	17.5	37.2	15.9	55.3	135,991	**117,260**	56,706
CenterTrack	2020	✔	28.2	53.6	27.9	**28.0**	111,206	**151,189**	6925
TraDes	2021	✔	27.3	**53.8**	28.0	**27.6**	99,521	167,834	5270
MFACNet (ours)	2023	✔	**33.1**	**55.4**	**31.9**	31.0	52,792	189,089	**1743**

The comparative experimental results demonstrate that our model achieved the best performance on the test dataset. Compared to JDE models like CenterTrack and TraDes, the MFACNet significantly reduced the FPs and IDSWs while introducing a small number of FNs, resulting in substantial improvements in both the MOTA and IDF1. In contrast to the TBD model, leveraging advanced detectors often leads to lower false positive detections (FPs) and identity switches (IDSWs), which can be attributed to the higher adaptability to different resolutions on advanced detectors. Furthermore, when compared to the ByteTrack method with similar performance, we observed that the MFACNet, incorporating rich contextual information, achieved significant improvements in detection performance over the TBD model and facilitated the management of object trajectories.

Figure 8 visualizes some representative tracking results of the MFACNet as well as the results from other comparative models, including CenterTrack, ByteTrack, and TraDes.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 8. Tracking results of the MFACNet and other comparative models visualized on the test set, some of the results are labeled with identity numbers; (**a**,**b**) shows the research areas and zoomed-in visualized ground truth. (**c**–**f**) presents the visualized tracking results of the MFACNet compared to CenterTrack, ByteTrack, and TraDes.

4. Discussion

To better evaluate the effectiveness of the design modules, we performed ablation experiments on our MFACNet. In the ablation experiments, we selected the CenterTrack model as the baseline and compared it with the proposed FAW, FMME, and MFACNet models. It is worth emphasizing that we used 50% of videos from the training set for training and the remaining 50% for performance validation in the experiments. The results of the ablation experiments are presented in Table 4.

Table 4. Ablation study for feature-mapping motion estimation (FMME), feature aggregation wrapper (FAW), and IoU-prior cascade matching (IoU CM). ↑ indicates that higher is better and ↓ indicates that lower is better.

Baseline	FMME	FAW	IoU CM	MOTA ↑	IDF1 ↑	FN ↓	MT ↑	IDS ↓
✔	✔			26.6	42.9	502,392	30.1	8882
✔	✔	✔		28.4	43.2	438,951	33.8	6866
✔	✔	✔	✔	28.4	43.3	438,956	34.3	6809
	CenterTrack			26.3	42.7	511,462	31.0	11,741

Compared with the baseline model, the inclusion of the FMME module led to an improvement of 0.2% in the IDF1, with a significant reduction of 24.4% in the IDSWs. Furthermore, there was a 1.8% enhancement in the MOTA, 0.3% in the IDF1, and 3.7% in the MTs, along with an additional reduction of 2016 in the IDSWs when both the FMME and FAW were simultaneously incorporated. Additionally, by incorporating IoU-prior cascaded matching, the integration of the MFACNet reduced the IDSWs and improved the IDF1 in the aforementioned performance. CenterTrack utilizes two consecutive frames and contextual cues as inputs, relying solely on localization information for trajectory matching. In comparison, the MFACNet significantly improved the MT while reducing both the FNs and IDSWs, resulting in a 2.1% improvement in MOTA performance and a 0.6% improvement in the IDF1 compared to CenterTrack.

4.1. Effectiveness of the FMME

Unlike the MFACNet, CenterTrack uses the fused information from two consecutive frames and heatmaps of the previous detection as inputs, directly predicting object displacements. After incorporating the FMME module, the model achieved a slight improvement in both the MOTA and IDF1, while also observing a significant decrease of 2859 in the IDSWs. This demonstrates that the FMME module provides better displacement predictions through inter-feature mapping construction and ReID sub-network optimization. This helps to reduce identity changes caused by erroneous associations between the trajectories and detection, enhancing the ability to manage target identities. It is worth noting that the appearance features generated using the ReID sub-network are not utilized in the association process between the detection and trajectories.

4.2. Effectiveness of the FAW Module

After incorporating the FAW module, the model exhibited a considerable decrease in FNs by 12.6% and IDSWs by 22.7%, respectively, resulting in a 3.7% improvement in MTs, a 1.8% improvement in the MOTA, and a 0.3% improvement in the IDF1. This demonstrates that the FAW module reduces missed detection for tracking, thus leading to an improvement in the MOTA, which primarily measures the detection performance. Additionally, both a decrease in IDSWs and an increase in MTs indicate that the model's ability to track complete trajectories is improved. This further confirms that the FAW module utilizes features from the previous two frames to enhance the current frame's feature representation, thereby improving the tracking robustness and consistency, ultimately reflected in the improvement of the IDF1.

4.3. Effectiveness of IoU-Prior Cascade Matching

By incorporating IoU-prior cascade matching, the complete MFACNet further reduces a small number of IDSWs and improves MTs, which illustrates its ability to recover lost tracklets through multi-source information and delicate logic design. IoU-prior cascade matching utilizes the high overlap characteristics of objects in video frames to filter out the same targets using an IoU threshold in the first step. In the final association step, ReID information is also leveraged to associate the remaining detections with unmatched tracklets. This alleviates, to some extent, the issues of erroneous matching and trajectory interruptions caused by relying solely on localization information and using the greedy order algorithm to associate the detections with tracklets.

5. Conclusions

In this paper, we presented an online deep learning framework, MFACNet, for multi-object tracking with video satellites. The framework enhances the feature representation of the current frame by inheriting multi-frame deep features. It also constructs inter-frame feature mapping by calculating the correlation volume between the current frame and the preceding one, enabling the prediction of object displacements. Finally, a novel matching algorithm based on object motion and appearance features was employed to complete the trajectory construction. Qualitative and quantitative experiments demonstrated that our JDE algorithm, MFACNet, effectively reduces false positives (FPs) in detection, while efficiently utilizing motion and appearance information to construct and manage object trajectories, leading to a great reduction in identity switches (IDSWs). Improvements were observed in both the detection and tracking performance. Moreover, comparative experiments showed that our tracking framework achieved superior performance compared to popular multi-object tracking models.

For the MFACNet, our improvement came at the cost of slightly increased FNs, indicating the requirement to explore more effective ways of utilizing features. Additionally, during the training of the MFACNet, we observed a certain degree of overfitting in the tracking vehicle targets, where only vehicles with prominent features obtained satisfactory tracking results. This necessitates a reevaluation of our training strategy and training data. Lastly, current MOT methods are tailored for finely processed video sequence products and have not been formally deployed on video satellites. Normally, video sequences require the excessive occupation of satellite communication bandwidth and complex post-processing procedures. To provide real-time information for efficient MOT, we are still investigating the feasibility of deploying the MFACNet on video satellite platforms.

Author Contributions: Conceptualization, H.Z., Z.W. and Q.Z.; Methodology, H.Z., Y.S. and Q.Z.; Dataset Construction, H.Z. and Y.S.; Validation, H.Z.; Writing the original draft, H.Z. and Q.Z.; Formal analysis, Q.Z.; Writing, review, editing, and supervision H.Z., Z.W. and Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (No. 2022YFE0209300) and the Shenzhen Science and Technology Program (No. ZDSYS20210623091808026).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. Keck, M.; Galup, L.; Stauffer, C. Real-time tracking of low-resolution vehicles for wide-area persistent surveillance. In Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV), Clearwater Beach, FL, USA, 15–17 January 2013; pp. 441–448.
2. Karimi Nejadasl, F.; Gorte, B.G.H.; Hoogendoorn, S.P. Optical flow based vehicle tracking strengthened by statistical decisions. *ISPRS J. Photogramm. Remote Sens.* **2006**, *61*, 159–169. [CrossRef]

3. Zhang, J.; Jia, X.; Hu, J.; Tan, K. Satellite Multi-Vehicle Tracking under Inconsistent Detection Conditions by Bilevel K-Shortest Paths Optimization. In Proceedings of the 2018 Digital Image Computing: Techniques and Applications (DICTA), Canberra, Australia, 10–13 December 2018; pp. 1–8.
4. Zhang, J.; Zhang, X.; Tang, X.; Huang, Z.; Jiao, L. Vehicle Detection and Tracking in Remote Sensing Satellite Vidio based on Dynamic Association. In Proceedings of the 2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp), Shanghai, China, 5–7 August 2019; pp. 1–4.
5. Ahmadi, S.A.; Ghorbanian, A.; Mohammadzadeh, A. Moving vehicle detection, tracking and traffic parameter estimation from a satellite video: A perspective on a smarter city. *Int. J. Remote Sens.* **2019**, *40*, 8379–8394. [CrossRef]
6. Ao, W.; Fu, Y.W.; Hou, X.Y.; Xu, F. Needles in a Haystack: Tracking City-Scale Moving Vehicles From Continuously Moving Satellite. *Ieee Trans. Image Process.* **2020**, *29*, 1944–1957. [CrossRef]
7. Wei, J.; Sun, J.; Wu, Z.; Yang, J.; Wei, Z. Moving Object Tracking via 3-D Total Variation in Remote-Sensing Videos. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3506405. [CrossRef]
8. Piccardi, M. Background subtraction techniques: A review. In Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583), The Hague, The Netherlands, 10–13 October 2004; pp. 3099–3104.
9. Bruhn, A.; Weickert, J.; Schnörr, C. Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *Int. J. Comput. Vis.* **2005**, *61*, 211–231. [CrossRef]
10. Singla, N. Motion detection based on frame difference method. *Int. J. Inf. Comput. Technol.* **2014**, *4*, 1559–1565.
11. Shao, J.; Du, B.; Wu, C.; Yan, P. PASiam: Predicting Attention Inspired Siamese Network, for Space-Borne Satellite Video Tracking. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 1504–1509.
12. Yang, T.; Wang, X.; Yao, B.; Li, J.; Zhang, Y.; He, Z.; Duan, W. Small moving vehicle detection in a satellite video of an urban area. *Sensors* **2016**, *16*, 1528. [CrossRef] [PubMed]
13. Barnich, O.; Droogenbroeck, M.V. ViBe: A Universal Background Subtraction Algorithm for Video Sequences. *IEEE Trans. Image Process.* **2011**, *20*, 1709–1724. [CrossRef]
14. Yin, Q.; Hu, Q.; Liu, H.; Zhang, F.; Wang, Y.; Lin, Z.; An, W.; Guo, Y. Detecting and tracking small and dense moving objects in satellite videos: A benchmark. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5612518. [CrossRef]
15. He, Q.; Sun, X.; Yan, Z.; Li, B.; Fu, K. Multi-object tracking in satellite videos with graph-based multitask modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5619513. [CrossRef]
16. Li, S.; Zhou, Z.; Zhao, M.; Yang, J.; Guo, W.; Lv, Y.; Kou, L.; Wang, H.; Gu, Y. A Multi-task Benchmark Dataset for Satellite Video: Object Detection, Tracking, and Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5611021.
17. Wang, Q.; Zheng, Y.; Pan, P.; Xu, Y. Multiple object tracking with correlation learning. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3876–3886.
18. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
19. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
20. Maher, A.; Taha, H.; Zhang, B. Realtime multi-aircraft tracking in aerial scene with deep orientation network. *J. Real-Time Image Process.* **2018**, *15*, 495–507. [CrossRef]
21. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
22. 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.
23. Kalman, R.E. A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
24. Wang, Y.; Cheng, H.; Zhou, X.; Luo, W.; Zhang, H. Moving Ship Detection and Movement Prediction in Remote Sensing Videos. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *43*, 1303–1308. [CrossRef]
25. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
26. Xiao, C.; Yin, Q.; Ying, X.; Li, R.; Wu, S.; Li, M.; Liu, L.; An, W.; Chen, Z. DSFNet: Dynamic and static fusion network for moving object detection in satellite videos. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 3510405. [CrossRef]
27. Li, Y.; Jiao, L.; Tang, X.; Zhang, X.; Zhang, W.; Gao, L. Weak Moving Object Detection In Optical Remote Sensing Video With Motion-Drive Fusion Network. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 5476–5479.
28. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
29. Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; Vaughan, J.W. A theory of learning from different domains. *Mach. Learn.* **2010**, *79*, 151–175. [CrossRef]
30. Feng, J.; Zeng, D.; Jia, X.; Zhang, X.; Li, J.; Liang, Y.; Jiao, L. Cross-frame keypoint-based and spatial motion information-guided networks for moving vehicle detection and tracking in satellite videos. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 116–130. [CrossRef]
31. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. *Int. J. Comput. Vis.* **2018**, *128*, 642–656. [CrossRef]

32. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.-K.; Woo, W.-C. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In Proceedings of the Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015.
33. Date, K.; Nagi, R. GPU-accelerated Hungarian algorithms for the linear assignment problem. *Parallel Comput.* **2016**, *57*, 52–72. [CrossRef]
34. 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 October 2022; pp. 1–21.
35. Dai, Y.; Hu, Z.; Zhang, S.; Liu, L. A survey of detection-based video multi-object tracking. *Displays* **2022**, *75*, 102317. [CrossRef]
36. Cao, J.; Pang, J.; Weng, X.; Khirodkar, R.; Kitani, K. Observation-centric sort: Rethinking sort for robust multi-object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023, Vancouver, BC, Canada, 18–22 June 2023; pp. 9686–9696.
37. 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.
38. 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]
39. Zhou, X.; Koltun, V.; Krähenbühl, P. Tracking objects as points. In Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 474–490.
40. Wu, J.; Cao, J.; Song, L.; Wang, Y.; Yang, M.; Yuan, J. Track to detect and segment: An online multi-object tracker. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12352–12361.
41. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
42. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
43. Yu, F.; Wang, D.; Darrell, T. Deep Layer Aggregation. In Proceedings of the Default Cover Image 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2403–2412.
44. Perreault, H.; Bilodeau, G.-A.; Saunier, N.; Héritier, M. FFAVOD: Feature fusion architecture for video object detection. *Pattern Recognit. Lett.* **2021**, *151*, 294–301. [CrossRef]
45. Qiu, H.; Ma, Y.; Li, Z.; Liu, S.; Sun, J. BorderDet: Border Feature for Dense Object Detection. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020.
46. Scharstein, D.; Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42. [CrossRef]
47. Chang, J.-R.; Chen, Y.-S. Pyramid stereo matching network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5410–5418.
48. Yang, J.; Mao, W.; Álvarez, J.M.; Liu, M. Cost Volume Pyramid Based Depth Inference for Multi-View Stereo. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4876–4885.
49. Im, S.; Jeon, H.-G.; Lin, S.; Kweon, I.-S. DPSNet: End-to-end Deep Plane Sweep Stereo. *arxiv*, 2019; arXiv:1905.00538.
50. Laga, H.; Jospin, L.V.; Boussaid, F.; Bennamoun, M. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1738–1764. [CrossRef]
51. Teed, Z.; Deng, J. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
52. Dendorfer, P.; Osep, A.; Milan, A.; Schindler, K.; Cremers, D.; Reid, I.D.; Roth, S.; Leal-Taixé, L. MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking. *Int. J. Comput. Vis.* **2020**, *129*, 845–881. [CrossRef]
53. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arxiv*, 2014; arXiv:1412.6980.
54. Bernardin, K.; Stiefelhagen, R. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 246309. [CrossRef]
55. 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 13th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2016; pp. 17–35.