Multivehicle Object Tracking in Satellite Video Enhanced by Slow Features and Motion Features

Jialian Wu, Xin Su, *Member, IEEE*, Qiangqiang Yuan, *Member, IEEE*, Huanfeng Shen, *Senior Member, IEEE*, and Liangpei Zhang, *Fellow, IEEE*

***Abstract*— With the development of video satellites, multimoving object tracking in satellite video is possible and has become a new challenging task. The difficulties are mainly caused by the characteristics of satellite videos: 1) small objects; 2) low contrast between objects and background; and 3) background in a state of continuous motion. These characteristics make it difficult for the advanced multiobject tracking algorithms in the natural video to give full play to their advantages, resulting in vast false alarms, missed objects, ID switches, and low-confidence bounding boxes. To tackle these problems, a novel multimoving object tracking method considering slow features (SFs) and motion features has been proposed in this research, named SF and motion feature-guided multiobject tracking (SFMFMOT), which realizes the continuous tracking of moving vehicles in satellite videos. A nonmaximum suppression (NMS) module guided by bounding box proposals based on SFs is designed to assist the object detection part by utilizing the sensitivity of SF analysis to the changed pixels. While removing a large number of static false alarms and supplementing missed objects, it improves the recall rate by increasing the confidence score of the correctly detected object bounding boxes. In order to improve the tracking performance, a set of optimization strategies based on motion features and time accumulation information are proposed to smooth the trajectory, remove static false alarms, and duplicate bounding boxes. The proposed method is evaluated in three satellite videos and its superiority is demonstrated.**

***Index Terms*— Motion feature, multiobject tracking, object detection, satellite video, slow feature (SF).**

Manuscript received November 26, 2021; accepted December 20, 2021. Date of publication December 28, 2021; date of current version March 8, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 41922008, in part by the Excellent Youth Foundation of Hubei Scientific Committee under Grant 2020CFA051, and in part by the National Natural Science Foundation of China under Grant 61801332. *(Corresponding author: Xin Su.)*

Jialian Wu is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: 2015301610325@whu.edu.cn).

Xin Su is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: xinsu.rs@whu.edu.cn).

Qiangqiang Yuan is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China, and also with the Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: qqyuan@sgg.whu.edu.cn).

Huanfeng Shen is with the School of Resource and Environmental Science, Wuhan University, Wuhan 430079, China, also with the Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China, and also with the Key Laboratory of Geographic Information System, Ministry of Education, Wuhan University, Wuhan 430079, China (e-mail: shenhf@whu.edu.cn).

Liangpei Zhang is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China, and also with the Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: zlp62@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2021.3139121

I. Introduction

MULTIPLE object tracking or multiple target tracking (MOT or MTT) in satellite video is an important but very challenging task. Recently, video satellites have developed rapidly, and many countries and international organizations have been successively studying and launching video satellites to obtain very high-resolution (VHR) space-borne videos. For example, the SkySat-1 and SkySat-2 satellites launched by Planet, USA, can get 1.1 m of spatial resolution VHR satellite videos. Jilin No. 1 video satellite launched by China Changguang Satellite Technology Co. can get 0.92 m spatial resolution VHR satellite videos. As a new type of earth observation satellites, video satellites could gaze at a certain area to obtain the continuous dynamic picture of the object change. This raises the possibility of tracking ground and near-ground objects from a satellite perspective, such as vehicles, trains, planes, and so on. By using information obtained from a more global and broader perspective, multiobject tracking in satellite videos can play an essential role in the following scenarios [1]: 1) in the military aspect, it can monitor suspicious objects to obtain enemy intelligence [2]; 2) in the fire monitoring aspect, it can find the trend of wildfire and accurately locate firefighters [3]; and 3) in the traffic monitoring aspect, it can track and analyze the moving vehicles and other objects as well as realize density flow estimation [4]. In addition, it is a more economical way compared to ground-based video surveillance equipment, without installing and maintaining a large number of sensors and cameras. Meanwhile, its stability is also higher, avoiding the failure to collect information in emergencies such as power failure quickly.

At present, the object tracking research based on satellite video mainly focuses on the single object tracking task [1], [5]–[11]. To the best of our knowledge, there are relatively few studies on the topic of multiobject tracking based on satellite videos, which is mainly limited by the great challenge and the availability of data, including benchmarks for performance evaluation. Recently, a considerable research has grown up around the theme of multiobject tracking based on nature videos. According to the way the objects are initialized, most existing MOT methods can be classified into two categories: detection-based tracking (DBT) mode and detection-free tracking (DFT) mode. In addition, track-before-detect (TBD) mode is also used in a few methods. For the DBT

framework, given a video sequence, the object hypotheses are first obtained by detecting the moving objects in each frame and then they are associated with the existing trajectory to complete the tracking [12]–[14]. Unlike the DBT mode, DFT mode requires manually initializing a fixed number of objects in the first frame and then positioning these objects in subsequent frames [15]–[18]. It is often limited to a fixed number of specified objects and unable to handle new objects that appear in subsequent frames. By contrast, the DBT mode, which can automatically add new objects and delete missed objects, is more popular and widely used in various multiobject tracking algorithms. In the TBD mode, the class-agnostic object tracking is performed first, followed by a tracker-guided object detector. Due to the lack of prior information of specific categories, such methods tend to produce false alarms and missed detection, and generally need to add an object classification process to introduce prior knowledge to obtain specific category objects. To sum up, the DBT framework is chosen in this study to carry out research. As an attempt, the advanced multiobject tracking algorithm based on the DBT framework is directly used in satellite videos after retraining in this research. However, these results are not very encouraging, including vast false alarms, missed objects, ID switches, and low-confidence bounding boxes. The characteristics of satellite videos could be the contributing factor to the poor performance.

1) *Small objects.* Limited by the low resolution, moving objects in satellite videos can only occupy a few to more than a dozen pixels with a very low proportion in the whole image. As a result, there are few distinct features available to the vehicle object, and the discriminant features commonly used in the object tracking methods based on natural videos, such as texture and color, cannot play a role. For MOT tasks, it is not only difficult to detect all objects accurately, but also easy to follow the wrong object, resulting in a large number of ID switches.
2) *Low contrast between objects and background.* Due to the low resolution and complex background of satellite videos, the contrast between the object and the background is very low. On the one hand, this further increases the difficulty of extracting complete continuous objects from the background. On the other hand, objects that have been successfully detected can easily be removed due to low confidence.
3) *The satellite is in a state of continuous motion.* The motion of the satellite causes the background part of the image not to remain stationary, which increases the difficulty of dynamic object detection for the introduction of false alarms.

The problems related to camera movement, such as background changes, also exist in other mobile camera systems. Taking drones as an example, the effective problems can be divided into the following aspects [19]–[22]: 1) demand real-time; 2) demand lightweight; 3) camera motion problem; 4) target scale change problem; and 5) flexible operability.

In addition, the target detection and tracking task based on Wide Area Motion Imagery (WAMI) also has the problems of low resolution and small target [23]–[25]. The main challenges

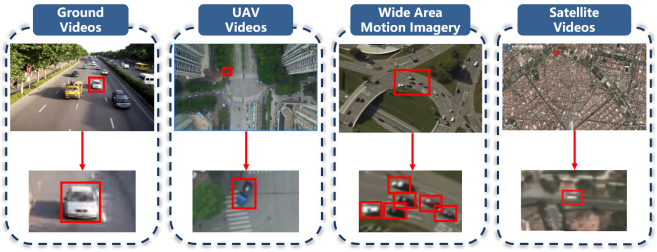


Fig. 1. Comparison of sample targets from different data acquisition platforms.

are as follows: 1) low frame rate; 2) low resolution; 3) camera motion problem; and 4) light changes, joint artifacts.

In contrast, the satellite-based video has different characteristics from drone videos and WAMI due to the extremely high orbital height and stable orbit of the satellite.

1) Low demand for real-time, lightweight. The on-board real-time processing is the development trend of the on-board remote-sensing image processing, but it is limited by the large width of the on-board image and the large amount of information contained in the image. At present, most on-board remote-sensing image processing adopts off-line processing and has low requirements on real-time and lightweight.
2) Camera motion problems are negligible. Due to the stability of the flight platform, there is almost no camera motion between adjacent frames of satellite video, so there is no need to implement the camera motion compensation strategy. Besides, camera compensation in UAV and WAMI usually computes the SIFT-like descriptor based on Harris corner points and then computes the transformation matrix through feature descriptor matching. The camera compensation method is not applicable due to the lack of appearance features of moving targets in satellite videos.
3) The scale of the target is almost fixed. Due to the stability of satellite altitude, the scale variation of the target is slight, so scale invariance is not a key consideration. Therefore, small target detection by multiscale feature mapping, commonly used in UAV-based object tracking, is unsuitable for satellite video.
4) Inflexible operability. The satellite sensor cannot be improved randomly according to the requirements as in the UAV, so it is not feasible to calculate the target position based on IMU, GPS, and other multisource information.
5) High frame rate. Compared with WAMI, which has a frame rate of 1–2 frames per second (FPS), the frame rate of satellite videos is as high as 20–30 FPS, so the target displacement between adjacent frames is small and there is no noticeable change in appearance.
6) Extremely low resolution. In contrast, objects handled by drones and WAMI still have some usable details, such as textures and colors, which are almost wholly lost in satellite videos, as shown in Fig. 1.

To sum up, the problem of tiny target and false alarms caused by background movement in multiobject tracking based on satellite videos have not been fully studied in the

applications based on UAV and WAMI. In addition, compared with WAMI, the satellite video has a higher frame rate and smaller target motion displacement, which can provide motion information as an auxiliary to explore better detection and tracking performance. Therefore, it is of great significance to design a multiobject tracking method suitable for small target characteristics of satellite videos.

In this work, we present a novel method for multivehicle object tracking in satellite videos that provides effective solutions to the aforementioned problems, named slow feature and motion feature guided multiobject tracking (SFMFMOT). Two steps of object detection and data association are studied, respectively, in this research. For the object detection part, based on the nonmaximum suppression (NMS) [26] post-processing operation, an improved NMS module guided by bounding box proposals based on slow features (SFs) [27], [28] is designed to realize the enhancement of object detection. Furthermore, to improve the accuracy of the preliminary tracking results obtained by data association, SFMFMOT deeply studies the moving features of the vehicle objects and false alarms and the time accumulation information. A series of tracking optimization strategies are proposed, including smoothing interrupt trajectory by the Kalman filter [29], removing most of the remaining static false alarms and deleting the duplicate bounding boxes. The proposed method is evaluated in three satellite videos and its superiority is demonstrated. The overall process of the proposed tracker is shown in Fig. 2.

The main contributions of this research are summarized as follows.

1) In this research, an effective multivehicle object tracking method for satellite videos is proposed, named SFMFMOT. The quantitative evaluation of actual satellite video datasets shows that the proposed method can significantly improve the accuracy of multiobject tracking on satellite videos.
2) An NMS module guided by bounding box proposals based on SFs is introduced to enhance object detection. The bounding box proposals for moving objects are obtained by utilizing SF analysis, which provides enhanced change detection. Furthermore, the adaptive combination of bounding boxes based on the improved NMS module is performed. Experimental results show that the enhancement processing can greatly improve the object detection performance, success to remove most false alarms and supplement some missed objects. Also, the recall rate is improved by increasing the confidence score of the correctly detected object bounding boxes with low confidence.
3) A set of tracking optimization strategies based on motion features and time accumulative information is proposed. These strategies achieve the removal of false alarms and duplicate bounding boxes, which improves the accuracy and integrity of object trajectories. Experiments show that the motion information of the objects and false alarms is very effective in the multiobject tracking task of satellite videos.

The rest of this article is organized as follows. In Section II, we briefly introduce the related work. In Section III, we detail the proposed multiobject tracking algorithm for satellite videos. The experimental results and discussion are presented in Section IV, which proves the effectiveness of SFMFMOT. Finally, we summarize this research in Section V. Appendix A supplements the detailed analysis of object detection performance and the selection of relevant parameters, and Appendix B supplements the analysis of difficult scenarios.

II. Background

A. Tracking in Very High-Resolution Satellite Videos

With the rapid development of VHR satellite videos, a new research field of single object tracking based on satellite videos has emerged gradually [1], [5]–[11]. Xuan *et al.* [5] combined the Kalman filter with the motion estimation algorithm of trajectory averaging to solve the problem of tracking failure when the moving object is partially or entirely occluded. Wang *et al.* [6] established the filter training mechanism of objects and background and built the object feature model based on the Gabor filter, so as to improve the recognition ability of weak feature objects. Besides, a tracking state evaluation index is employed to avoid tracking drift. HRSiam [10] uses a lightweight parallel network with high spatial resolution pruned based on HRNet [30] to obtain fine-grained appearance features to achieve accurate object tracking and positioning results in SiamRPN [31]. In addition, a pixel-level refined model is used to detect moving objects based on motion information, and adaptive fusion of tracking and detection results is performed to cope with challenging scenarios such as occlusion. In view of the small size of the objects in satellite videos, PASiam [11] uses a shallow full convolutional siamese network to obtain the fine-grained appearance features of the object. Moreover, the Gaussian Mixture Model [32] and Kalman filter are utilized to correct the position of the object to deal with occlusion and blur. In [33], a multimorphological-cue-based discrimination strategy is used to detect moving targets from the complex background in the satellite videos.

To our best knowledge, due to the characteristics of satellite videos and the challenge of multitarget tracking tasks, there is scarce research on deep learning-based multiobject tracking tasks using satellite videos. The object detection method for satellite videos based on traditional methods [4], [34] generally regards object detection as foreground and background segmentation. Background subtraction, frame difference, and other methods were used to obtain foreground pixels, and then the threshold segmentation and morphological processing were used to obtain vehicle detection results. However, traditional methods are easily affected by moving background and brightness changes, and these scenes are ubiquitous in satellite videos, resulting in many false alarms and missed objects.

Zhang *et al.* [35] proposed a two-step global data association method based on probability, which solved the detection inconsistency caused by the low spatial resolution of satellite videos. Methods based on probability generally focus on the data association part and obtain continuous trajectory based on rigorous and complex mathematical probability model,

but ignore the improvement of detection part. However, the performance of the detection part has an important impact on multiobject tracking based on the DBT framework.

Jie *et al.* [36] proposed a framework combining a keypoint-based cross-frame detection network (CKDNet) and a spatial motion information-guided tracking network (SMTNet) for the detection and tracking of moving vehicles in satellite videos. Deep neural network has a strong capability of feature representation and can enhance both the detector and the tracker. However, this method adopts a double LSTM structure with high computation cost and requires future frames as input, so it cannot be processed online.

B. Visual Multiobject Tracking

Most existing MOT methods can be classified into two categories according to how the object is initialized: the DBT mode and DFT mode.

In addition, the TBD mode is also used in a few methods [37]–[39]. Different from the DBT framework, Wong *et al.* [37] do not introduce any prior information about the object in the detection and tracking stage, but track all possible objects of interest and independently track each target with a separate shape estimation filter (SEF). Due to the lack of prior information, the tracking system does not know the scale range of the target, which only connects a group of consistent observation results (in shape, position, and velocity) with a single SEF. It is easy to regard a dense group as a single target and generate false positives and false negatives, so it is inappropriate for satellite videos with small and dense targets. The mechanism to track all possible objects of interest would also be computationally heavy in the case of the large width of satellite videos. Overall, the TBD mode is not suitable for the small object tracking task in satellite videos.

Considering that the current mainstream research framework of MOT algorithms is Detection-based Tracking, the following introduction mainly focuses on the DBT framework.

For the DBT framework, the quality of the object detector affects tracking performance to a large extent. With the rapid development of deep learning in object detection in recent years, a large number of existing object detection methods have been used in MOT to perform object initialization and object location of each frame. Classical methods fall into the following two categories: The R-CNN serial algorithms [40]–[42] based on the two-stage framework, which achieve high detection performance, but are difficult to meet the real-time requirements. The YOLO serial algorithms [43]–[46] based on the single-shot framework, which greatly accelerate the speed of detection, but the object position is not very accurate, resulting in relatively low detection accuracy. Most of the methods based on the two structures adopt the anchor-based approach, introducing a large number of hyperparameters. According to [47], using anchors often causes feature unfairness and feature conflicts between the detection task and Re-ID task, leading to the decrease of accuracy. In contrast, the anchor-free structure significantly reduces the number of hyperparameters by omitting the sliding window process and greatly improves the training speed without loss of accuracy [48]–[51]. For example, CenterNet [52] abandons the use of axis alignment boxes and models the object as the center point of the bounding box. Because there is only one central point per object, nonmaximum suppression (NMS) [26] postprocessing operation is no longer required, which makes the method end-to-end differentiable, simpler, faster, and more accurate.

For the data association part, the traditional techniques always need to build a complex model, such as multihypothesis tracker (MHT) [53] and joint probabilistic data association filter (JPDAF) [54]. At present, most trackers use the following common strategy. First, the cost function is calculated based on a single measure such as appearance feature, direct distance, intersection over union (IoU) distance, and so on, or a combination of multiple measures. Then, the matching algorithm such as the Hungarian algorithm [55] is used to realize the data association between the detection bounding boxes and the existing trajectories. There are also a few approaches [56]–[58] that utilize more complex association strategies, such as group models and recurrent autoregressive networks.

Among many multiobject tracking methods, the multiobject tracker FairMOT [47] based on the DBT framework stands out, which obtains advanced multiobject tracking performance in the natural video. FairMOT uses CenterNet [52] as the object detector and then inputs the detection results into the Re-ID module to obtain the Re-ID features. By calculating the cost function based on Re-ID features and IoU measure, data correlation is realized through the Kalman filter and Hungarian matching [55]. Based on this, FairMOT is used as the benchmark in this research. Nevertheless, limited by the characteristics of satellite videos, it is difficult for FairMOT to achieve the expected performance in satellite videos, which produces a large number of false positives, missed objects, ID switches, and low-confidence bounding boxes.

In this research, the SF analysis (SFA) [28] method is used to increase the contrast between objects and background, so as to obtain the bounding boxes proposals of the moving object. Next, an NMS module guided by bounding box proposals is designed to realize the removal of false alarms and the supplement of missed objects in object detection results. In addition, by using the bounding box proposals to verify the reliability of the detection bounding boxes, the module increases the confidence score of the correctly detected bounding boxes, which significantly improves the recall rate of the object detection. Furthermore, a series of corresponding strategies are proposed to solve the problems of trajectory interruption, false alarms, and duplicate bounding boxes in data association. By using the motion information and time accumulative information, these strategies greatly improve the robustness and accuracy of multiobject tracking in satellite videos.

III. Proposed Approach

A. Overall Architecture

Based on the MOT method FairMOT, this research proposes a novel MOT algorithm, SFMFMOT, designed for satellite videos. The architecture of the method is shown in Fig. 2,

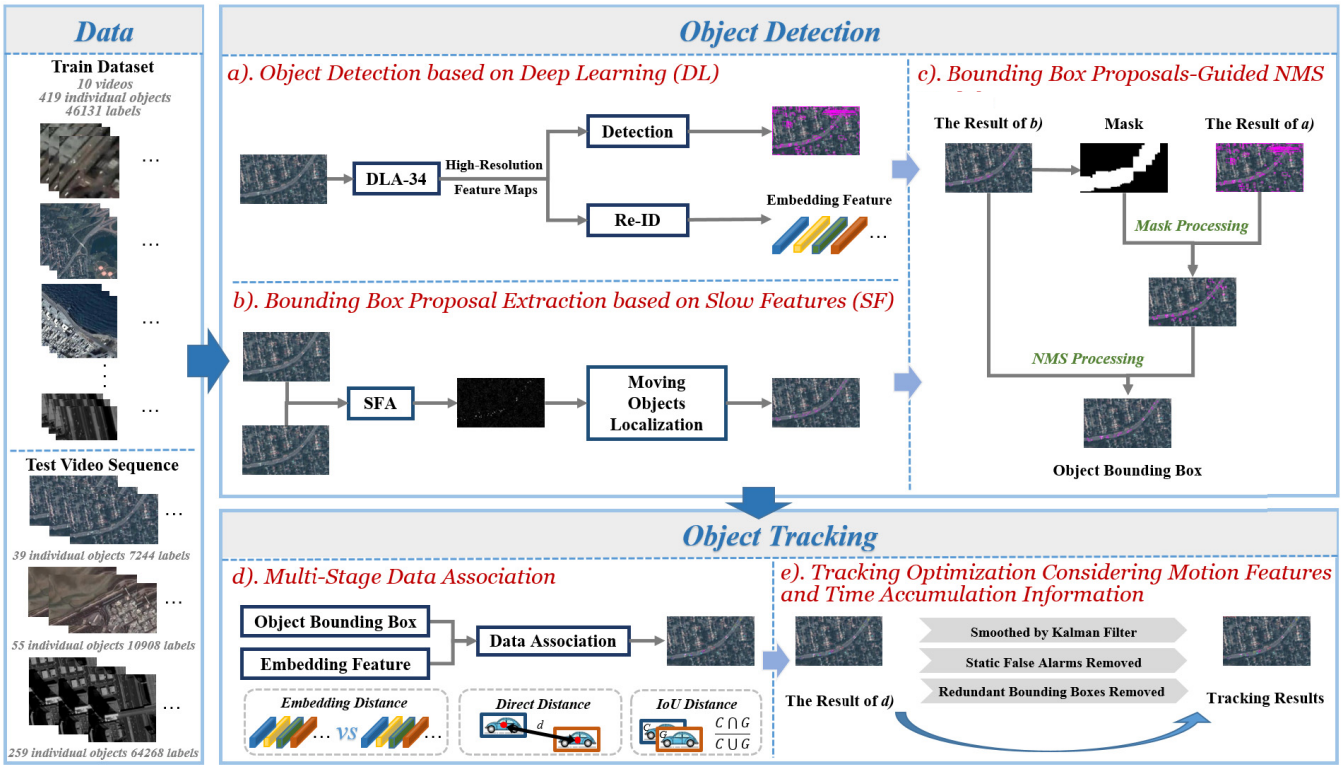


Fig. 2. Overview of the proposed SFMFMOT tracker. The dataset used in this study is provided on the left, including ten training videos and three test videos. The network is divided into two parts: object detection and object tracking. The detection part is composed of two branches: object detection based on deep learning and bounding box proposals extraction based on SFs. The results of the two parts are combined by the bounding box proposals-guided NMS module to obtain the final more accurate object bounding boxes. Then, the detection bounding boxes and Re-ID embedding features are used to perform multistage data association to obtain preliminary tracking results. Finally, the motion features and time accumulation information are used to improve the accuracy and completeness of the tracking part.

which consists of two modules: detection and tracking. The detection part includes two branches. One branch uses the enhanced version DLA-34 [52] of deep layer aggregation (DLA) [59] to generate high-resolution feature maps and further obtains preliminary object detection results and Re-ID embedding features. The other branch performs bounding box proposal extraction based on SFs, which requires inputting a set of image pairs, adjacent, or with N frames interval. The moving object bounding box proposals can be obtained after extracting the moving object pixel points between the two input images based on SFs. Finally, the results of the two branches are combined through a bounding box proposals-guided NMS module to obtain more accurate object detection results. In the tracking part, the object bounding boxes and the Re-ID embedding features are used as input to execute the data association module to realize the correlation matching between the detection boxes in the current frame and the existing trajectories. Then, the object motion features and time accumulative information are analyzed to optimize and correct the preliminary tracking results, which further improve the accuracy and integrity of multiobject tracking.

B. Enhanced Object Detection Based on Slow Features

1) Object Detection Based on Deep Learning: In this research, ResNet-34 [60] is used as the backbone to execute feature extraction to obtain the high-resolution feature maps, which can achieve a good balance between accuracy and speed. Refer to [52], the variant DLA-34 of DLA [59] is applied to the backbone network. By adding skip connections between low-level features and high-level features, the adaptation to different scales is realized. In addition, the deformable convolutional layer is used to replace the convolutional layer in the upper sampling module to realize the adaptive adjustment of the receptive field with the change of the object scale. Denote the size of the input image as $H \times W$, and the size of the output feature map is $C \times H/4 \times W/4$.

High-resolution feature maps processed by the backbone network flow to object detection and Re-ID identity embedding branches, as shown in Fig. 3. The object detection branch draws on the idea of anchor-free, referring to CenterNet [52], which regards object detection as a center point-based bounding box regression task on the feature map. It is composed of Heatmap Head, Center Offset Head and Box Size Head, with dimensions of $1 \times H \times W$, $2 \times H \times W$, and $2 \times H \times W$, respectively. The Heatmap Head outputs the heatmap, in which the response value of each pixel decays exponentially with the distance between the position of the pixel and the center of the object. This branch is used to estimate the object's center position by extracting the key peak points through NMS. The Center Offset Head is responsible for positioning the object more accurately, mainly to deal with the quantization error

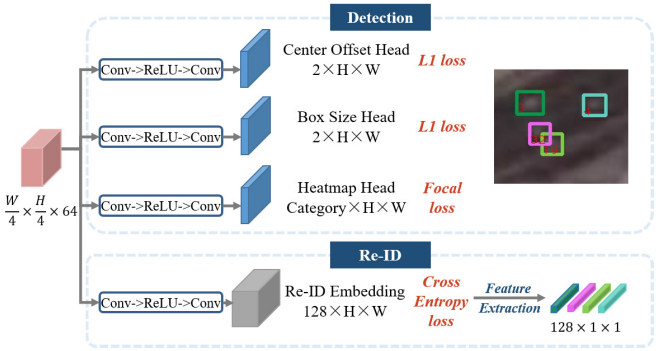


Fig. 3. Diagram of object detection branch and Re-ID embedding branch. The high-resolution feature map obtained by the DLA-34 network is used as the input of each branch. Each head branch consists of two convolutional layers sandwiched by a ReLU activation layer. In the detection branch, the L1 loss is employed in the center offset head and box size head, while the focal loss is employed in the heatmap head. In the Re-ID branch, the cross-entropy loss is used. And Re-ID features with a size of 128 × 1 × 1 at the center point of the predicted object are extracted for tracking. Only vehicle categories are studied in this article, so the value of the category is set to 1.

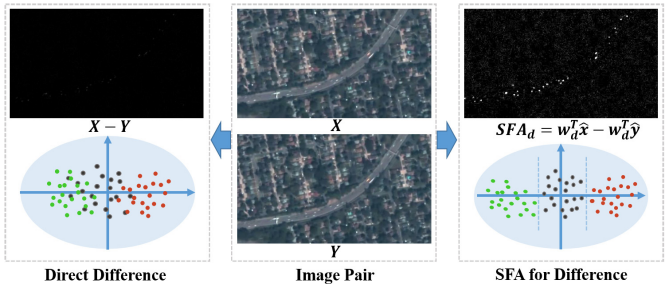


Fig. 4. Direct difference versus SFA for difference. The red and green dots are the changed pixels, while the gray dots are the unchanged pixels. After SFA processing, the separability between unchanged pixels and changed pixels is greatly improved.

introduced by the step size of 4 in the feature mapping. The box size head is adopted to estimate the height and width of the object bounding boxes. The identity embedding branch is used to extract the Re-ID features at the center of the estimated object to distinguish different objects, with the dimension of $128 \times H \times W$. In addition, for each detected object, a confidence score is set, which indicates the possibility that the object belongs to the vehicle class. In particular, only bounding boxes with a confidence score greater than the specified threshold are retained (0.1 is used in this article). Refer to Appendix A for detailed instructions on the selection of this parameter.

In the actual study, we found that the above deep learning-based object detection method would produce a large number of false alarms, missed objects, and ID switches. Besides, due to the lack of distinct features and the poor contrast between objects and background, some correctly detected moving object bounding boxes have very low confidence scores, which would be filtered out by the confidence score screening. In the light of the significance of object detection results for the overall performance of MOT, bounding box proposals based on SFs are extracted in this research to improve the accuracy and integrity of object detection. The specific operation is introduced as follows.

2) Bounding Box Proposal Extraction Based on Slow Features: This section is mainly divided into the following two steps: 1) moving object pixel points extraction based on SFs and 2) obtain the object bounding boxes according to the results of the first step, as the bounding box proposals for object detection.

First, moving object pixel points are extracted based on the SFA method proposed in [28]. This method makes the SFA method more automatic by iterating and repeating the process of SF processing and selecting invariant pixels for training and learning transformation matrix. Different weights are applied to each pixel in the iteration, where large weight values are assigned to unchanged pixels and small weight values are assigned to changing pixels. In this way, unchanged pixels play an increasingly important role in learning strategies, and the separability between unchanged pixels and changed pixels becomes stronger. The core idea of the slowness principle is to find and extract slowly changed output signals from rapidly changed input signals [27]. Unlike the traditional SFA problem, the object processed here is not a sequence signal that changes with time, but two image frames of different moments. Mathematically, the corresponding pixels in two image frames X and Y at different times are, respectively, represented by $x^i = [x_1^i, \ldots, x_M^i]$, $y^i = [y_1^i, \ldots, y_M^i]$, where M is the number of image bands and i is the number of pixels. As shown in Fig. 4, for the input image pairs, the goal is to highlight the variant pixels by suppressing the invariant pixels, so that the part that changes between the two frames, namely the moving object, can be separated more correctly.

In the general linear case, the objective function can be expressed as follows:

$$\min_{w_d} : \frac{1}{n}\sum_{i=1}^{n}\left(w_d^T \hat{x}_i - w_d^T \hat{y}_i\right)^2 \tag{1}$$

where w^T is expressed as the transpose of the transformation vector w. Subscript d represents the number of bands. $\hat{x}_d^i$ and $\hat{y}_d^i$ represent the normalized pixel value of image X and image Y using zero mean value μ and unit variance σ, respectively. n represents the total number of pixels.

Furthermore, by using the generalized eigenvalue problem, w can be calculated. Finally, the difference between the two images after conversion can be calculated as follows:

$$\mathrm{SFA}_d = w_d^T \hat{\boldsymbol{x}} - w_d^T \hat{\boldsymbol{y}} \tag{2}$$

where subscript d represents the number of bands, and there are three bands in this research. Taking a random image in Video 1 as an example, the results of three bands processed by the SFA method, namely SFA_d in Formula 2, are shown in Fig. 5. It can be seen that the change information is mainly concentrated in the first band, so we use the result of the first band as the extraction result of moving object pixel points, which is denoted as I_{SF}.

Ideally, we hope that the result I_{SF} obtained based on SF processing only contains moving object pixel points. In fact, under the influence of light changes and satellite motion,

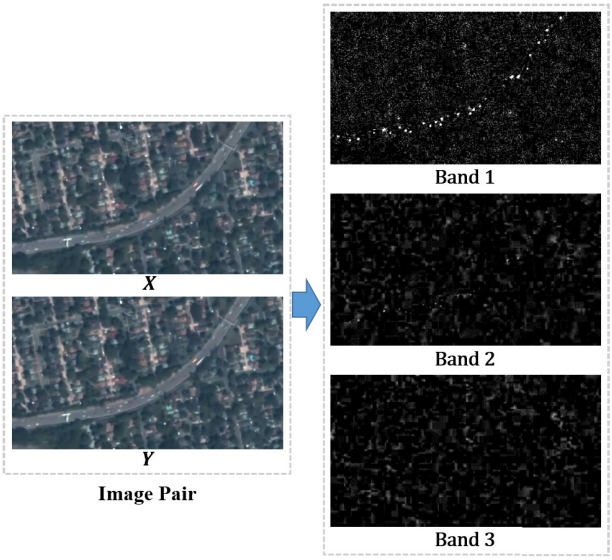


Fig. 5. Result diagram of the SFA method for each band. Bands 1–3 correspond to three sequential bands of the difference map between image X and image Y processed by the SFA method, respectively.

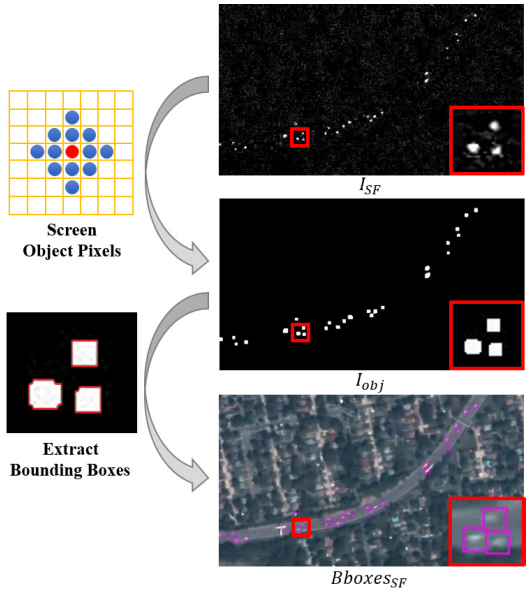


Fig. 6. Diagram of the process of obtaining the object bounding box proposals from the result of moving object pixel points extraction based on SFs. The red dot in the yellow grid represents the current point, and the blue dots represent the 12 adjacent pixel points around it. After contour extraction and minimum bounding box extraction, $\text{Bboxes}_{\text{SF}}$ can be obtained from I_{obj}.

background changes cannot be completely suppressed. There is still a lot of interference information in the processing result I_{SF}, namely background pixels. To obtain accurate moving object bounding box proposals, a vehicle object pixel point screening measure (denoted as OPS) is designed to eliminate interference information, as shown in Fig. 6.

First, a binarization image (donated as I_b) of the result I_{SF} is obtained through the Otsu threshold segmentation method. The Otsu method is also called the maximum between-class variance method [61]. Based on global image information, the optimal segmentation threshold is obtained by maximizing the between-class variance of each pixel class after segmentation. It is worth noting that, after experimental analysis, three times the threshold T calculated by the Otsu algorithm is employed to perform the segmentation task, namely $3T$. By observation, it can be found that the moving vehicles in the result I_{SF} are generally presented in the form of changed pixel blocks, as opposed to the changed background which appears as some discrete single-pixel points. Based on this, 12 surrounding pixels are chosen as its adjacent reference area for each changed pixel. Next, perform the vehicle object pixel point screening on the binary image I_b, this is, if more than nine points in this reference area belong to changed pixels, the central changed pixel is retained. The screening result is donated as I_{obj}, in which the white pixels are regarded as the location of the vehicle objects. By further extracting the white pixel area in I_{obj}, the bounding box proposals of the moving vehicle objects can be obtained, which are denoted as $\text{Bboxes}_{\text{SF}}$. $\text{Bboxes}_{\text{SF}}^{i}$ records the coordinates of the upper left corner and the lower right corner $[x_1, y_1, x_2, y_2]$ of the ith bounding box i. The whole process of bounding box proposal extraction can be expressed by the following formula:

$$\text{Bboxes}_{\text{SF}} = f_{\text{Box}}(f_{\text{OPS}}(f_{\text{Binarize}}(I_{\text{SF}}))) \tag{3}$$

where f_{Binarize} represents the binarization step, f_{OPS} represents the object pixel point screening step, and f_{Box} represents the bounding box extraction step.

Furthermore, it can be found that the extracted bounding box proposals have roughly the same size after implementing the screening measures of vehicle object pixel points. Therefore, a fixed size $R \times R$ is set for the bounding boxes in $\text{Bboxes}_{\text{SF}}$. The optimal value of R varies from video to video. See the experimental analysis section for details.

3) Bounding Box Proposals-Guided NMS Module: Through observation, we found that the bounding box proposals $\text{Bboxes}_{\text{SF}}$ based on SFs are less accurate than the correctly detected bounding boxes in the deep learning-based object detection $\text{Bboxes}_{\text{DL}}$. Based on this, a bounding box proposals-guided NMS module is designed to combine the object bounding boxes of the two parts. It is dominated by object detection results based on deep learning and supplemented by the bounding box proposals based on SFs. The specific process is shown in Fig. 2.

The NMS module is mainly performed in two parts. First, an object mask image is generated on $\text{Bboxes}_{\text{SF}}$, named as I_{mask}, to perform a filter operation on $\text{Bboxes}_{\text{DL}}$, as shown

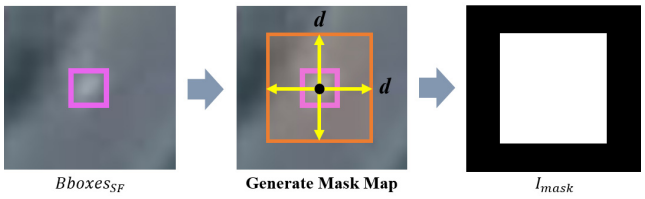


Fig. 7. Diagram of the generating mask. For each proposed bounding box, a $d \times d$ mask region is generated centered on its center point. Set d to 0.2 times the image area, which is larger than the area of the bounding box.

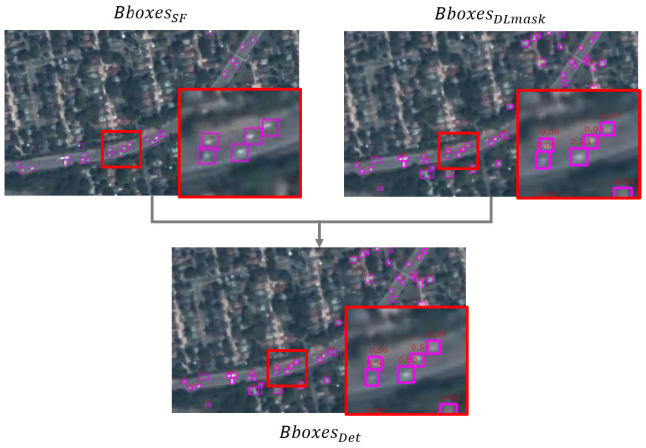


Fig. 8. Diagram of the bounding box proposals-guided NMS module. It takes into account the difference in the accuracy of bounding boxes, which is measured by the center point error between the ground truth and the predicted bounding box. The red number on the box shows the confidence score belonging to the vehicle. Only the detection boxes with a confidence score higher than 0.1 can be kept. As can be seen, after NMS module processing, the confidence score of the correct bounding box increases to 0.5.

in Fig. 7. At the same time, the ranges of the area, aspect ratio, length, and width are restricted to delete the wrong objects that are too large, too small, or too long. Here, the maximum aspect ratio is set to 4, the width and height are limited to between 5 and 45 pixels, and the maximum area of the bounding box is set to $(45 \times 45)/2.5$ pixels. The bounding boxes after filtering are recorded as $\text{Bboxes}_{\text{DLmask}}$. By making use of the characteristic that moving object pixel points is only sensitive to the changed part, most static false alarms in $\text{Bboxes}_{\text{DL}}$ can be filtered out. It can be expressed as follows:

$$\text{Bboxes}_{\text{DLmask}} = f_{\text{Filter}}(\text{Bboxes}_{\text{DL}}, I_{\text{mask}}). \tag{4}$$

In the second part, by executing the NMS module guided by the object bounding box proposals $\text{Bboxes}_{\text{SF}}$, some of the missed objects in $\text{Bboxes}_{\text{DLmask}}$ are supplemented, as shown in Fig. 8, while the processing result is denoted as $\text{Bboxes}_{\text{Det}}$. This step follows the principle that $\text{Bboxes}_{\text{DLmask}}$ is dominant and $\text{Bboxes}_{\text{SF}}$ is auxiliary. If an object is only detected by $\text{Bboxes}_{\text{SF}}$ or $\text{Bboxes}_{\text{DL}}$, the bounding box will be retained without changing the score. For the object detected in both parts, only the detection result in $\text{Bboxes}_{\text{DLmask}}$ is retained. Furthermore, it is reasonable to assume that the probability of the object being correct is greater than that detected only by one part. Based on this, the confidence score of the bounding box is increased to 0.5 accordingly. A detailed analysis of the choice of this re-set confidence score is provided in Appendix B. It can be expressed as follows:

$$\text{Bboxes}_{\text{Det}} = f_{\text{NMS}}(\text{Bboxes}_{\text{DLmask}}, \text{Bboxes}_{\text{SF}}) \tag{5}$$

where f_{NMS} represents executing the bounding box proposals-guided NMS module.

After performing the bounding box proposals-guided NMS module, false alarms are greatly suppressed, and some missed objects are supplemented. Besides, the correctly detected bounding boxes can be retained with a higher confidence score.

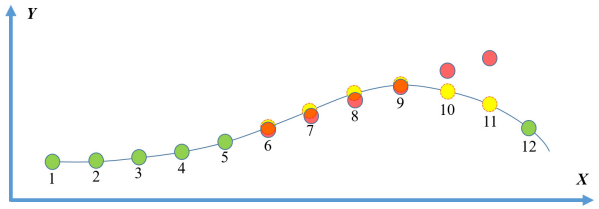


Fig. 9. Diagram of the filling of interrupt trajectory with the prediction results of the Kalman filter. The trajectory over time is shown from left to right, and the number below the object indicates the life span of the trajectory. The green ball represents the object tracked successfully, the yellow ball represents the tracking is interrupted, and the red ball represents the object position predicted based on the Kalman filter.

The great promotion of the overall detection accuracy provides better preparation for the follow-up multiobject tracking part.

C. Multiobject Tracking Considering Motion Features and Time Accumulation Information

1) Multistage Data Association: The association matching part adopts the standard online tracking algorithm with a multistage matching strategy (refer to [12]). In the first frame, the trajectories are initialized using the detection bounding boxes. In the subsequent frame, the detection bounding boxes are matched with the existing trajectories according to the cost matrix in three steps. First, the cost matrix calculated by the Re-ID features is used to perform matching. Second, the IoU metric is chosen to calculate the cost matrix for the matching, which is performed on the detection bounding boxes and trajectories that have not been matched in the first step. Third, the matching is performed once again according to the cost matrix calculated by the IoU metric, executing only on those not previously matched. Note that since the cost matrix calculated in the first step does not consider spatial information, the sensitivity processing for excessive distance matching is added. For each trajectory, the Euclidean distance between the position of the matching detection bounding box in the current frame and the position predicted by the Kalman filter is calculated. When the distance is overlarge, the corresponding value in the cost matrix is set to infinity to suppress the unreasonable matching with a large spatial distance span. Moreover, the appearance features are also updated at each time step to handle appearance changes. The matching process will not be introduced in detail here (refer to [12]).

2) Trajectory Smoothing Based on Kalman Filter: Interrupted trajectories can be found in the preliminary tracking results. A primary reason is that the bounding box of the trajectory is not detected in the object detection stage of the current frame. Considering the continuity of motion, the Kalman filter is used to predict the position of the object to fill in the missed part, which improves the integrity of trajectories, as shown in Fig. 9. The Kalman filter as an efficient autoregressive filter, powerful and universal, can estimate the state of a dynamic system from combined information with many uncertainties. The extremely high acquisition height of satellite videos makes vehicles move slowly, and the motion direction of the vehicle

has little difference between adjacent frames. Based on this, we assume that most vehicles show linear motion, which is consistent with the assumption made in the object motion model used by the Kalman filter. In this research, based on the known motion information of the previous frame, the Kalman filter can predict the position and speed of the object in the current frame. Of particular concern is as the predicted length increases, the probability of object drift increases, so it is necessary to select an appropriate filling length. See the experimental part for detailed analysis.

3) False Alarm Removal Based on Motion Features and Time Accumulation Information: Although the result of bounding box proposal extraction based on SFs has been used to generate a mask to filter out false alarms, there are still some static false alarms. It could be argued that the issue is due to the fact that the region of the mask area is set larger than the scope of the bounding box, which will inevitably retain some false alarms around the moving objects.

Ideally, the main difference between static false alarms and moving objects lies in the different motion states—the objects are in a slow-motion state, while the static false alarms always stay in the same position. However, due to the background movement, light changes, and other influences, static false alarms will appear slight jitter and not always in the same position. Moreover, in satellite videos, the vehicle object moves very slowly with normal brake pauses. Consequently, if the velocity or displacement threshold is directly and simply adopted for screening false alarms, some moving objects will be removed by mistake. In addition, it can be found that as time progresses, the amplitude of the false alarm jitter presents an upward trend, and the same false alarm always appears continuously in the whole video and keeps the same ID value. Based on this observation, a static false alarm removal strategy is designed in this research. By utilizing the motion characteristics of false alarms and the time accumulated information of removal, the removal rate of false alarms is improved without accidentally deleting the moving object. This strategy adopts the three-stage judgment principle, as shown in Fig. 10. In particular, the remove-num property is set for each track to record the number of times the track has been removed as a false alarm so far in the video, with an initial value of 0. In the first stage, for trajectories that have been continuously tracked for four frames or more, the velocity change between adjacent frames in the x and y directions is judged. If the absolute values in both directions are less than 0.1 pixels, the object is considered as a static false alarm and removed from the current frame, while its remove-num value is incremented by 1. It is important to note that the removal strategy will only be implemented when the track has been continuously tracked for four frames or more. For the reason that some moving objects are not very stable in the initial tracking, it is easy to mistakenly remove moving objects if the tracking length requirement is not set. In the second stage, if the condition of the first stage is not met, but remove-num ≥ 5 and the absolute values of the velocity change in both directions are less than 0.2 pixels, the object is also considered as a static false alarm. In the third stage, if the conditions of the above two stages are not met, but remove-num ≥ 50 and the absolute values

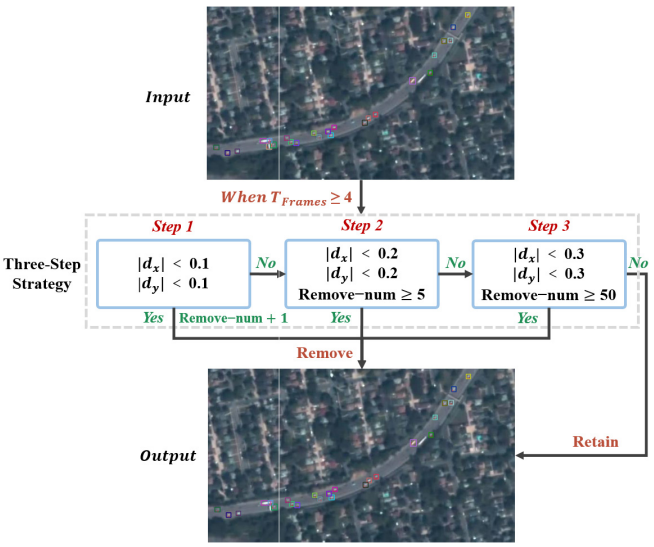


Fig. 10. Diagram of the three-stage strategy to remove false alarms. d_x and d_y represent the velocity change of the object between adjacent frames in the x and y directions. The remove-num property records the number of times the track has been removed as a static false alarm, with an initial value of 0.

of velocity change in both directions are less than 0.3 pixels, regard the object as a static false alarm, too.

The remove-num attribute plays a vital role in the overall strategy, and to ensure rigor, its value is only adjusted in the first stage when the velocity limit is the most stringent. With the increase of remove-num, the possibility that the trajectory belongs to a static false alarm gradually increases so that the velocity change threshold can be relaxed. By adapting to the change of false alarm motion state and employing the accumulated removal information of previous frames, the three-step judgment principle achieves a high false alarm recognition rate. And it can well handle the phenomenon that the drift degree of the static false alarm increases over time. In addition, to reduce the damage caused by the wrong removal of the motion trajectories, the false alarm removal strategy only removes the tracking result of the trajectory in the current frame. In the next frame, the trajectory will be added to the data association step again.

4) Duplicate Bounding Boxes Removal Considering Motion Features: In the preliminary tracking results, in addition to the static false alarms mentioned above, there are also some duplicate bounding boxes, that is, some objects have two bounding boxes.

One possible explanation for this might be the conflict between the bounding box supplemented by the bounding box proposal based on SFs and the prediction bounding box by the Kalman filter. Specifically, when the trajectory S is interrupted, a moving object bounding box proposal Box_{SF} based on SFs will supplement it. However, Box_{SF} fails to match with the corresponding interrupt trajectory S during the data association phase and is assigned a new ID as a new object. Then, the trajectory smoothing strategy based on the Kalman filter is performed, obtaining a prediction bounding box Box_{KF} with the same ID value as the trajectory S. Therefore, redundancy is formed between Box_{SF} and Box_{KF}. In this case, it is clear that

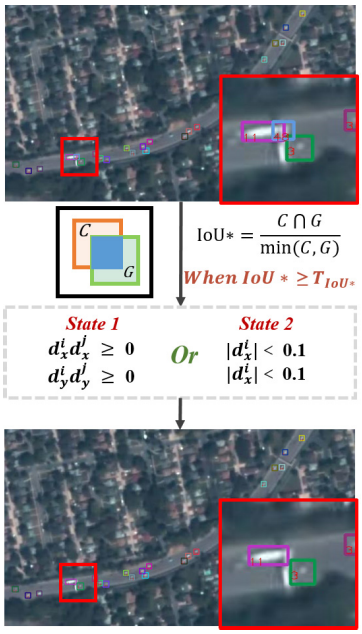


Fig. 11. Diagram of the removal of duplicate bounding boxes. C and G represent the area of the orange and green bounding boxes, respectively, and the blue area represents the overlap of the two bounding boxes. (d_x^i, d_y^i) and (d_x^j, d_y^j), respectively, represent the velocity changes of the bounding box Box_i and Box_j in the x and y directions. When IoU* is greater than the specified threshold $T_{\text{IoU}*}$, the motion state of the bounding box is judged. If one of the two states is satisfied, the redundancy box removal step is performed. For example, the bounding boxes with ID 11 and ID 45 in the figure detect the same object, and the bounding box with ID 45 is removed after processing.

the bounding box Box_{SF} based on SFs is not accurate enough and needs to be removed. Given that the newly assigned ID value must be greater than the ID value of the trajectory S, it is reasonable to maintain the bounding box with a smaller ID value, that is, the more accurate bounding box Box_{KF}.

In addition, there are also some cases of duplicate bounding boxes due to false alarms. To deal with both the two cases, a duplicate bounding box removal strategy is designed based on the NMS [26] method, in which the motion feature is introduced to improve the accuracy, as shown in Fig. 11.

First, for each bounding box in the current frame, calculate the IoU* metric between the ith bounding box Box_i and all other bounding boxes $\text{Box}_j (j \neq i)$ sequentially. It should be noted here that what we calculate is not the IoU metrics in the traditional sense, and the denominator only contains the smaller area in the two bounding boxes, as shown in Fig. 11. This is designed according to the actual situation, when the size of two duplicate bounding boxes is greatly different, the traditional calculation means of IoU will get a small value, so that the duplicate box cannot be screened out. By adopting the improved calculation mean, this problem can be solved.

However, if the IoU* threshold is simply directly used for screening, some moving objects will be removed by mistake. For example, when two different moving objects are very close, the threshold judgment for IoU* will also be met. In this case, additional information is needed to filter further. Motion information is introduced to deal with this problem in this research. Considering that if the duplicate box is caused by repeated detection of the same object, the motion direction of the two bounding boxes should be the same. If the static false alarm causes the duplicate box, it should only sway in place at a very slow pace. Thus, for the bounding box pair that satisfies IoU* threshold filtering, the following judgment will be made on its motion state: 1) The motion directions of the two bounding boxes are the same in x and y directions; 2) The velocity change of the bounding box Box_i in both x and y directions is less than 0.1 pixel. When one of these two states is satisfied, the removal of the duplicate box will be carried out. When the first case is met, the bounding box with the larger ID will be deleted, and when the second case is met, the bounding box Box_i will be deleted.

Fig. 12. Diagram of three test videos. Video 1 and video 2 are from Jilin No. 1 satellite video, and video 3 is from SkySat-1 satellite video.

IV. Experimental Results

A. Datasets and Metrics

1) Datasets: As new research, multiobject tracking of satellite videos has no public data such as CalTech [62], MOT16 [63], CUHK-SYSU [64], and PRW [65] in natural video object tracking. For better research and evaluation, a multivehicle object tracking dataset of satellite videos has been labeled in this research, named SateMVT, consisting of ten videos for training and three videos for testing. The data composition is introduced as follows.

a) Jilin No. 1 satellite video dataset.: The dataset was constructed based on Jilin No. 1 satellite video, and the original video frame format is 4096 × 2160 pixels. We cropped 9 videos for training and two videos for testing with a resolution of 1.13 m and a frame rate of 25 FPS. The frame format of training videos is 643 × 667 pixels, including 1998 frames, with 168 individual objects and 14 758 labels. The frame format of the two test videos is 640 × 360 pixels, with a duration of 300 frames for each video, as shown in Fig. 2.

b) SkySat satellite video dataset.: The dataset was constructed based on Skysat-1 satellite videos, and the original video frame format was 1920 × 1080 pixels. Two regions were cropped for training and testing, and each video was composed of 1800 frames at 30 FPS with a spatial resolution of 1.1 m.

TABLE I

OVERVIEW OF EVALUATION METRICS. ↑ (RESP. ↓) MEANS THAT THE PERFORMANCE IS BETTER WHEN THE VALUE IS GREATER (RESP. SMALLER)

Metric	Description	Note
MOTA	Multiple Object Tracking Accuracy. Combine false negatives, false positives and identity switches.	↑
MOTP	Multiple Object Tracking Precision. The misalignment between the estimated positions and the ground truth.	↓
IDF1	Identification F-Score. The harmonic mean of IDP and IDR.	↑
IDP	Identification Precision. Precision of object ID identification in each bounding box.	↑
IDR	Identification Recall. Recall of object ID identification in each bounding box.	↑
Precision	Ratio of correctly matched detections to total result detections.	↑
Recall	Ratio of correctly matched detections to ground-truth detections.	↑
MT	Mostly tracked objects. Percentage of ground-truth trajectories which are covered by track hypotheses for more than 80% of their length.	↑
ML	Mostly lost objects. Percentage of ground-truth trajectories which are covered by track hypotheses for less than 20% of their length.	↓
IDs	The total number of the identity switches.	↓
FM	Number of times that trajectories are interrupted.	↓
FP	The total number of false positives.	↓
FN	The total number of false negatives.	↓

The frame format of the training video is 240 × 135 pixels, containing 251 individual objects and 31 373 labels, while the frame format of the test video is 400 × 400 pixels, containing 259 individual objects and 64 268 labels, as shown in Fig. 2. To facilitate comparison, the test video was consistent with the test region adopted by Zhang *et al.* [35] and Jie *et al.* [36]. Since neither of the two comparison methods discloses the label file of the test video, we label the label file by ourselves.

The label files of both the training and test sets provide the center coordinates, size, and identity annotation of the vehicle object bounding box. Of particular concern is that to increase the diversity of the video environment and avoid the over-fitting phenomenon in the training process, we increased the number of labeled videos. Considering the time cost of data annotation, we did not mark all the vehicles that appeared in each training video, but the trajectory of each marked vehicle was complete. In addition, there is no overlap between training data and test data. For convenience, the naming sequence of the three test videos is shown in Fig. 12.

2) Evaluation Methodology: To verify the performance of this method, we used the metrics in MotChallenge[1] Benchmark for quantitative analysis. The specific meanings of the different metrics are shown in Table I. Multiple object tracking precision (MOTP), Recall, Precision, the total number of false positives (FP), and the total number of false negatives (FN) measure the performance of the detection part, mainly representing the accuracy of the detection box compared with the ground truth. MULTIPLE object tracking accuracy (MOTA), ID switches (IDs), IDF1 Score, ID Precision (IDP), ID Recall (IDR), mostly tracked objects (MT), mostly lost objects (ML), and fragmentation (FM) measure the performance of the tracking part. Among them, MT, ML, FM reflect the integrity of the tracking, and IDs reflects the robustness of the tracking. Specifically, MOTA comprehensively considers false positive rate, false negative rate, and mismatch rate, while IDF1 comprehensively considers IDP and IDR. As important comprehensive measures, they are the focus of this research.

B. Implementation Details

The default backbone of the method proposed in this research is the variant DLA-34 presented in [52]. The COCO detection dataset [66] is used for pre-training to initialize the model. Based on the Adam optimizer, the model performs the training of 30 epochs, and the initial learning rate is 1e-4, which attenuates to 1e-5 and 1e-6, respectively, at the 20th and 27th epoch. The batch size is set to 2. Plus, standard data enhancement techniques such as rotation, scaling, and color dithering are performed. The input image size is adjusted to 1088 × 608 before inputting to the model, and the resolution of feature images obtained in the head stage is 272 × 152. On an RTX 2080 TI GPU, training takes about 3 h.

C. Ablation Study

In this section, ablation experiments are adopted to provide rounded, detailed illustrations of the effectiveness of the method. This experiment takes Video 1 and Video 2 as examples to analyze the following three questions. First, the effectiveness of using bounding box proposals based on SFs to assist object detection. Second, how to combine the bounding box proposals with object detection results based on deep learning to achieve a better detection effect. Third, the effectiveness of motion features and time accumulative information for satellite video multiobject tracking.

1) With Bounding Box Proposals vs. Without Bounding Box Proposals: To prove the effectiveness of bounding box proposal extraction based on SFs (denoted as BbeSF) to assist object detection, a traditional version is implemented for comparison. It combines the frame difference method with the background subtraction method to implement moving object pixel points extraction. Then, the bounding box proposals are obtained in the same way as BbeSF (denoted as BbeTRA). The experimental results are shown in Table II. To be fair, only the detection part is changed in the experiment. The original step of baseline FairMOT is used for the tracking part. For the two versions with bounding box proposals, the frame interval of the input image pair is 2. In addition, the size of the bounding box proposals in BbeSF is fixed at 12 × 12.

Visibly, it can be seen that compared with the baseline FairMOT, most of the metrics of the bounding box proposals assisted methods are improved, indicating the effectiveness of using bounding box proposals. Comparing the two methods with bounding box proposals, the BbeTRA version brings relatively little improvement, and BbeSF has greater advantages from various metrics, such as MOTA (78.3% vs. 49.7% for

[1]The official MOTChallenge web page is available at https://motchallenge.net

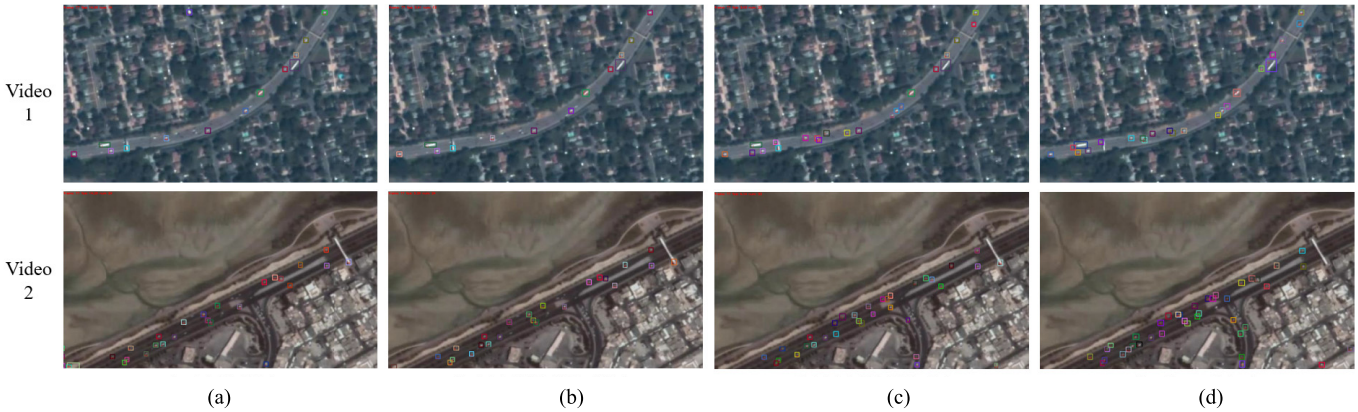


Fig. 13. Visual contrast of the improved effect of different bounding box proposal extraction methods on multiobject tracking. Bounding boxes with different colors represent different objects, and each box is labeled with its ID value. The first and second rows represent the results of Video 1 and Video 2, respectively. Columns A to C correspond to the three versions in Table II and column D represents ground truth. (a) FairMOT. (b) FairMOT-BbeTRA. (c) FairMOT-BbeSF. (d) GroundTruth.

TABLE II

COMPARISON OF THE USE AND ABSENCE OF BOUNDING BOX PROPOSALS TO ASSIST OBJECT DETECTION. NOTE THAT THE TRACKING PART OF THE LAST TWO VERSIONS IS CONSISTENT WITH THE BASELINE FAIRMOT

Method	IDF1↑	IDP↑	IDR↑	Rcll↑	Prcn↑	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓	MOTA↑	MOTP↓
Video 1													
FairMOT	66.00%	88.30%	52.70%	54.90%	92.00%	13	6	348	3269	39	330	49.50%	3.386
FairMOT-BbeTRA	65.40%	**90.60%**	51.20%	53.40%	94.40%	12	6	**230**	3377	**36**	359	49.70%	**3.358**
FairMOT-BbeSF	**84.20%**	90.10%	**79.00%**	**83.40%**	**95.10%**	**27**	**2**	309	**1202**	62	**279**	**78.30%**	3.409
Video 2													
FairMOT	62.50%	76.00%	53.10%	53.40%	76.50%	24	18	1999	5673	6	**277**	36.90%	3.917
FairMOT-BbeTRA	65.40%	**91.00%**	51.00%	51.20%	91.40%	21	21	588	5934	**5**	300	46.40%	3.855
FairMOT-BbeSF	**74.30%**	89.70%	**63.40%**	**66.20%**	**93.70%**	**28**	**11**	**538**	**4115**	78	461	**61.10%**	**3.798**

Video 1, 61.1% vs. 46.4% for Video 2) and IDF1 (84.2% vs. 65.4% for Video 1, 74.3% vs. 65.4% for Video 2). The main reason is that the traditional method is not sensitive enough to the tiny movement of the vehicles in satellite videos, and the changes that can be detected are limited. By contrast, the method based on SFs can better detect the changed part in the satellite video, namely moving object pixel points.

The visual results are shown in Fig. 13, and different colors represent the bounding boxes of different objects. It is apparent that after adding bounding box proposals, many static false alarms are filtered out and some missed objects are added, which greatly reduces the number of false positives and false negatives. The improvement of detection performance further promotes the performance of subsequent multiobject tracking.

2) Combination Strategy of Two Parts Object Bounding Boxes: This section is designed to study how to combine the bounding box proposals based on SFs with the object detection result based on deep learning to achieve the best improvement effect. Reviewing Section III, before combination, it is fundamental to extract the bounding box proposals $\text{Bboxes}_{\text{SF}}$ from the result of the moving object pixel points extraction based on SFs, namely I_{SF}. There are two solutions: directly extract the bounding box proposals from I_{SF} (denoted as Raw) or first process it with the object pixel point screening strategy designed in this research (denoted as OPS), and then extract the bounding box proposals. Furthermore, four strategies have been adopted for combination. Regarding whether to consider the difference in the accuracy of the detection bounding boxes of the two parts and whether to use the bounding box proposals to generate the mask, it is divided into the following four variants: 1) InDisFuse (regardless of the priority differences and do not use the mask); 2) DisFuse (account of the priority differences and do not use the mask); 3) InDisFuse-Mask (regardless of the priority differences and adopt the mask); and 4) DisFuse-Mask (account of the priority differences and adopt the mask). In summary, there are eight variants as shown in Table III.

As can be seen from the table, the versions with the OPS strategy can obtain better performance by comparing the first four items with the last four. The main reason is that the extraction results of moving object pixel points based on SFs are affected by satellite motion and light change, containing a lot of background information. If the bounding box proposals are extracted directly, some wrong bounding boxes may be introduced, as shown in Fig. 14. This further shows that the OPS strategy designed in this research has a practical inhibitory effect on nonobject information. Comparing OPS-Disfuse with OPS-Disfuse-Mask, OPS-Disfuse-Mask has the better performance, especially the drop in FP (from 352 to 309 in Video 1 and from 744 to 538 in Video 2), which indicates that the mask can filter out some static false alarms and improve the accuracy of detection. Comparing OPS-InDisfuse-Mask with OPS-Disfuse-Mask, the

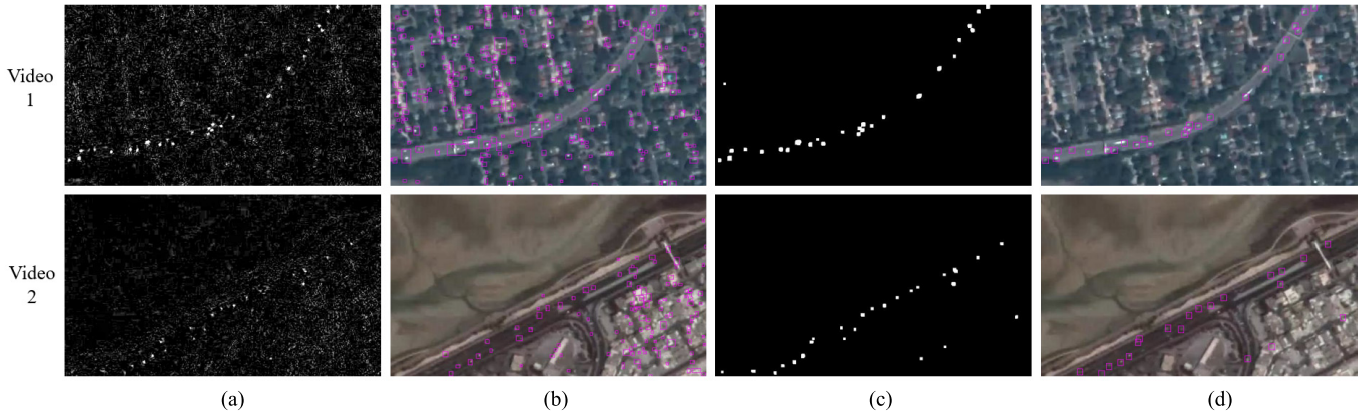


Fig. 14. Visual comparison diagram of the extraction of the bounding box proposals based on SFs. (a) Result of moving object pixel points extraction based on SFs. (b) Result obtained by directly extracting the bounding boxes from (a). To facilitate observation, the bounding boxes are displayed on the original image. (c) Result of object pixel point screening strategy for (a). (d) Result of detection boxes extraction for (c).

TABLE III

COMPARISON OF EIGHT COMBINATION STRATEGIES. RAW REPRESENTS THE EXTRACTION OF BOUNDING BOXES DIRECTLY FROM THE RESULTS OF MOVING OBJECT PIXEL POINTS EXTRACTION BASED ON SFS. OPS MEANS THAT THE BOUNDING BOX IS EXTRACTED AFTER PROCESSING WITH THE OBJECT PIXEL POINT SCREENING STRATEGY

Method	IDF1↑	IDP↑	IDR↑	Rcll↑	Prcn↑	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓	MOTA↑	MOTP↓
Video 1													
Raw-InDisFuse	65.40%	87.90%	52.10%	54.50%	92.00%	12	6	342	3296	44	344	49.20%	3.382
Raw-DisFuse	53.20%	49.10%	60.50%	63.40%	49.70%	13	4	4642	2649	69	492	-1.60%	3.420
Raw-InDisFuse-Mask	65.40%	88.20%	52.00%	54.40%	92.30%	12	6	329	3305	40	347	49.30%	3.370
Raw-DisFuse-Mask	53.20%	49.20%	60.40%	63.30%	49.80%	13	4	4625	2660	69	493	-1.50%	3.404
OPS-InDisFuse	65.90%	89.90%	52.00%	54.30%	93.90%	13	6	255	3311	**35**	341	50.30%	3.377
OPS-DisFuse	83.90%	89.50%	**79.00**%	**83.40**%	94.50%	**27**	**2**	352	**1200**	62	**279**	77.70%	3.415
OPS-InDisFuse-Mask	66.10%	**90.80**%	51.90%	54.30%	94.90%	13	6	**213**	3313	**35**	342	50.80%	**3.368**
OPS-DisFuse-Mask	**84.20**%	90.10%	**79.00**%	**83.40**%	**95.10**%	**27**	**2**	309	1202	62	**279**	**78.30**%	3.409
Video 2													
Raw-InDisFuse	66.10%	89.30%	52.50%	52.70%	89.70%	22	18	733	5751	5	**273**	46.70%	3.838
Raw-DisFuse	63.50%	76.40%	55.20%	57.30%	77.50%	22	16	2021	5192	42	400	40.40%	3.810
Raw-InDisFuse-Mask	66.30%	90.90%	52.20%	52.50%	91.30%	22	18	609	5783	5	297	47.40%	3.848
Raw-DisFuse-Mask	63.60%	77.30%	57.90%	57.10%	78.60%	22	16	1896	5224	42	422	41.10%	3.819
OPS-InDisFuse	66.20%	89.60%	52.50%	52.70%	90.00%	22	18	714	5751	5	**273**	46.80%	3.838
OPS-DisFuse	73.70%	87.60%	**63.60**%	**66.50**%	91.60%	**28**	**11**	744	**4079**	79	435	59.70%	**3.791**
OPS-InDisFuse-Mask	66.60%	**92.20**%	52.20%	52.40%	92.60%	21	18	**511**	5789	**4**	291	48.20%	3.849
OPS-DisFuse-Mask	**74.30**%	89.70%	63.40%	66.20%	**93.70**%	**28**	**11**	538	4115	78	461	**61.10**%	3.798

TABLE IV

COMPARISON OF THE FRAME INTERVAL OF THE INPUT IMAGE PAIR IN BOUNDING BOX PROPOSAL EXTRACTION BASED ON SFS. THIS RESEARCH DEFINES THE INTERVAL AS THE NUMBER OF FRAMES BETWEEN TWO IMAGE FRAMES. FOR EXAMPLE, $(t, t+4)$ IS AN IMAGE PAIR WITH AN INTERVAL OF THREE FRAMES

Interval	IDF1↑	IDP↑	IDR↑	Rcll↑	Prcn↑	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓	MOTA↑	MOTP↓
Video 1													
0	75.80%	89.10%	66.00%	70.60%	95.30%	17	3	**250**	2132	**51**	644	66.40%	**3.330**
1	82.90%	89.90%	76.90%	81.60%	**95.50**%	23	**2**	280	1330	70	342	76.80%	3.388
2	**84.20**%	**90.10**%	**79.00**%	83.40%	95.10%	27	**2**	309	1202	62	279	**78.30**%	3.409
3	81.30%	86.40%	76.80%	**84.10**%	94.60%	**28**	**2**	351	**1153**	84	**276**	78.10%	3.426
Video 2													
0	61.70%	**94.20**%	45.90%	46.40%	**95.30**%	15	20	**281**	6519	**11**	1005	44.00%	3.926
1	72.30%	92.70%	59.30%	60.00%	93.80%	24	15	479	4862	26	573	55.90%	3.833
2	74.30%	89.70%	63.40%	66.20%	93.70%	28	11	538	4115	78	461	61.10%	3.798
3	**76.70**%	89.90%	**66.90**%	**69.60**%	93.50%	**29**	**8**	586	**3697**	76	**407**	**64.20**%	**3.781**

performance of OPS-Disfuse-Mask is superior, especially the significant decrease of FN (from 3313 to 1202 in Video 1 and from 5789 to 4115 in Video 2). The main reason is that when the two parts detect the same object, the strategy taking the accuracy priority into account retains the more accurate object detection result based on deep learning. Moreover, it also improves the confidence score, which allows the correctly detected object bounding box with low confidence to be retained in the confidence score screening step. In terms of comprehensive comparison, the OPF-Disfuse-Mask version, this is, the bounding boxes proposals-guided NMS module, achieves the optimal results.

TABLE V

COMPARISON OF THE FIXED SIZE OF THE BOUNDING BOX PROPOSALS OBTAINED BY SFS. FOR EXAMPLE, 8 MEANS THAT THE BOUNDING BOX SIZE IS 8×8

Fix-size	IDF1↑	IDP↑	IDR↑	Rcll↑	Prcn↑	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓	MOTA↑	MOTP↓
Video 1													
8	82.30%	87.60%	77.60%	**83.90%**	94.90%	**28**	**2**	330	**1164**	76	**256**	**78.30%**	3.408
10	82.30%	87.80%	77.40%	83.70%	94.90%	**28**	**2**	323	1178	83	265	78.10%	**3.405**
12	**84.20%**	**90.10%**	**79.00%**	83.40%	95.10%	27	**2**	309	1202	**62**	279	**78.30%**	3.409
14	83.70%	90.00%	78.30%	82.80%	**95.20%**	25	**2**	**302**	1244	71	304	77.70%	3.413
Video 2													
8	**77.40%**	90.50%	**67.60%**	**69.80%**	**93.50%**	**29**	**8**	587	**3673**	74	400	**64.40%**	**3.779**
10	76.80%	89.80%	67.00%	**69.80%**	**93.50%**	**29**	**8**	590	3678	76	395	64.30%	3.782
12	76.70%	89.90%	66.90%	69.60%	**93.50%**	**29**	**8**	586	3697	76	407	64.20%	3.781
14	77.00%	**90.90%**	66.80%	68.70%	**93.50%**	28	10	**584**	3813	**54**	**391**	63.40%	3.796

TABLE VI

COMPARISON OF DIFFERENT IMPROVEMENT STRATEGIES BASED ON MOTION INFORMATION AND TIME ACCUMULATIVE INFORMATION. FAIRMOT* REPRESENTS THE OPTIMAL VERSION OF OPS-DISFUSE-MASK IN TABLE III, NOT BASELINE FAIRMOT

Method	IDF1↑	IDP↑	IDR↑	Rcll↑	Prcn↑	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓	MOTA↑	MOTP↓
Video 1													
FairMOT*	84.20%	90.10%	79.00%	83.40%	95.10%	27	2	309	1202	62	279	78.30%	3.409
FairMOT*-ReFA	85.20%	92.30%	79.00%	83.20%	97.30%	27	2	168	1214	58	278	80.10%	3.369
FairMOT*-ReFA-ReRB	85.20%	**92.60%**	78.90%	83.20%	**97.50%**	27	2	**154**	1220	58	282	80.20%	**3.364**
FairMOT*-ReFA-KF	87.20%	89.60%	**85.00%**	**89.70%**	94.60%	**31**	**1**	370	**748**	**24**	81	84.20%	3.497
FairMOT*-ReFA-ReRB-KF	**87.60%**	90.80%	84.60%	89.30%	95.90%	**31**	2	279	776	25	**79**	**85.10%**	3.450
Video 2													
FairMOT*	77.40%	90.50%	67.60%	69.80%	93.50%	29	8	587	3673	74	400	64.40%	3.779
FairMOT*-ReFA	79.10%	**95.30%**	67.60%	69.80%	**98.40%**	29	8	138	3674	72	399	68.10%	**3.776**
FairMOT*-ReFA-ReRB	79.00%	**95.30%**	67.50%	69.80%	**98.40%**	29	8	**135**	3679	72	399	68.10%	**3.776**
FairMOT*-ReFA-KF	83.70%	91.20%	**77.40%**	**79.70%**	94.00%	**36**	**5**	624	**2472**	**17**	**47**	74.40%	3.946
FairMOT*-ReFA-ReRB-KF	**84.00%**	92.70%	76.80%	79.60%	96.00%	**36**	**5**	399	2485	26	49	**76.10%**	3.976

What is more, the number of optimal interval frames between two input frames for bounding box proposal extraction based on SFs has been further analyzed, as shown in Table IV. For Video 1, the performance is optimal when the interval is 2, and for Video 2, the optimal value is 3. We also analyzed the optimal fixed size of bounding box proposals with OPS strategy, as shown in Table V. It can be seen that for Video 1, the performance is optimal when the fixed size is 12×12, and for Video 2, the optimal value is 8×8. For different videos, the optimal parameters will be different, which we guess may be affected by the average size and speed of objects.

3) Effectiveness of Motion Features and Time Accumulative Information: In this research, motion features and time accumulative information are used to process the preliminary tracking results obtained from data association, including three parts: 1) trajectory smoothing based on the Kalman filter (denoted as KF); 2) removal of static false alarms (denoted as ReFA); and 3) removal of repeated object bounding boxes (denoted as ReRB). To verify the effectiveness of each component, ablation experiments are conducted for each component, as shown in Table VI. To be fair, for all versions, the optimal version of OPS-Disfuse-Mask in Table III is used for the detection portion, while the original step of FairMOT is used for the tracking portion in the first base version.

Through observation, it can be seen that after adding each improvement strategy, the MOTA measurement of the two videos will improve or remain unchanged, among which ReFA and KF bring the greater improvement effect. One interesting finding is that when comparing FairMOT*-ReFA and FairMOT*-ReFA-ReRB, the improvement of MOTA is very small and even caused a slight deterioration of other metrics. However, when comparing FairMOT*-ReFA-KF with FairMOT*-ReFA-ReRB-KF, MOTA improves significantly (85.1% vs. 84.2% for Video 1 and 76.1% vs. 74.4% for Video 2). This finding suggests that the ReRB part has a great improvement effect on the part of KF, and the combination of the two parts will yield better performance than either alone. A possible explanation for this might be that when the Kalman filter is used to fill the trajectories, some wrong detection boxes may be introduced due to inaccurate predictions, which can be partly removed by the ReRB part.

Besides, the optimal trajectory smoothing length is further compared and analyzed, as shown in Table VII. Note that the base version without the KF module is the version FairMOT*-ReFA-ReRB in Table VI. Clearly, Video 1 and Video 2 reach the optimal MOTA value when the frame length is 3 and 11, respectively. Compared with the basic version, FN, IDs, and FM gradually decrease with the increase of smoothing length. This proves that using the prediction result based on the Kalman filter to fill the interrupt trajectory can reduce many missed objects and effectively prevent the occurrence of ID switches and tracking loss after trajectory interruption. In addition, with the increase of the predicted frame length, the deviation degree between the predicted position and the real trajectory gradually increases, the prediction accuracy decreases, and more false alarms will be introduced. This shows that it is necessary to choose the appropriate smoothing length according to the situation of the video.

TABLE VII

COMPARISON OF THE DIFFERENT LENGTHS OF TRAJECTORY FILLING WITH THE KALMAN FILTER. THE "WITHOUT" VERSION IS FAIRMOT*-REFA-RERB IN TABLE VI WITHOUT USING THE KALMAN FILTER

Frames	IDF1↑	IDP↑	IDR↑	Rcll↑	Prcn↑	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓	MOTA↑	MOTP↓
Video 1													
without	85.20%	**92.60**%	78.90%	83.20%	**97.50**%	27	2	**154**	1220	58	282	80.20%	**3.364**
1	86.90%	92.00%	82.30%	86.70%	97.00%	29	2	195	963	40	140	83.50%	3.390
3	**87.60**%	90.80%	84.60%	89.30%	95.90%	31	2	279	776	25	79	**85.10**%	3.450
5	87.50%	89.50%	85.50%	90.30%	94.60%	31	**1**	373	701	23	65	84.90%	3.509
7	87.20%	88.40%	86.00%	90.90%	93.40%	**34**	**1**	462	659	18	62	84.30%	3.546
9	86.90%	87.30%	86.40%	91.40%	92.30%	**34**	**1**	549	624	19	60	83.50%	3.568
11	86.70%	86.50%	**86.80**%	**91.70**%	91.50%	**34**	**1**	621	**598**	**17**	**56**	82.90%	3.597
Video 2													
without	79.00%	**95.30**%	67.50%	69.80%	**98.40**%	29	8	**135**	3679	72	399	68.10%	**3.776**
1	80.90%	94.80%	70.50%	73.00%	98.20%	29	7	162	3281	52	204	71.30%	3.810
3	82.40%	94.30%	73.10%	75.80%	97.80%	32	**5**	208	2945	39	115	73.80%	3.862
5	83.20%	93.90%	74.60%	77.40%	97.40%	34	**5**	251	2753	35	77	75.00%	3.900
7	83.60%	93.50%	75.50%	78.30%	96.90%	35	**5**	301	2639	31	62	75.60%	3.924
9	83.80%	93.10%	76.20%	79.00%	96.50%	35	**5**	352	2556	30	54	75.90%	3.939
11	**84.00**%	92.70%	**76.80**%	**79.60**%	96.00%	**36**	**5**	399	**2485**	**26**	**49**	**76.10**%	3.976

TABLE VIII

COMPARISON OF THE STATE-OF-THE-ART MOT METHODS. IT IS NOTEWORTHY THAT YOLOV4-DEEPSORT IS THE METHOD FOR REPLACING THE OBJECT DETECTION PART OF THE ORIGINAL DEEPSORT [12] METHOD WITH THE YOLOV4 [46] OBJECT DETECTION METHOD. ALL THE METHODS ARE TRAINED BASED ON THE SATEMVT DATASET

Method	IDF1↑	IDP↑	IDR↑	Rcll↑	Prcn↑	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓	MOTA↑	MOTP↓
Video 1													
FairMOT	66.00%	88.30%	52.70%	54.90%	92.00%	13	6	348	3269	39	330	49.50%	**3.386**
YOLOv4 [46]-DeepSORT [12]	12.40%	**91.40**%	6.60%	6.60%	91.40%	2	37	**45**	6763	**0**	**20**	6.00%	13.486
SFMFMOT	**87.60**%	90.80%	**84.60**%	**89.30**%	**95.90**%	**31**	**2**	279	**776**	25	79	**85.10**%	3.450
Video 2													
FairMOT	62.50%	76.00%	53.10%	53.40%	76.50%	24	18	1999	5673	**6**	277	36.90%	**3.917**
YOLOv4 [46]-DeepSORT [12]	5.40%	28.30%	3.00%	3.80%	36.50%	0	50	809	11702	9	93	-2.90%	17.054
SFMFMOT	**84.00**%	**92.70**%	**76.80**%	**79.60**%	**96.00**%	**36**	**5**	**399**	**2485**	26	**49**	**76.10**%	3.976

D. Comparisons With the State-of-the-Arts

The purpose of this study is to verify the superiority of the method proposed in this research compared with other advanced multiobject tracking methods in satellite videos. We combined two representative methods, the object detection method YOLOv4 [46] and the multiobject tracking method DeepSORT [12], as a SOTA multiobject tracking method YOLOv4-DeepSORT for comparison.

1) Test Results of the Jilin No. 1 Test Video: The visual results of sequence images and completed trajectories are shown in Figs. 15 and 16, respectively. First, by comparing the FairMOT with Yolov4-DeepSORT, it is clear that more vehicles can be detected and tracked by FairMOT. This result may be explained by the fact that through multilayer feature fusion, FairMOT greatly alleviates the feature unfairness and feature conflict between detection task and Re-ID task in the two-step methods such as YOLOv4-DeepSORT. This is also an important reason why FairMOT is chosen as the baseline in this research. Then, in comparison, the SFMFMOT method proposed in this research detects the most objects and can achieve the most continuous tracking of multiple objects.

The quantitative evaluation results are shown in Table VIII. It can be seen that the main evaluation measures of the proposed method, MOTA and IDF1, surpass other advanced trackers by a large margin, which demonstrates the superiority for vehicle object tracking in satellite video. Additionally, the proposed tracker has also achieved the best performance in other metrics such as IDR, Recall, Precision, and FN. Among these indicators, the lowest FN value and the highest IDR and Recall values prove that more objects could be detected in the object detection part assisted by bounding box proposal extraction based on SFs. And the highest Precision indicates that the highest ratio of correctly matched detections to total detection results could be guaranteed at the same time. The optimal MT and ML values indicate the superior robustness of the proposed tracker for continuous tracking of the object. Furthermore, given the relative concepts of indicators such as FP, IDs, FM, and so on, it is unreasonable to compare only the absolute value. For example, when no trajectory is detected and tracked, FP, IDs, and FM get the seemingly optimal result of 0, which is clearly unreasonable. Hence, considering that the number of objects detected by the proposed method is far more than that detected by YOLOv4-DeepSORT, the slight increase in the value of FP, IDs, FM, and so on can still prove the superiority of the proposed method.

2) Test Results of the SkySat-1 Test Video: Considering that Zhang *et al.* [35] and Jie *et al.* [36] both provided qualitative and quantitative evaluation results for the same region of Video 3, they were added as comparison methods in this section.

In order to ensure the fairness of experimental comparison, the following two points should be explained first:

a) Only the first 700 frames were used for comparison: According to Zhang *et al.* [35] and Jie *et al.* [36], only

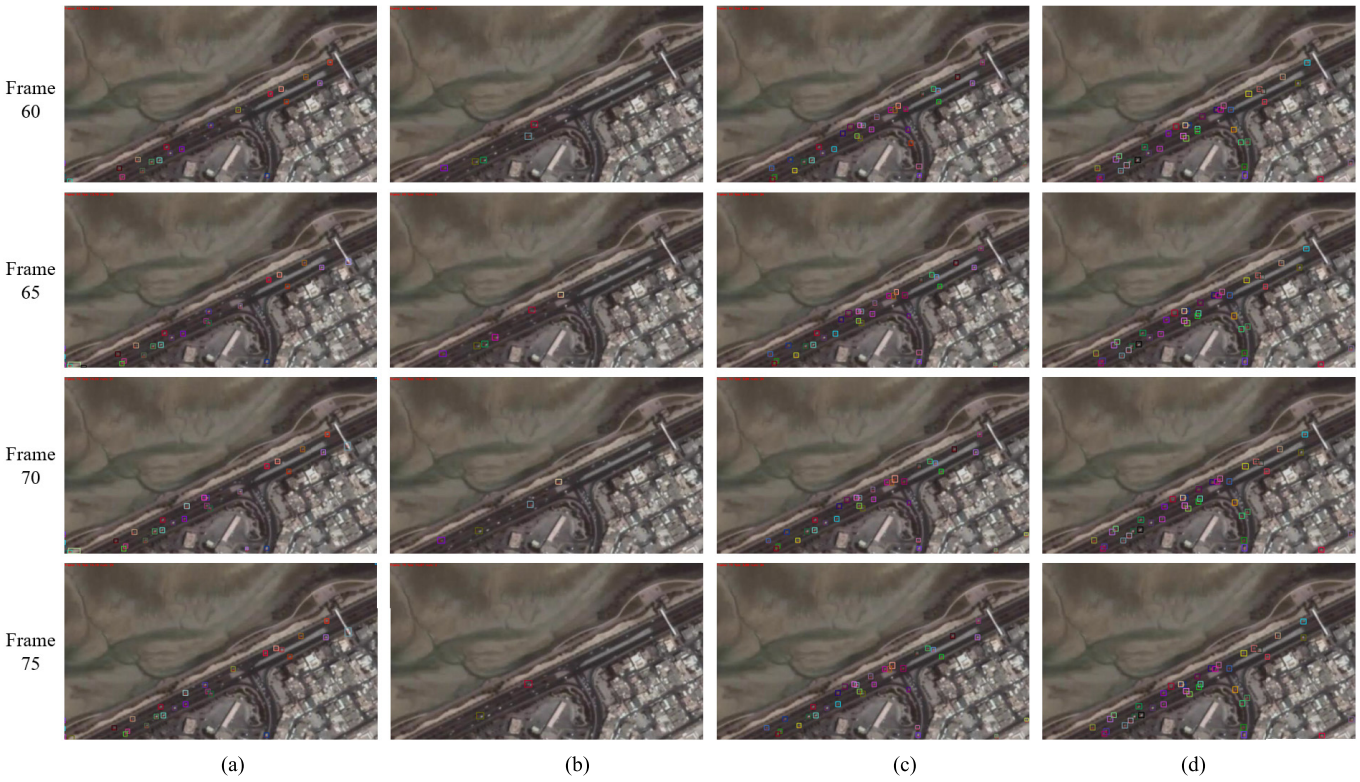


Fig. 15. Visual comparison diagram of the tracking results of different multiobject tracking methods, taking Video 2 as an example. From top to bottom, each column displays the tracking results in chronological order at intervals of five frames. (a) FairMOT. (b) YOLOv4-DeepSORT. (c) SFMFMOT. (d) GroundTruth.

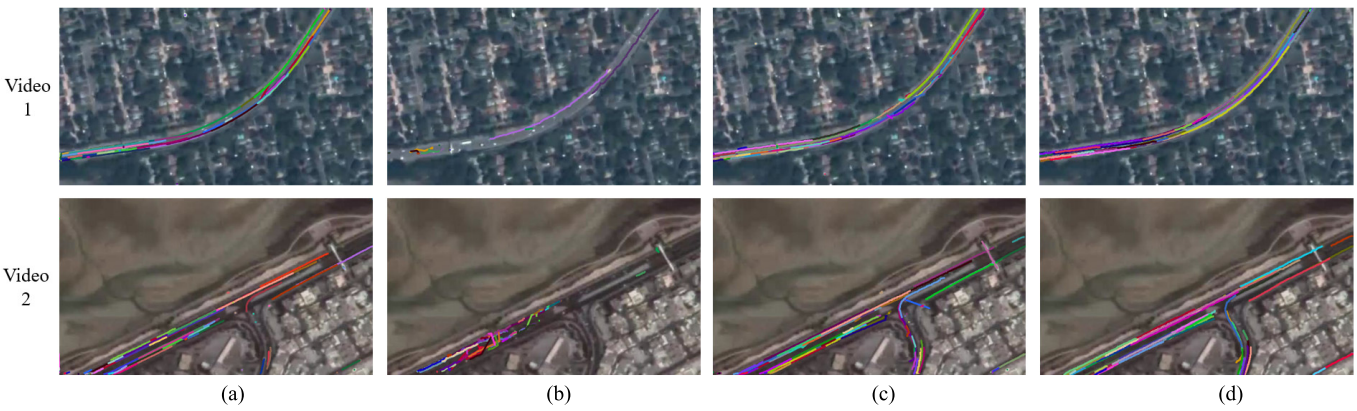


Fig. 16. Trajectories of moving vehicles on the test video of Jilin No. 1 satellite. (a) FairMOT. (b) YOLOv4-DeepSORT. (c) SFMFMOT. (d) GroundTruth.

TABLE IX

TRACKING PERFORMANCE OF DIFFERENT METHODS (FIRST 700 FRAMES). THE QUALITATIVE AND QUANTITATIVE RESULTS OF ZHANG *et al.* [35] AND JIE *et al.* [36] METHODS WERE OBTAINED DIRECTLY FROM THE ORIGINAL ARTICLE BECAUSE THE SOURCE CODE WAS NOT PROVIDED

Method	IDF1↑	IDP↑	IDR↑	Rcll↑	Prcn↑	GT	MT↑	FP↓	FN↓	IDs↓	FM↓	MOTA↑	MOTP↓
DeepSORT*	40.10%	54.80%	31.60%	38.20%	66.20%	214	33	5580	17731	262	644	17.80%	4.99
FairMOT*	59.70%	55.80%	64.20%	67.40%	58.60%	214	102	13672	9355	130	965	19.20%	**2.02**
SFMFMOT*	**88.00**%	**89.30**%	**86.70**%	**89.10**%	**91.70**%	214	**177**	2311	3128	100	498	**80.70**%	2.13
Zhang et al. [35]	62.60%	72.70%	55.0%	-	-	171	81	3101	8638	**33**	**80**	48.10%	-
Feng et al. [36]	84.10%	84.40%	83.80%	86.70%	87.30%	193	157	**665**	**706**	37	-	73.40%	-

the first 700 frames were used in the test stage. To be fair, two groups of qualitative and quantitative evaluation results are provided. One group is used to compare with the methods of Zhang *et al.* [35] and Jie *et al.* [36], and only the first 700 frames of Video 3 are tested. The trajectory diagram and the qualitative evaluation results are shown in Fig. 17 and Table IX. The other group tested the complete tracking results of Video 3, and the trajectory diagram

TABLE X
TRACKING PERFORMANCE OF DIFFERENT METHODS (1800 FRAMES)

Method	IDF1↑	IDP↑	IDR↑	Rcll↑	Prcn↑	GT	MT↑	FP↓	FN↓	IDs↓	FM↓	MOTA↑	MOTP↓
DeepSORT	31.20%	43.00%	24.50%	41.20%	72.40%	259	28	10082	37793	632	1652	24.50%	4.97
FairMOT	43.20%	39.90%	47.10%	71.10%	60.30%	259	118	30154	18560	504	2163	23.40%	**1.933**
SFMFMOT	59.80%	60.70%	58.90%	**89.00%**	**91.80%**	259	215	**5097**	**7038**	455	1248	80.40%	2.039
DeepSORT*	45.20%	62.30%	35.50%	41.20%	72.40%	432	67	10082	37793	498	1518	24.70%	4.97
FairMOT*	62.40%	57.70%	68.10%	71.10%	60.30%	432	219	30154	18560	346	2011	23.70%	**1.933**
SFMFMOT*	**87.00%**	**88.30%**	**85.70%**	**89.00%**	**91.80%**	432	**350**	**5097**	**7038**	**286**	**1146**	**80.70%**	2.039

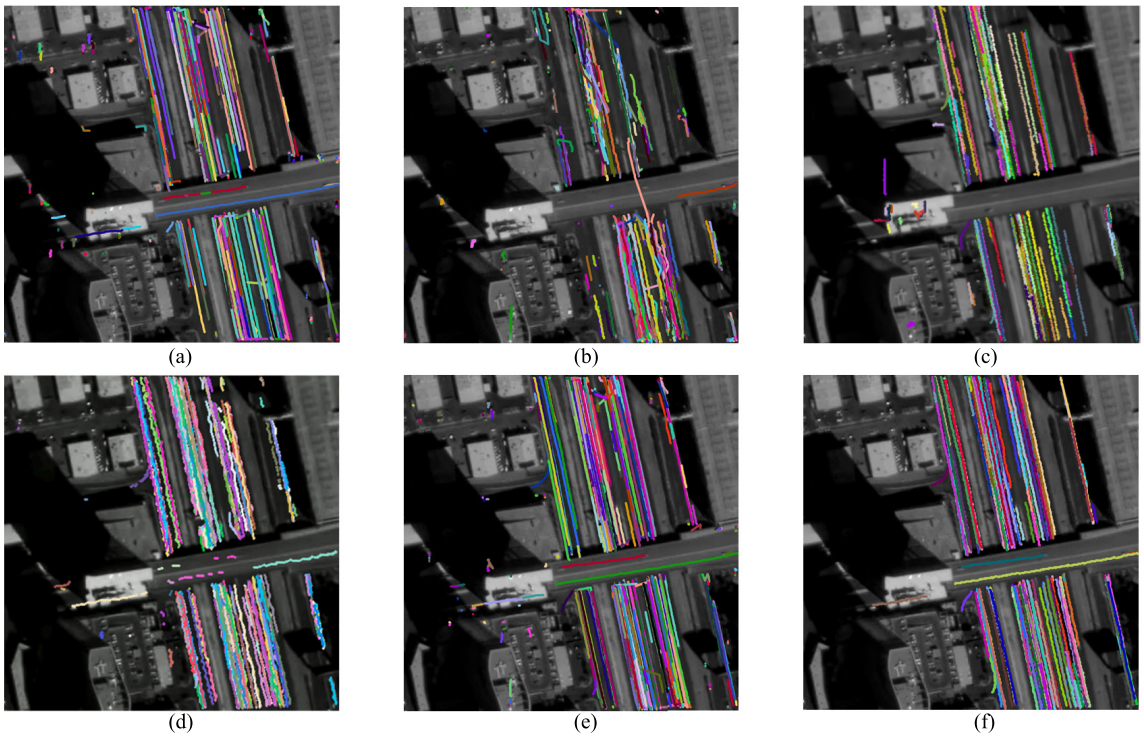


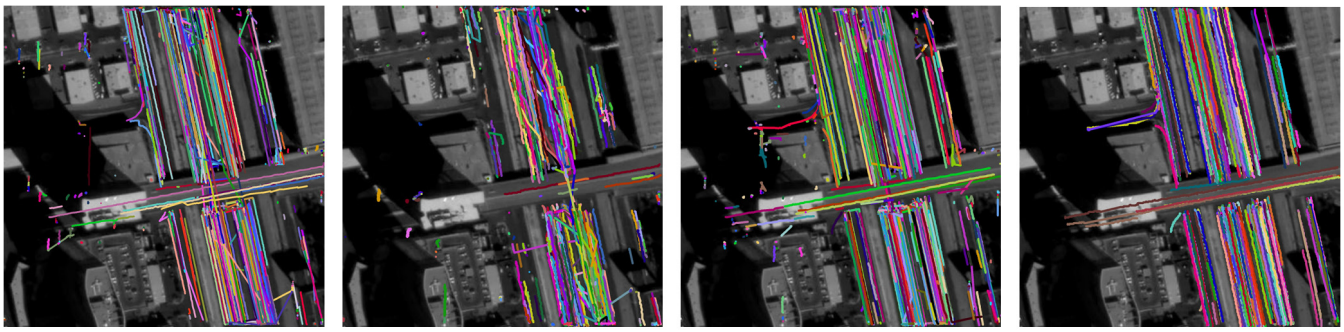
Fig. 17. Trajectories of moving vehicles in video B (first 700 frames). (a) FairMOT. (b) YOLOv4-DeepSORT. (c) Zhang *et al.* (d) Feng *et al.* (e) SFMFMOT. (f) GroundTruth.

and qualitative evaluation results are shown in Fig. 18 and Table X.

b) Analysis of the ID labeling strategy: As shown in Tables IX and X, the GT column represents the total number of real moving targets. According to Table IX, the first 700 frames of Video 3 labeled by ourselves contained 137 moving targets, while Zhang *et al.* [35] and Jie *et al.* [36] reached 171 and 193, respectively. Considering that the principle of labeling all moving targets as far as possible was adopted in this research, the number of targets labeled by the two methods should not be more than 137 for the same region. In addition, as shown in Fig. 17, the trajectory graphs of the two methods were inconsistent with the IDF1 index. After occlusion by overpasses, the colors of trajectories of most targets changed, indicating a large number of ID switches, but a high IDF1 value was obtained. Therefore, we speculated that in methods Zhang *et al.* [35] and Jie *et al.* [36], the principle of assigning different IDs to the same target before and after the overpass could be adopted. To make a fair comparison, as shown in Tables IX and X, we added a group of experiments based on the different-ID labeling strategy and marked with * to distinguish from the results using the same-ID labeling strategy. It is worth noting that the DeepSORT in the table is the method after replacing the detection part with the YOLOv4 method, which is abbreviated as DeepSORT for brevity.

As shown in Table IX, FairMOT and DeepSORT methods designed based on optical video are compared with Zhang *et al.* [35], Jie *et al.* [36] and SFMFMOT designed based on satellite videos. It can be seen that the method based on satellite videos has a better performance on the SkySat-1 test video, which reflects the significance of designing a multiobject tracking method specifically considering the characteristics of objects in satellite videos.

From the comparison of quantitative indexes in various aspects as shown in Table IX, SFMFMOT achieved the best results in MOTA, IDF1, Recall, Precision, and other important indexes. From the perspective of visual results, as shown in Fig. 17, curves of the same color represent the trajectory of the same target. Compared with Zhang *et al.* [35] and the ground truth, the trajectory integrity rate of SFMFMOT is higher in terms of density and continuity, and most of the



(a) (b) (c) (d)

Fig. 18. Trajectories of moving vehicles in video B (1800 frames). (a) FairMOT. (b) YOLOv4-DeepSORT. (c) SFMFMOT. (d) GroundTruth.

moving targets have been detected and tracked. Compared with Jie *et al.* [36], the trajectory results obtained by the proposed method are more stable and continuous. Taking the vehicles on the overpass as an example, Zhang *et al.* [35] and YoloV4-DeepSort almost failed to detect the moving vehicles on the overpass. Although FairMOT, Jie *et al.* [36] kept tracking the target, the interruption problem existed. Only SFMFMOT can keep the continuous tracking of vehicles.

For the long-time series tracking up to 1 min, as shown in Table X and Fig. 18, SFMFMOT can still obtain the best quantitative results and the most stable and continuous visual tracking results with MOTA as high as 80.7%. It shows that SFMFMOT has high robustness for long-time tracking.

In conclusion, in the SkySat-1 satellite test video, the proposed method achieves the most superior performance compared with other methods, which may mainly benefit from the focus on exploring the characteristics of moving targets in satellite videos.

Discuss: Although the proposed method in this research has achieved the optimal performance on the test video of the SkySat-1 satellite and Jilin No. 1 satellite, there are still some missed objects and a high ID switch rate under long-term occlusion. As shown in Table X, for the two different versions of the proposed method, the quantitative indicators of SFMFMOT that are related to the consistency of ID, such as IDF1, IDP, and IDR, have a significant decrease compared with SFMFMOT*. How to further reduce the miss rate of weak and small targets in satellite videos and improve the algorithm's robustness to the long-term occlusion problem is an important direction to be solved in the future.

V. Conclusion

Multiobject tracking based on satellite videos is a very challenging task. Compared with the ground video multiobject tracking task, the satellite video has the characteristics of small objects, low contrast between objects and background, complex and moving background, which brings more obstacles to the object extraction and continuous tracking.

In this research, a multiobject tracking algorithm, SFMFMOT, is proposed for the multivehicle object tracking task in satellite videos. Starting from object detection and data association modules, the accuracy and robustness from detection to tracking are comprehensively improved. By employing the NMS module guided by bounding box proposals based on SFs to object detection, the accuracy of the target extraction of interest is greatly improved. Through the deep mining of object motion features, strategies for trajectory smoothing, false alarm, and duplicate object detection removal are designed, which further enhance the integrity and robustness of the tracking part. Moreover, a multivehicle object tracking dataset of satellite videos, namely SateMVT, is annotated to facilitate research and quantitative analysis. Experimental results show that the proposed multiobject tracking method based on satellite videos achieves a new benchmark performance on the new dataset (SateMVT), with better performance than other methods.

Despite these promising results, some questions remain. Limited by the characteristics of satellite videos, there are still some objects that have not been detected and tracked. Further improving the integrity of object detection and tracking in satellite videos and reducing the number of false negatives is an important topic for future research.

Appendix A
Supplementary Details of the Object Detection Section

A. Evaluation of the Object Detection Performance

Object detection is an essential part of the multiobject tracking method based on the DBT mode, greatly influencing tracking performance. Therefore, this section conducts a specific evaluation of object detection performance based on the precision–recall curve.

As shown in Fig. 19, comparison diagrams of precision–recall curves of DeepSORT, FairMOT, and SFMFMOT are shown. It is worth noting that the DeepSORT here is the method after replacing the detection part with YOLOv4, which is abbreviated as DeepSORT for brevity.

The precision–recall curves reflect the relationship between precision and recall at different thresholds. Detector with a precision–recall curve closer to the upper right has a better performance. Correspondingly, the area under the curve (AUC) enclosed by the curve and the coordinate axes is larger. By observing the curve of DeepSORT, it can be seen that the AUC of this curve is significantly lower than that of the other

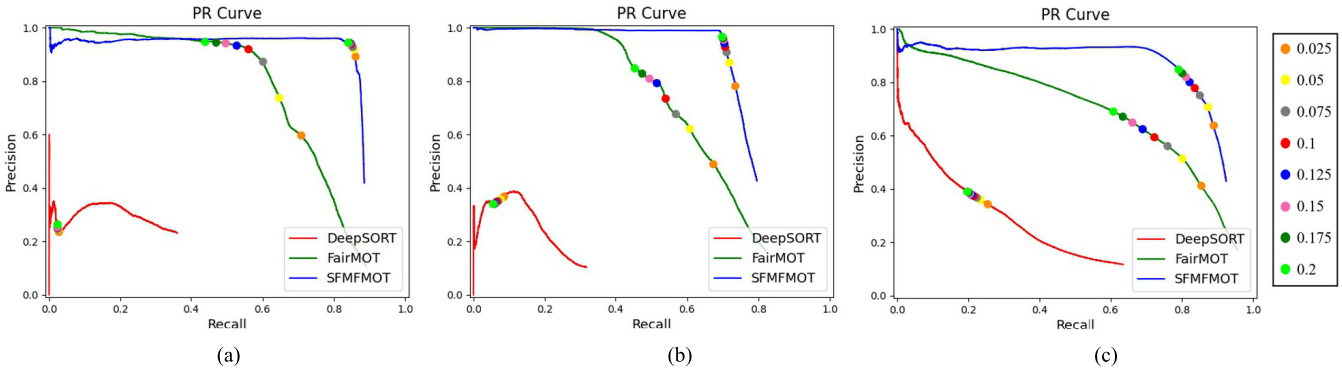


Fig. 19. Precision–recall curves for different multiobject tracking methods on three test videos. Eight confidence thresholds were selected with intervals of 0.025. The precision–recall values corresponding to different confidence thresholds are displayed on the precision–recall curve with different colored dots. (a) Video 1. (b) Video 2. (c) Video 3.

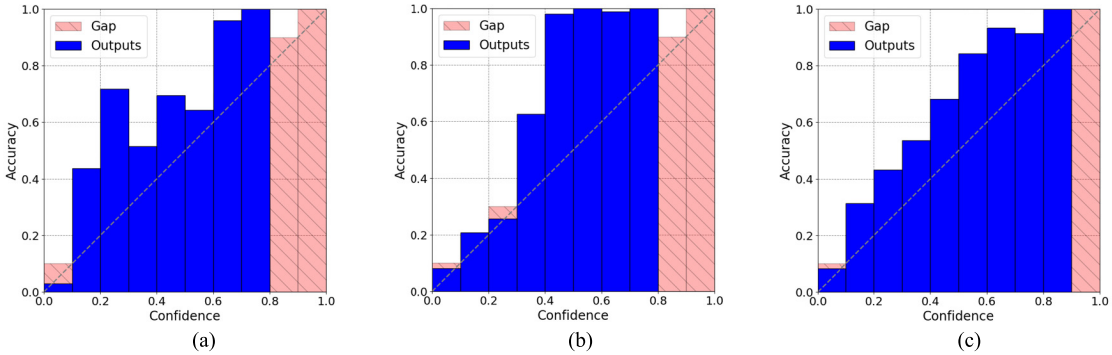


Fig. 20. Reliability diagrams for different test videos. All prediction bounding boxes were divided into different intervals according to the confidence threshold with an interval of 0.1, and the accuracy of prediction boxes in each interval was calculated to draw the bar chart. The red bins are distributed along the diagonal and only for reference. Refer to [67] for details. (a) Video 1. (b) Video 2. (c) Video 3.

methods, indicating that the DeepSORT method has the worst detection performance and a large number of false alarms. In addition, the maximum Recall value achieved by this curve is low, indicating that there are many missed objects. Both the baseline FairMOT method and the proposed method maintain high precision in the high confidence threshold part. As the confidence threshold decreases, the curve of FairMOT first presents a downward trend. For Video 1 and Video 2, when Recall reaches about 0.4, the curve begins to show a downward trend. For Video 3, the curve even presents a downward trend from the beginning. In contrast, with the decrease of the threshold value, SFMFMOT proposed in this research can still ensure high precision while gradually improving Recall. It is not until the Recall reaches 0.7–0.8 that the curve begins to show a significant downward trend. By comprehensive comparison, the SFMFMOT method achieves the maximum AUC value, indicating that the SF-based object detection enhancement strategy proposed in this research has great advantages in satellite videos and achieves the best object detection performance.

B. Description of Confidence Threshold Selection

For all prediction boxes output by the target detection network, only those with confidence values greater than the specified threshold are regarded as vehicle targets and retained, otherwise they will be removed as nonvehicle targets. In order to compare performance at different confidence thresholds, as shown in Fig. 19, 0.025 was used as an interval to select eight confidence thresholds. The precision–recall values corresponding to different confidence thresholds are displayed on the precision–recall curve with different color dots.

Through observation, it is found that for the proposed SFMFMOT method, with the decrease of confidence threshold, the precision changes very gently at the beginning. When the confidence threshold reaches about 0.2, the precision begins to plummet. This phenomenon shows that most of the detection boxes whose confidence threshold is greater than 0.2 are correct detection, otherwise most of them are false detection. In addition, it can be seen that when the confidence is between 0.075 and 0.2, the corresponding dot positions are very close, which indicates that the selection of threshold in this range will not have a great impact on the detection precision.

Furthermore, referring to method [67], all prediction boxes are divided according to different confidence score intervals, and the accuracy values are calculated, respectively, to generate the reliability graph as shown in Fig. 20. When the confidence value is between 0 and 0.1, the accuracy is very low, but after 0.1, the accuracy is significantly improved. In order to balance precision and recall and maintain high accuracy, 0.1 was selected as the confidence threshold.

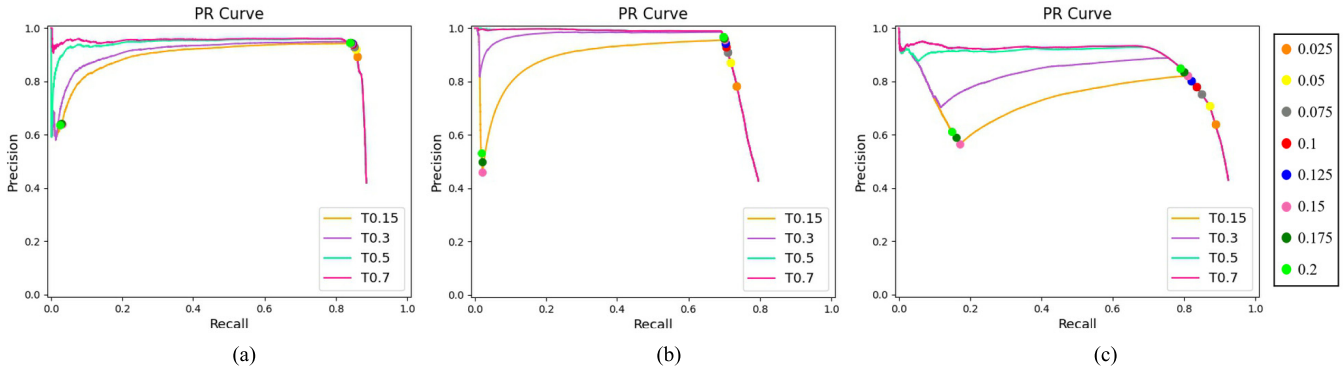


Fig. 21. Precision–recall curves for different reset confidences on the three test videos. (a) Video 1. (b) Video 2. (c) Video 3.

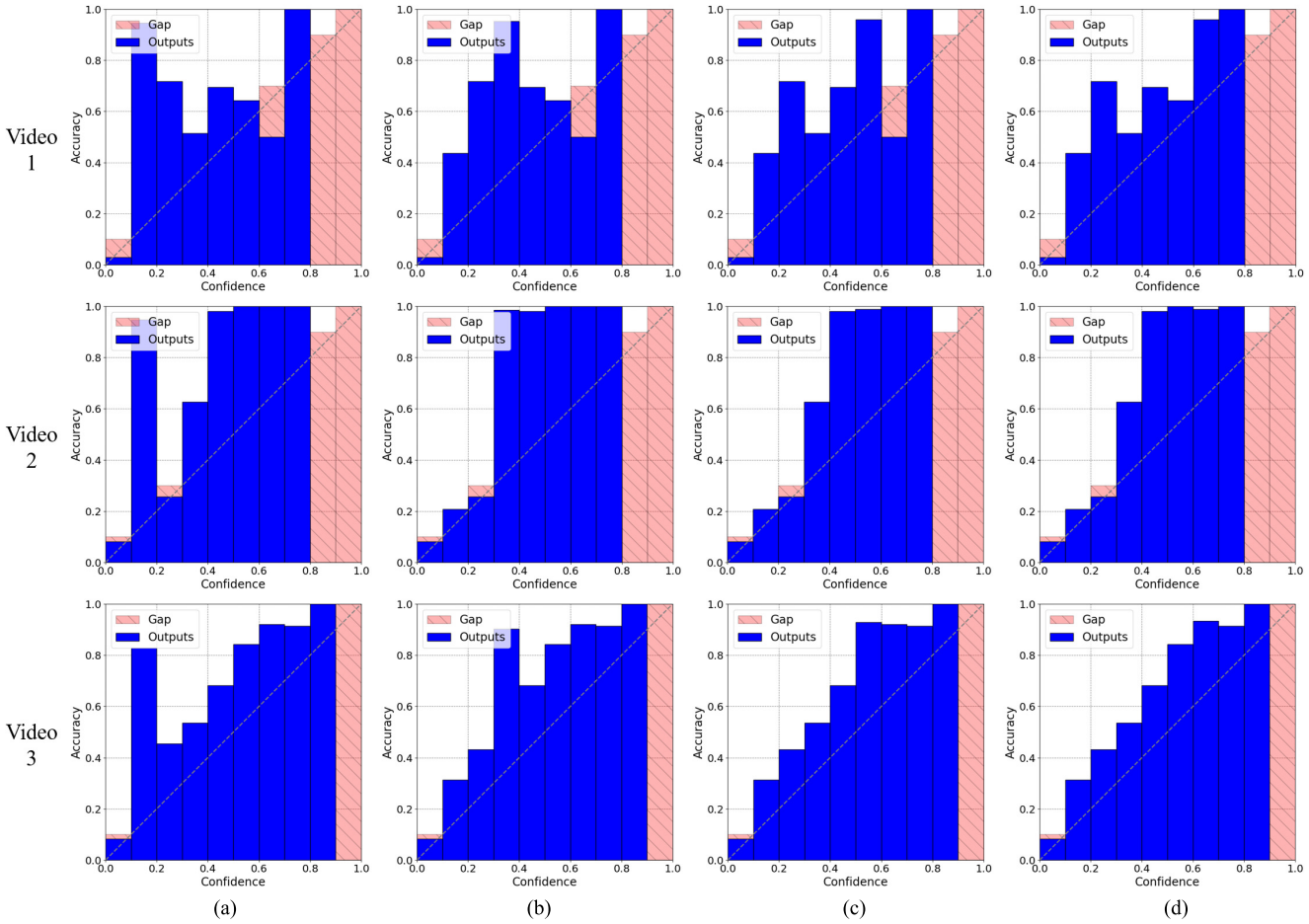


Fig. 22. Reliability diagrams for different reset confidences on the three test videos. (a) 0.15. (b) 0.3. (c) 0.5. (d) 0.7.

C. Description of Reset Confidence Parameter Selection

In the Bounding Box Proposals-Guided NMS Module proposed in this research, for the object detected in both the bounding box proposals $\text{Bboxes}_{\text{SF}}$ based on SF and the bounding boxes $\text{Bboxes}_{\text{DL}}$ detected in the deep learning-based object detection, the confidence score of the object is increased to 0.5. This section makes a detailed analysis of the selection of reset confidence score.

In order to compare the influence of reset confidence value S_{new} on the detection performance, several values were selected for comparison, which were 0.15, 0.3, 0.5, and 0.7, respectively. As shown in Fig. 21, the precision–recall curves of different parameters were plotted on the three test videos, and curves of different colors correspond to different reset confidence. It can be observed the following.

1) Some Curves Suddenly Drop and Then Rise, and This Phenomenon Gradually Alleviates With the Increase of Reset Confidence Value: As shown in the reliability diagram in Fig. 22, the accuracy of some high confidence intervals is lower than 1, indicating the existence of error detection with

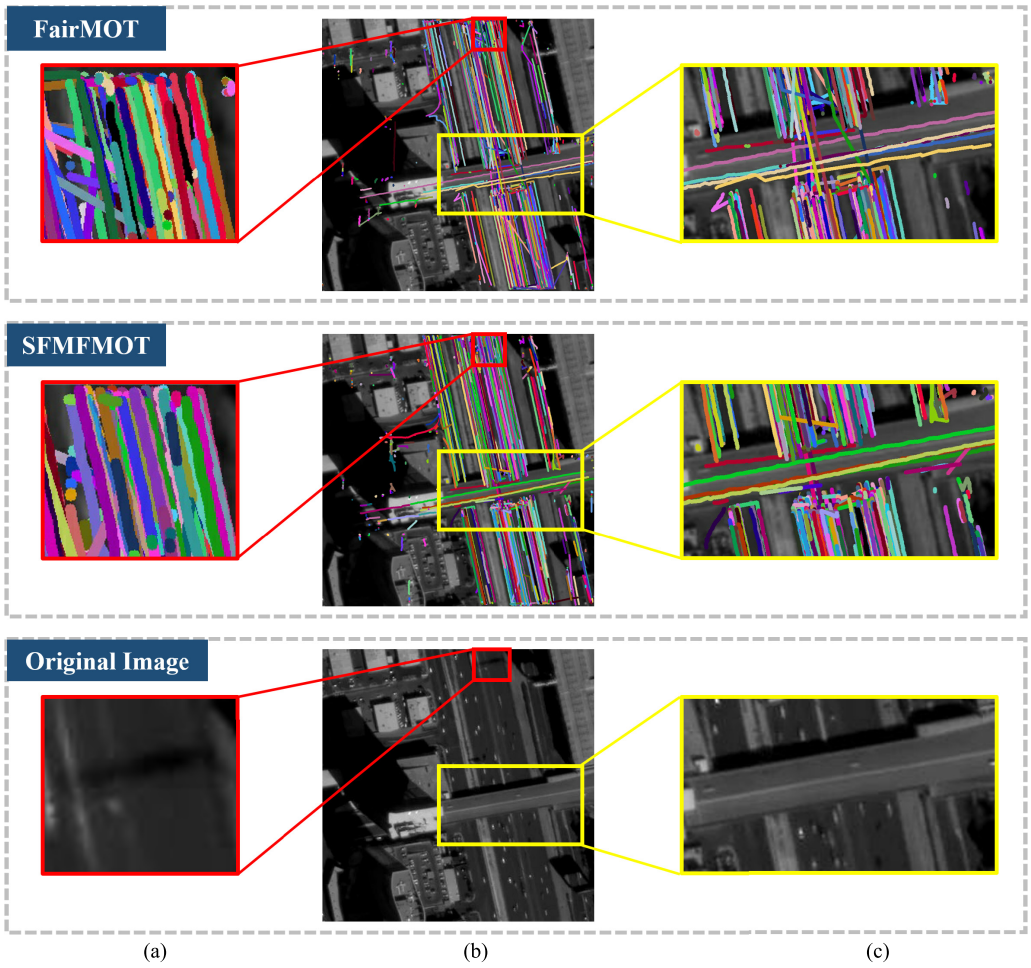


Fig. 23. Example diagram of the occlusion. The red box area on the left is a weak occlusion area, with a narrow building causing short-term occlusion to vehicles on the road. The yellow box area on the right is the strong occlusion area, and there is an overpass with a width of about six times the length of the vehicle, which causes long-term occlusion to the passing vehicles. (a) Weak occlusions. (b) Trajectory diagram. (c) Strong occlusions.

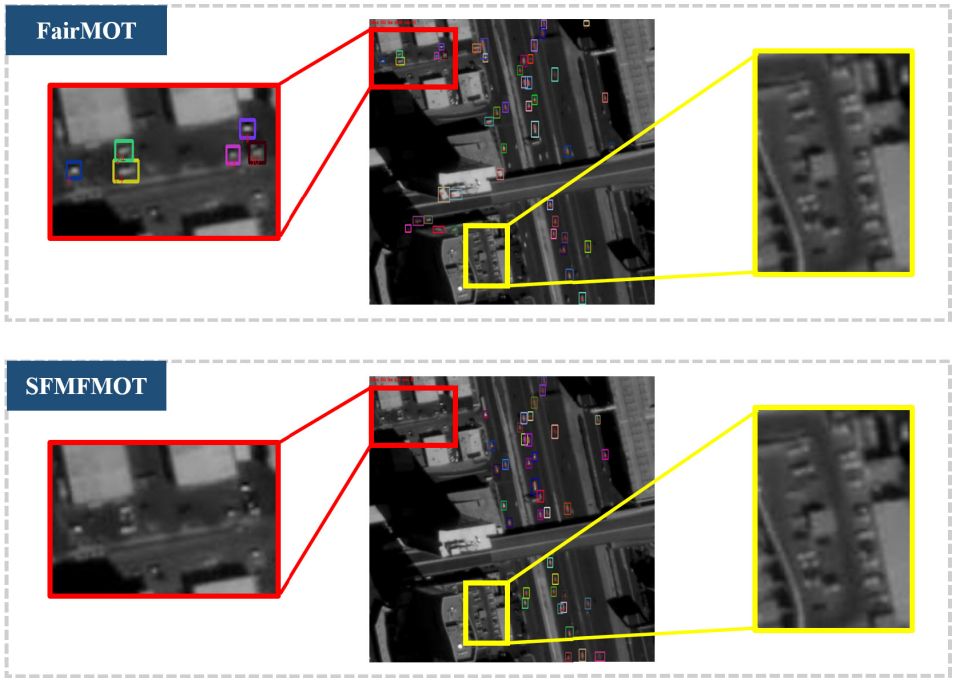


Fig. 24. Example diagram of parked vehicles. The red box area on the left is a roadside parking area, and the yellow box area on the right is a parking area.

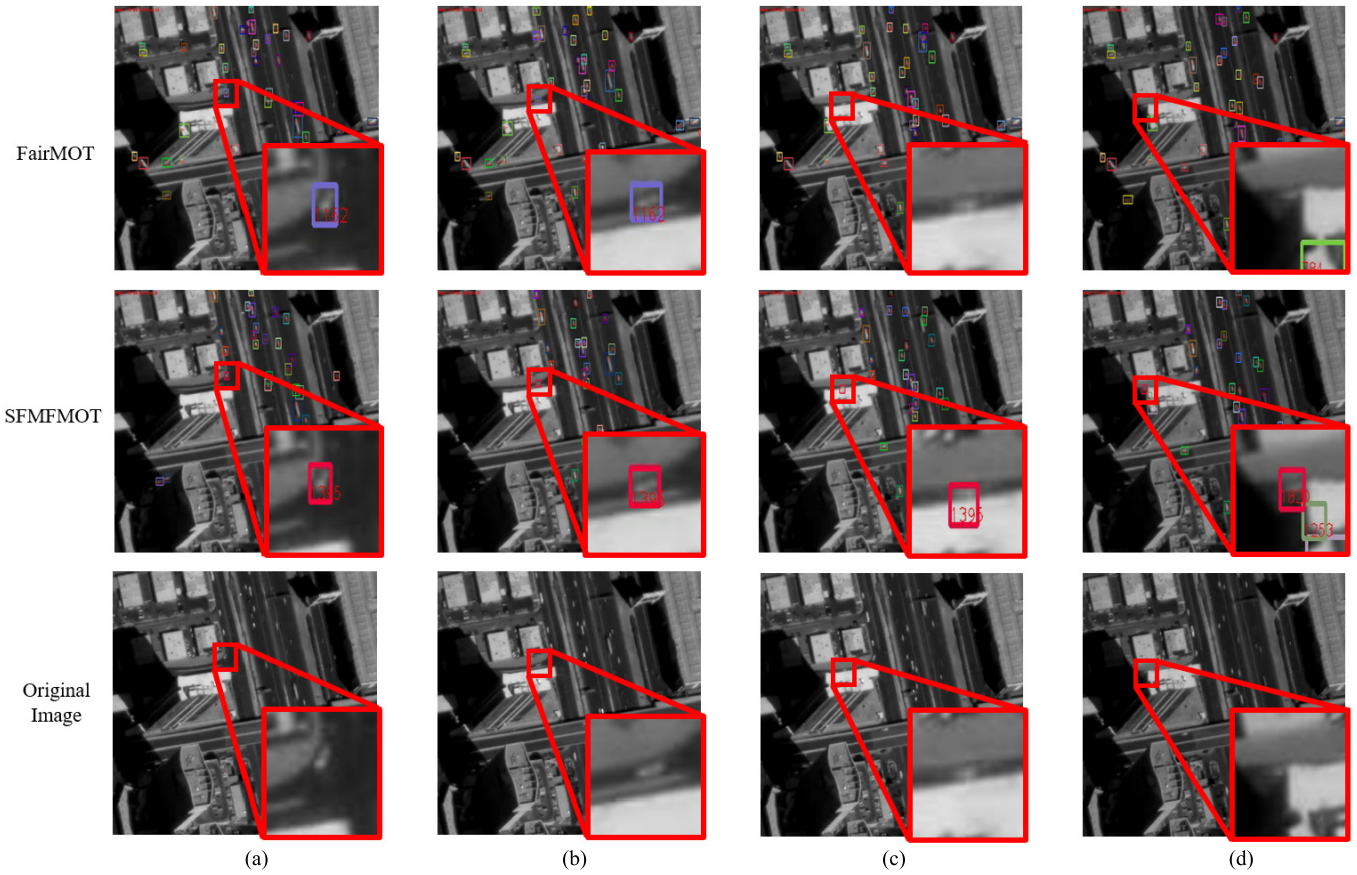


Fig. 25. Example diagram of objects disappearing from view caused by buildings. (a) 1155 Frame. (b) 1213 Frame. (c) 1264 Frame. (d) 1315 Frame.

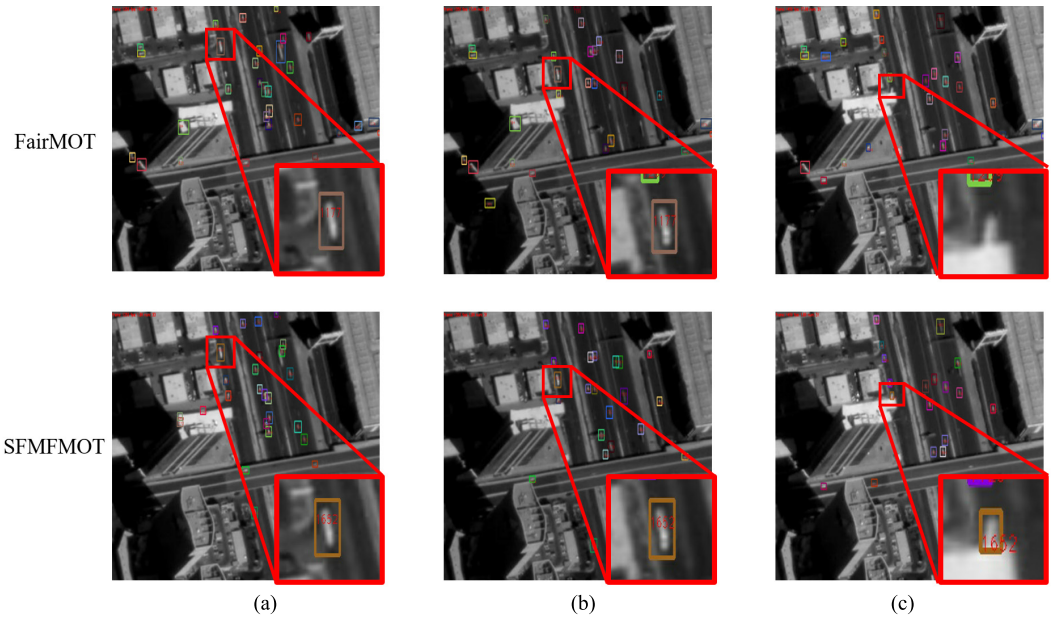


Fig. 26. Example diagram of light variation. (a) 1250 Frame. (b) 1350 Frame. (c) 1450 Frame.

a high confidence score. According to the drawing principle of the precision–recall curve, detection bounding boxes need to be sorted in the order of confidence value from high to low. The sorted detection bounding boxes sequence is denoted as $D = [\text{det}_1, \text{det}_2, \ldots, \text{det}_n]$, where n represents the number of all detection bounding boxes in the whole video. Then, according to the sequence D, each detection bounding box det_i is used as the demarcation point in sequence to allocate the

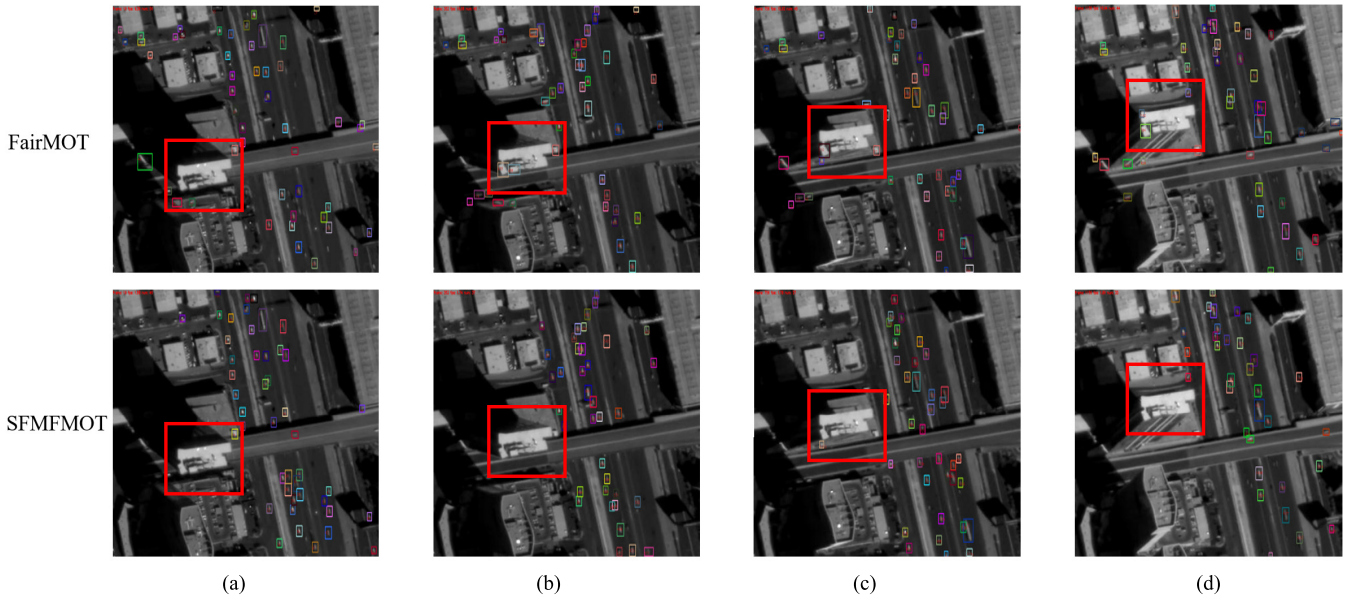


Fig. 27. Visual comparison of moving false alarm problems in high-rise buildings. Red boxes highlight the detection and tracking of targets in high-rise building areas. (a) 19 Frame. (b) 352 Frame. (c) 756 Frame. (d) 1156 Frame.

prediction category of all detection bounding boxes in D. The prediction category of the detection bounding box det_j before the demarcation point ($j <= i$) is a vehicle target, otherwise a nonvehicle target ($j > i$). At this point, the calculation formula for precision_i and recall_i is as follows:

$$\text{precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} = \frac{\text{TP}_i}{\text{Num}_i} = \frac{\text{TP}_i}{i} \tag{6}$$

$$\text{recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} = \frac{\text{TP}_i}{\text{Num}_{\text{gt}}} \tag{7}$$

where Num_{gt} is the number of real vehicle targets in the whole test video, which is a fixed value. Num_i is the number of detection bounding boxes predicted by the detector as vehicle targets, so when the compute node is det_i, $\text{Num}_i = i$.

It is easy to know that as i increases, the denominator of precision_i increases gradually, and the same change of the numerator has less influence on precision_i, this is, the influence of an error detection on precision_i decreases. If there is an error detection with a high confidence score, it will be ranked relatively high in D. At this time, the value i is small, and the error detection will have a significant impact on precision_i, leading to a sharp drop of the curve.

If the re-set confidence is high, the correct detection ranking will be advanced, and the error detection with high confidence score will be ranked later, which will reduce the influence of error detection with high confidence score on precision_i and avoid the sudden drop of the curve.

2) As Long as the Reset Confidence Value Is Higher Than the Confidence Threshold (Set as 0.1 in This Research), the Value Difference Will Not Affect the Final Detection and Tracking Performance: As shown in Fig. 21, for different reset parameters, the difference of precision–recall curves mainly lies in the part of the high-confidence threshold. With the decrease of threshold, curves of different reset parameters will overlap. In detail, for a given confidence threshold T, as long as $S_{\text{new}} > T$, it will not change the prediction category of the bounding box under different values of S_{new}, so it will not affect the precision and recall. In the case that 0.1 is selected as the threshold value in this research, as long as the reset confidence value is greater than 0.1, different values will obtain the same precision–recall value.

To sum up, the selection of reset confidence value S_{new} will not affect the detection performance under the threshold value of T while $S_{\text{new}} > T$.

APPENDIX B
ANALYSIS OF DIFFICULT SCENARIOS

In order to further analyze the performance of the proposed method in difficult scenes and obtain enlightenment for the focus of future research, this study carried out detailed analysis from the aspects of occlusion, parked objects, objects disappearing from view caused by buildings, light variation, and false alarms generated by moving buildings.

A. Problem With Occlusion

As shown in Fig. 23, the trajectory diagram of the 1-min test sequence is shown, in which the red box and yellow box correspond to the two sample areas of weak occlusion and strong occlusion, respectively. It is easy to draw the following conclusions from observation: In the case of weak occlusion, most vehicles can maintain the continuous tracking, and the trajectory results of the proposed method have stronger continuity; In the case of strong occlusion, neither of the two methods can guarantee the continuous tracking of vehicles entering and leaving the overpass. This shows that the proposed method has strong robustness for short-term occlusion, but there are still deficiencies in long-term tracking problems.

B. Problem With Parked Vehicles

The research focus of this study is moving target detection and tracking. In the algorithm design, the interference of static object false alarm can be reduced by introducing the assistance of motion information. As shown in the two sample areas in Fig. 24, some stationary vehicles are still detected by the FairMOT method. In contrast, SFMFMOT does not detect and track stationary vehicles in parking lots and parked on the roadside, indicating that the proposed method has a stronger ability to resist interference from stationary targets.

C. Problem With Objects Disappearing From View Caused by Buildings

As shown in Fig. 25, the test area contains high-rise buildings, which will produce apparent displacement over time, shielding vehicles on the road and making targets disappear from view. In the method proposed in this research, the target that goes out of sight will not be deleted immediately. Instead, the record of the historical position will be removed only when the target does not appear again after a specified frame length $T_{\max}$. In this way, when the target appears again, the target can be tracked quickly, ensuring the continuity of tracking. As shown in Fig. 25, after passing through the occlusion of buildings, SFMFMOT can track the moving object that reappears in the field of view again quickly, while the baseline FairMOT loses the tracked object.

D. Problem With Light Variation

There is a slow light change in satellite videos, and the intensity change between adjacent frames is almost negligible. As shown in Fig. 26, light gradually brightens in the three sequential images over time, but it is not obvious when observed by naked eyes. Observing the target in the highlighted area marked in the red box shows that for the same moving target, the change of light does not interfere with its tracking process. Compared with the baseline FairMOT method, the proposed method has a higher detection rate under the influence of light change, indicating a stronger adaptability to the light changes in satellite videos.

E. Problems With False Alarms Generated by Moving Buildings

As shown in Fig. 27, the position of the buildings in the red box shows a significant shift from south to north over time. Compared with the baseline FairMOT method, SFMFMOT can remove most of the false alarms of moving objects caused by the deviation of high-rise buildings. However, there are still a small number of false alarms that have not been completely removed.

In conclusion, the proposed method shows strong robustness for problems such as weak occlusion, light change, parked vehicles, and objects out of sight caused by buildings, and can remove a large number of false alarms generated by moving buildings. However, this method is still inadequate for long-term occlusion problems such as severe high-rise building occlusion, and there are still some moving false alarms that cannot be completely removed. Therefore, better solutions to these particular problems need to be further explored in the future.

REFERENCES

[1] B. Du, Y. Sun, S. Cai, C. Wu, and Q. Du, "Object tracking in satellite videos by fusing the kernel correlation filter and the three-frame-difference algorithm," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 168–172, Feb. 2018.

[2] H. Li and Y. Man, "Moving ship detection based on visual saliency for video satellite," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1248–1250.

[3] C. Yuan, Z. Liu, and Y. Zhang, "UAV-based forest fire detection and tracking using image processing techniques," in *Proc. Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, Jun. 2015, pp. 639–643.

[4] S. A. Ahmadi, A. Ghorbanian, and A. Mohammadzadeh, "Moving vehicle detection, tracking and traffic parameter estimation from a satellite video: A perspective on a smarter city," *Int. J. Remote Sens.*, vol. 40, no. 22, pp. 8379–8394, Nov. 2019.

[5] S. Xuan, S. Li, M. Han, X. Wan, and G. S. Xia, "Object tracking in satellite videos by improved correlation filters with motion estimations," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1074–1086, Feb. 2020.

[6] Y. Wang, T. Wang, G. Zhang, Q. Cheng, and J.-Q. Wu, "Small target tracking in satellite videos using background compensation," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7010–7021, Oct. 2020.

[7] X. Chen and H. Sui, "Real-time tracking in satellite videos via joint discrimination and pose estimation," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLII-2/W12, pp. 23–29, 2019.

[8] B. Du, S. Cai, and C. Wu, "Object tracking in satellite videos based on a multiframe optical flow tracker," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 3043–3055, Aug. 2019.

[9] J. Shao, B. Du, C. Wu, and L. Zhang, "Can we track targets from space? A hybrid kernel correlation filter tracker for satellite video," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8719–8731, Aug. 2019.

[10] J. Shao, B. Du, C. Wu, M. Gong, and T. Liu, "HRSiam: high-resolution Siamese network, towards space-borne satellite video tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 3056–3068, 2021.

[11] J. Shao, B. Du, C. Wu, and Y. Pingkun, "PASiam: Predicting attention inspired Siamese network, for space-borne satellite video tracking," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Oct. 2019, pp. 1504–1509.

[12] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.

[13] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 474–490.

[14] D. Riahi and G.-A. Bilodeau, "Online multi-object tracking by detection based on generative appearance models," *Comput. Vis. Image Understand.*, vol. 152, pp. 88–102, Nov. 2016.

[15] W. Hu, X. Li, W. Luo, X. Zhang, S. J. Maybank, and Z. Zhang, "Single and multiple object tracking using log-Euclidean Riemannian subspace and block-division appearance model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2420–2440, Dec. 2012.

[16] L. Zhang and L. van der Maaten, "Structure preserving object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1838–1845.

[17] L. Zhang and L. van der Maaten, "Preserving structure in model-free tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 756–769, Apr. 2014.

[18] M. Yang, T. Yu, and Y. Wu, "Game-theoretic multiple target tracking," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Mar. 2007, pp. 1–8.

[19] P. Chen, Y. Dang, R. Liang, W. Zhu, and X. He, "Real-time object tracking on a drone with multi-inertial sensing data," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 131–139, Jan. 2018.

[20] D. Cavaliere, V. Loia, A. Saggese, S. Senatore, and M. Vento, "A human-like description of scene events for a proper UAV-based video content analysis," *Knowl.-Based Syst.*, vol. 178, no. AUG. 15, pp. 163–175, 2019.

[21] V. Carletti, A. Greco, A. Saggese, and M. Vento, "Multi-object tracking by flying cameras based on a forward-backward interaction," *IEEE Access*, vol. 6, pp. 43905–43919, 2018.

[22] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep learning for UAV-based object detection and tracking: A survey," 2021, *arXiv:2110.12638*.

[23] L. W. Sommer, M. Teutsch, T. Schuchert, and J. Beyerer, "A survey on moving object detection for wide area motion imagery," in *Proc. IEEE winter Conf. Appl. Comput. Vis. (WACV)*, Oct. 2016, pp. 1–9.

[24] M. Teutsch and M. Grinberg, "Robust detection of moving vehicles in wide area motion imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 27–35.

[25] L. Sommer, W. Kruger, and M. Teutsch, "Appearance and motion based persistent multiple object tracking in wide area motion imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3878–3888.

[26] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, 2006, pp. 850–855.

[27] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Comput.*, vol. 14, no. 4, pp. 715–770, Apr. 2002.

[28] C. Wu, B. Du, and L. Zhang, "Slow feature analysis for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2858–2874, May 2014.

[29] G. Welch *et al.*, "An introduction to the Kalman filter," in *Proc. SIGGRAPH*, vol. 8, 2001, pp. 1–16.

[30] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Oct. 2019, pp. 5693–5703.

[31] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with Siamese region proposal network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.

[32] S. Zhang, L. Jiao, F. Liu, and S. Wang, "Global low-rank image restoration with Gaussian mixture model," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1827–1838, Jun. 2018.

[33] W. Ao, Y. Fu, X. Hou, and F. Xu, "Needles in a haystack: Tracking city-scale moving vehicles from continuously moving satellite," *IEEE Trans. Image Process.*, vol. 29, pp. 1944–1957, 2019.

[34] G. Kopsiaftis and K. Karantzalos, "Vehicle detection and traffic density monitoring from very high resolution satellite video data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 1881–1884.

[35] J. Zhang, X. Jia, J. Hu, and K. Tan, "Satellite multi-vehicle tracking under inconsistent detection conditions by bilevel K-shortest paths optimization," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, 2018, pp. 1–8.

[36] J. Feng *et al.*, "Cross-frame keypoint-based and spatial motion information-guided networks for moving vehicle detection and tracking in satellite videos," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 116–130, Jul. 2021.

[37] S. C. Wong, V. Stamatescu, A. Gatt, D. Kearney, I. Lee, and M. D. McDonnell, "Track everything: Limiting prior knowledge in online multi-object recognition," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4669–4683, Oct. 2017.

[38] S. Wong, A. Gatt, D. Kearney, A. Milton, and V. Stamatescu, "A competitive attentional approach to mitigating model drift in adaptive visual tracking," in *Proc. 29th Int. Conf. Image Vis. Comput. New Zealand*, 2014, pp. 1–6.

[39] V. Stamatescu, S. Wong, M. D. McDonnell, and D. Kearney, "Learned filters for object detection in multi-object visual tracking," *Proc. SPIE*, vol. 9844, May 2016, Art. no. 98440F.

[40] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[41] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.

[43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[44] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.

[45] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 1504–1509.

[46] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[47] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," 2020, *arXiv:2004.01888*.

[48] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[49] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.

[50] Z. Dong, G. Li, Y. Liao, F. Wang, P. Ren, and C. Qian, "CentripetalNet: Pursuing high-quality keypoint pairs for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10519–10528.

[51] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9657–9666.

[52] X. Zhou, D. Wang, and P. Krähenbhl, "Objects as points," 2019, *arXiv:1904.07850*.

[53] D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. Autom. Control*, vol. AC-24, no. 6, pp. 843–854, Dec. 1979.

[54] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE J. Ocean. Eng.*, vol. JOE-8, no. 3, pp. 173–184, Jul. 1983.

[55] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics*, vol. 2, nos. 1–2, pp. 83–97, Mar. 2010.

[56] N. Mahmoudi, S. M. Ahadi, and M. Rahmati, "Multi-target tracking using CNN-based features: CNNMTT," *Multimedia Tools Appl.*, vol. 78, no. 6, pp. 7077–7096, Mar. 2019.

[57] Z. Zhou, J. Xing, M. Zhang, and W. Hu, "Online multi-target tracking with tensor-based high-order graph matching," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, 2018, pp. 1809–1814.

[58] K. Fang, Y. Xiang, X. Li, and S. Savarese, "Recurrent autoregressive networks for online multi-object tracking," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 466–475.

[59] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[61] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[62] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[63] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," 2016, *arXiv:1603.00831*.

[64] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3415–3424.

[65] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1367–1376.

[66] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[67] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Sydney, NSW, Australia: PMLR, Aug. 2017, pp. 1321–1330.

Jialian Wu received the B.S. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2019, where she is currently pursuing the M.S. degree with the School of Geodesy and Geomatics.

Her major research interests include multitarget tracking, deep learning, and computer vision.

Xin Su (Member, IEEE) received the B.S. degree in electronic engineering from Wuhan University, Wuhan, China, in 2008, and the Ph.D. degree in image and signal processing from Télécom ParisTech, Paris, France, in 2015.

He was a Post-Doctoral Researcher with the Team SIROCCO, Institut National de Recherche en Informatique et en Automatique, Rennes, France. He is currently an Assistant Professor with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include multitemporal remote-sensing image processing, multiview image processing, and 3-D video communication.

Qiangqiang Yuan (Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively.

In 2012, he joined the School of Geodesy and Geomatics, Wuhan University, where he is currently a Professor. He has authored or coauthored more than 90 research articles, including more than 70 peer-reviewed articles in international journals, such as *Remote Sensing of Environment*, *ISPRS Journal of Photogrammetry and Remote Sensing*, the IEEE TRANSACTION ON IMAGE PROCESSING, and the IEEE TRANSACTION ON GEOSCIENCE AND REMOTE SENSING. His research interests include image reconstruction, remote-sensing image processing and application, and data fusion.

Dr. Yuan was a recipient of the Youth Talent Support Program of China in 2019, the Top-Ten Academic Star of Wuhan University in 2011, and the recognition of Best Reviewers of the IEEE GRSL in 2019. In 2014, he received the Hong Kong Scholar Award from the Society of Hong Kong Scholars and the China National Postdoctoral Council. He is an Associate Editor of five International Journals and has frequently served as a Referee for more than 40 international journals for remote sensing and image processing.

Huanfeng Shen (Senior Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

He is currently a Distinguished Professor with Wuhan University, where he serves as an Associate Dean with the School of Resource and Environmental Sciences. He was or is the Principal Investigator (PI) of two projects supported by the National Key Research and Development Program of China and six projects supported by the National Natural Science Foundation of China. He has authored or coauthored more than 150 peer-reviewed international journal articles, where over 60 appeared in IEEE Journals, and published four books as a Chief Editor. His research interests include remote-sensing image processing, multisource data fusion, and intelligent environmental sensing.

Dr. Shen is a fellow of the Institution of Engineering and Technology (IET), the Education Committee Member of Chinese Society for Geodesy Photogrammetry and Cartography, and the Theory Committee Member of Chinese Society for Geospatial Information Society. He was a recipient of the First Prize in Natural Science Award of Hubei Province in 2011, the First Prize in Nature Scientific Award of China's Ministry of Education in 2015, and the First Prize in Scientific and Technological Progress Award of Chinese Society for Geodesy Photogrammetry and Cartography in 2017. He is a Senior Regional Editor of the Journal of Applied Remote Sensing and an Associate Editor of *Geography and Geo-Information Science and Journal of Remote Sensing*.

Liangpei Zhang (Fellow, IEEE) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He is a Chang-Jiang Scholar Chair Professor appointed by the Ministry of Education of China in State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. He was a Principal Scientist for the China State Key Basic Research Project from 2011 to 2016 appointed by the Ministry of National Science and Technology of China to lead the Remote Sensing Program in China. He has authored or coauthored more than 700 research articles and five books. He is the Institute for Scientific Information (ISI) Highly Cited Author. He is the holder of 30 patents. His research interests include hyper-spectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a fellow of the Institution of Engineering and Technology (IET). He was a recipient of the 2010 Best Paper Boeing Award, the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing (ASPRS), and the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). His research teams won the top three prizes of the IEEE GEOSCIENCE AND REMOTE SENSING Society (GRSS) 2014 Data Fusion Contest, and his students have been selected as the Winners or the Finalists of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) Student Paper Contest in recent years. He is the Founding Chair of IEEE GRSS Wuhan Chapter. He also serves as an Associate Editor or an Editor of more than ten international journals. He is serving as an Associate Editor of the IEEE TRANSACTION ON GEOSCIENCE AND REMOTE SENSING.