HRMOT: two-step association based multi-object tracking in satellite videos enhanced by high-resolution feature fusion.

WU, Y., ZHANG, X., LIU, Q., XUE, D., SUN, H. and REN, J.

2024

HRMOT: Two-Step Association Based Multi-object Tracking in Satellite Videos Enhanced by High-Resolution Feature Fusion

Yuqi Wu[1,2], Xiaowen Zhang[1,2], Qiaoyuan Liu[1(✉)], Donglin Xue[1(✉)], Haijiang Sun[1], and Jinchang Ren[3]

[1] The Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
{liuqy,xuedl,sunhj}@ciomp.ac.cn
[2] The University of Chinese Academy of Sciences, Beijing 100049, China
{wuyuqi21,zhangxiaowen22}@mails.ucas.ac.cn
[3] The National Subsea Centre, Robert Gordon University, Aberdeen AB21 0BH, U.K.
jinchang.ren@ieee.org

Abstract. Multi-object tracking in satellite videos (SV-MOT) is one of the most challenging tasks in remote sensing, its difficulty mainly comes from the low spatial resolution, small target and extremely complex background. The widely studied multi-object tracking (MOT) approaches for general images can hardly be directly introduced to the remote sensing scenarios. The main reason can be attributed to: 1) the existing MOT approaches would cause a significant missed detection of the small targets in satellite videos; 2) it is difficult for the general MOT approaches to generate complete trajectories in complex satellite scenarios. To address these problems, a novel SV-MOT approach enhanced by high-resolution feature fusion (HRMOT) is proposed. It is comprised of a high-resolution detection network and a two-step based association strategy. In the high-resolution detection network, a high-resolution feature fusion module is designed to assist the detection by maintaining small object features in forward propagation. Based on high-quality detection, the densely-packed weak objects can be effectively tracked by associating almost every detection box instead of only the high score ones. Comprehensive experiment results on the representative satellite video datasets (VISO) demonstrate that the proposed HRMOT can achieve a competitive performance on the tracking accuracy and the frequency of ID conversion with the state-of-the-art (SOTA) methods.

Keywords: Multi-object Tracking · Satellite Video · High-resolution Feature Fusion · Data association

1 Introduction

With the rapid development of remote sensing earth observation technology: the successive launch of the video satellites such as Jilin-1, SkySat. etc. [1, 2], remote sensing video data is becoming increasingly easy to obtain and play an important role in the tasks of

military precision strike, civilian traffic flow estimation, fire detection, etc. [3–5]. Video satellites finish dynamic ground observation by gazing at specific areas for a certain period of time, could provide richer temporal information and a larger observation area. Therefore SV-MOT has become one of the hotspots in the satellite video processing and analysis. Different from the general MOT approaches, SV-MOT would have to face more challenges such as low resolution, complex background, especially targets in very small sizes [6–9].

In recent decades, general MOT have made significant breakthroughs, which could be divided into two branches: detection-based tracking (DBT) and joint track and detection (JDT). Specifically, the DBT approaches would first get detection boxes in every frame and then do association over time with the detection and association modules have no affection between each other, whose tracking performance is mainly depended on the quality of detection. However, the independent structure has low tracking efficiency and it could not make the joint optimization together, they can achieve a competitive performance easily but with a heavy training cost; while the JDT approaches would do detection and tracking with one network, so as to estimate the objects and learn re-ID features simultaneously. These models are more efficient and fast because of considering detection and association at the same time, but it is difficult to maintain a balance between detection and association, which may lead to a tracking performance decrease. SORT [10] are the most representative DBT methods, The SORT algorithm is based on a Kalmanfilter to establish an observation model foundation, which can accurately predict the position and motion speed of the target, and achieve tracking of the target. In order to reduce the number of ID conversions, Deepsort [11] introduced ReID features on the basis of SORT, which can better distinguish different feature targets in videos. Bytetrack first associates high score detection boxes and then associate low score detection boxes, finally distinguishes between real targets and false alarms through the similarity of trajectories. In the JDT branches, FairMOT [12] optimizes both the detection model and the ReID model, and achieves a balance between detection and ReID tasks through some strategies. CenterTrack [13] proposes an integrated network of detection and tracking based on key point. In the data association stage, CenterTrack performers data association based on the distance between the predicted offset of the center point and the center point of the tracked target in the previous frame. Since there is a big difference between the general videos and the satellite videos, directly apply any methods discussed above to satellite videos, there will be a significant performance degradation. This is because compared to natural scenes, the proportion of target pixels in satellite videos is low, the object size in satellite videos is about 10–100 pixels, increasing the difficulty of detection and the background of satellite videos is complex such as widespread cloudiness, reflective noise, which may cause more missed detection and false alarm.

To fill the gap between SV-MOT and MOT, a few tracking methods for satellite videos are proposed one after another. SV-MOT methods could also be divided into the branches of DBT and JDT. The DBT method such as CKDNet-SMTNet [14] proposed a spatial motion information-guided network for tracking performance enhancement, extracting spatial information of targets in the same frame and motion information of targets in consecutive frames. However, its double LSTM structure leads the approach unable

to track online. Since the JDT methods such as TGraM [15] proposed a graph based spatiotemporal inference module to explore potential high-order correlations between video frames, modeling multi-object tracking as a graph information inference process from the perspective of multi task learning. CFTracker [16] proposed a cross frame feature update module and a method for cross frame training. But their network cannot handle occlusion issues well and training is very time-consuming.

It is worth mentioning that although some progresses have been made by existing SV-MOT methods, The high-frequency tracking miss and ID switches while dealing with dense targets occlusion are still the key issues which have not been well handled. In our opinion the missing tracking problem mainly comes from a poor detection on small objects, and the high frequency of ID switches mainly come from a weak data association. Therefore, this paper focuses on both the detection and association parts, and proposes a two-step association based multi-object tracking approach for satellite video enhanced by high-resolution feature fusion (HRMOT). For a better tracking performance, we follow the branch of DBT, which do detection at first and then association. In detail, a novel detection module with high-resolution feature fusion for small targets is introduced, combined with a two-step association for tracklets prediction. Comprehensive experiment results on VISO datasets [17] demonstrate, the proposed SV-MOT can achieve a comparable performance with the SOTA methods.

2 Method

2.1 High-Resolution Detection for Network

In the satellite videos, the appearance information of object is weak because the object is very small in satellite scenarios and the background is complex, such as cloud interference, which result in a great deal of missed detection. A large-scale of missed detection is one of the main reasons that general MOT cannot be introduced into satellite videos. If there are not enough detection boxes generated at the beginning, any data association have to face a tracking failure. Therefore, in this paper the general detection module is replaced by a small object detection module to adapt the characteristics of satellite video. At present, most small target detectors obtain strong semantic information through downsampling, and then upsampling to recover high-resolution location information. However, this approach can lead to the loss of a large amount of effective information during the continuous upsampling and downsampling process. HRNet achieves the goal of strong semantic information and precise positional information by parallelizing multiple resolution branches and continuously exchanging information between different branches. So the most representative concept of multi-resolution is introduced for an essential improvement in tracking performance. Specifically, the most representative detection framework Yolov5 [18] is set as the baseline, and a high-resolution feature fusion strategy is designed to make full use of features in muti-resolution, whose network structure is shown in Fig. 1.

The idea of HRNet [19] is introduced to Yolov5 for high-resolution feature fusion. However, different from the original HRNet which only uses the highest resolution feature map as the output of the model, HRNetv2 [20] added a concatenate operation for feature maps with different resolutions, and then subsampled the combined feature

maps with average pooling to obtain multiple feature maps with different resolutions. Therefore, in this paper an operation similar to HRNetv2 is adopted. In specific, the HRNet retains high-level semantic information by adopting a parallel structure with different resolutions. Within the forward propagation, multiple resolutions interact between different branches, which further reduce the impact caused by the decreasing channel dimensions. The HRNet composes of basic blocks and fuse layers. The input of the basic block is $x_{in} \in \mathbb{R}^{H\times W\times C}$, and the output $x_{out} \in \mathbb{R}^{H\times W\times C}$ of each basic block can be expressed as:

$$x_{out} = x_{in} + f(f(x_{in})), \tag{1}$$

where $f(\cdot)$ denotes the convolutional layer with a Batch Normalization layer (BN) and ReLU activate function layer. As for the fuse layer, the input is combined with multi-resolution features, which are concatenated in channel dimension. Then the concatenated feature goes through an operation $f(\cdot)$ to generate the output, where the output channel varies with the position of the fuse layer.

Besides, to reduce the information loss of feature maps with different resolutions in the sampling process and reduce the calculation amount, the feature maps with the current resolution are also concatenated. Finally, three feature maps with different resolutions and channel numbers are obtained. These feature maps are then fed into the detection head for further prediction.

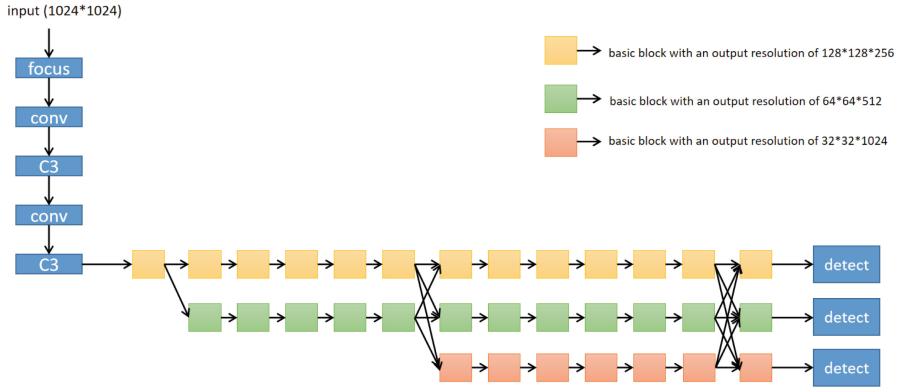


Fig. 1. Structure of the proposed high-resolution detection network.

2.2 Two-Step Association Based SV-MOT

Targets in satellite videos either with small size or affected by complex background would both easily get a low detection confidence. Common MOT associations would always first filter out the detection boxes with low confidence, and then do association. But in satellite scenarios, the real targets with small size cannot always be detected with

a high confidence. So, the filter operation based on confidence may cause a lot of small objects lost. But associating all the detection boxes may result in tracking a lot of false alarm. To tackle this issue, the similarity of the trajectories is applied to distinguish the false alarms of correct targets. Specifically, in order to make more objects in satellite video be associated to form a trajectory, a two-step association method is proposed which firstly associates the high-score detection frames then low-score detection frames, finally filters out false alarms through the similarity with trajectory. The flowchart can be seen from Fig. 2.

Below, we will specifically discuss data association methods. (1)According to an adaptive threshold, the detection boxes are divided into high-score detection boxes and low-score detection boxes. (2) the trajectory would be predicted in the new frame. (3) the high-score detection boxes in new frame are associated with the existing trajectory, The similarity between the predicted trajectory and the high-score detection boxes is calculated by IOU. The Hungarian algorithm [22] was used to solve the optimal matching between the track and the detection boxes. (4) the unmatched trajectories are associated with the low-score detection boxes in order to achieve better tracking performance on scale change, occlusion and motion ambiguity, the false alarms are distinguished from the real target by the similarity with trajectories. (5) the high-score detection boxes that have no match is initialized to a new track, and the tracklets which have not been matched over a certain number of frames would be deleted.

Here we will introduce our object motion model, Kalman filter [21], which is an algorithm that utilizes linear system state equations to perform optimal estimation of system state through system input and output observation data. Due to the inclusion of noise and interference in the observed data, the optimal estimation can also be seen as a filtering process. As long as it is a dynamic system with uncertain information, Kalman filter can make informed speculations about what the system will do next. Even with noise information interference, Kalman filter can usually figure out what is happening and find imperceptible correlations between images. Therefore, Kalman filtering is very suitable for constantly changing systems, and its advantages include small memory footprint (only retaining the previous state), fast speed, making it an ideal choice for real-time problems and embedded systems. In the paper, the state vector of each object was chosen to be a eight-tuple:

$$X = [x, y, a, h, \dot{x}, \dot{y}, \dot{a}, \dot{h}] \tag{2}$$

where x and y represent the horizontal and vertical coordinates of the object, while a and h represent the aspect ratio and the height of object's bounding box respectively. The last four parameters represent their rate of change, respectively.

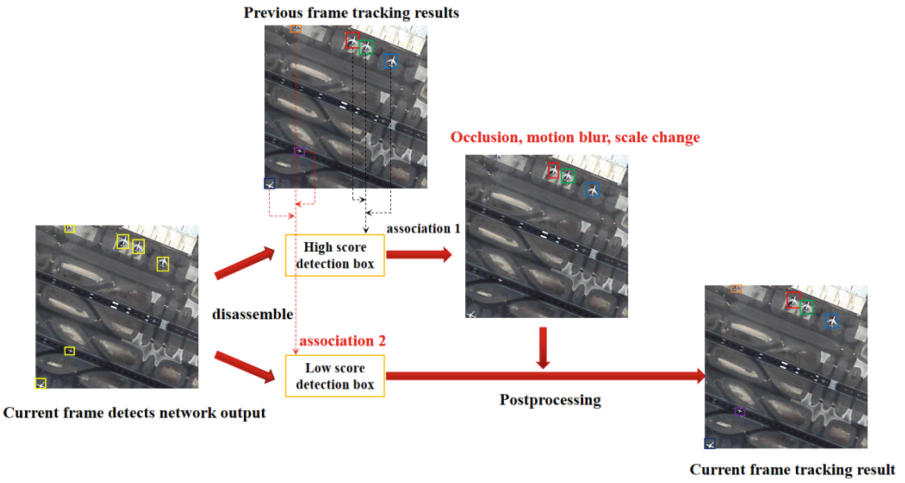


Fig. 2. Flowchart of the two-step association.

3 Experiments

In this section, we firstly introduce the details of implementation, including the dataset, metrics and experimental setup. Then ablation experiments are performed to verify the effectiveness of each module proposed. Finally, performance of HRMOT and start-of-art methods are compared, experimental results demonstrate the superiority of the proposed HRMOT.

3.1 Datasets

We validate our method on VISO dataset [17] which consists of 47 satellite videos captured by the Jilin-1 satellite. Jilin-1 satellite is a video satellite system launched by Chang Guang Satellite Technology Co., Ltd. The satellite video obtained by Jilin-1 satellite consists of a series of true color images. The coverage area of real scenes can reach several square kilometers. The majority of instances in VISO have a size smaller than 50 pixels and the videos cover different types of city-scale elements, such as roads, bridges, lakes, and a variety of moving vehicles. VISO is a diverse and comprehensive dataset which consist of different traffic situations such as dense lanes and traffic jams. Therefore, the satellite videos in dataset cover a wide range of challenges, including complex background, illumination variations, dense targets and so on. Every video is in the size of 1024×1024. The frame numbers are different from 324 - 326. We randomly selected 20 videos as training set and 5 unrepeated videos as testing set from the VISO dataset.

3.2 Evaluation Metrics

To evaluate the performance of different methods and the module that we proposed in multi-object tracking in satellite video. Six representative metrics are utilized for

evaluation, including false positives (FP), false negatives (FN), ID switches (IDs), ID F1 score (IDF1), multi-object tracking precision (MOTP) and multi-object tracking accuracy (MOTA), the discrimination of the metrics are as follow:

(1) FP: the number of false positives in the whole video.
(2) FN: the number of false negatives in the whole video.
(3) IDs: the total number of ID switches.
(4) IDF1: the ratio of correctly identified tests to the average true and calculated number of tests.
(5) The MOTP is defined as follow, in which d_i^t is the euclidean distance between the i-th target and the assumed position in frame t, c_t is the number of matches in frame t:

$$MOTP = \frac{\sum d_t^i}{\sum c_t} \tag{3}$$

(6) The MOTA is defined as follow, in which GT is the ground truth:

$$MOTA = 1 - \frac{FN + FP - IDs}{GT} \tag{4}$$

3.3 Experimental Setup

In our HRMOT, the detection part with high-resolution feature fusion is first trained with the VISO dataset, then followed by the two-step association strategy to get the trajectories of multi-objects in the satellite video. 20 videos datasets in VISO are trained on GPU 3090. During the training process, random gradient descent (SGD) is selected as the optimizer, the initial learning rate is 1e–2, and the weight attenuation parameter is 5e–4. In the tracking process, the maximum number of lost track frames is set to 30 frames, and the matching threshold of tracking is set to 0.8. We compare our method with the state-of-the-art trackers in natural and satellite scenarios, such as SORT, FairMOT, TGraM and Bytetrack. All methods use the same satellite video training and all experiments were conducted on RTX 3090 GPU to ensure a fair comparison.

3.4 Ablation Experiments

To verify the effectiveness of separate modules proposed in HRMOT, 5 representative videos in VISO datasets (as shown in Fig. 3) [17] with challenges like serious occlusion, complex background are selected as testing set. Yolov5 + SORT was selected as our baseline methods. Each module is integrated independently into the baseline for an effective evaluation. We compare the proposed HRNet + baseline method, YOLOX + baseline method, two-step association + baseline method and HRNet + two-step association + baseline method(HRMOT) with the baseline method under the above 5 test videos. The experimental result of adding each module to baseline are shown in Table 1, the tracking performance of the 5 tested videos produced by HRMOT are shown in Table 2.

It can be seen from Table 1 that compared with the baseline, compared with the baseline, Yolov5 + Byte (two-step association strategy added) has achieved a great

tracking enhancement by improving MOTA by 36% and IDF1 by 45.3%, at the same time IDs has also been significantly reduced, which verifies the effectiveness of the two-step association. We also try a latest detection model Yolox [23] for evaluation, compare to Yolox + Byte, Yolov5 + Byte achieves better performance by increasing the MOTA by 2.7% and IDF1 by 0.6%, which double confirmed the effective of Yolov5 in multi-object tracking in satellite videos. Therefore, the high-resolution feature fusion module is integrated into the framework of Yolov5, after integrating HRNet into Yolov5, We have achieved improvements in tracking performance by improving MOTA by 4.1% and IDF1 by 1%. It can be seen that the performance of HRMOT (Yolov5 + HRNet + Byte) is far superior to other methods, the MOTA reached 74% and the IDF1 reached 79.6%.

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

Fig. 3. Five selected test satellite videos in VISO

Table 1. Ablation study for different detection headers and data association methods.

	IDF1	FP	FN	IDs	MOTA	MOTP
Yolov5 + SORT (baseline1)	33.9%	16011	11248	14990	33.9%	0.38
Yolov5 + Byte	79.2%	1098	16206	199	72.6%	0.296
Yolox + Byte	78.6%	1405	17681	161	69.9%	0.307
HRMOT	79.6%	1020	15395	208	74%	0.294

Table 2. Quantitative results of our method on VISO test set.

	IDF1	FP	FN	IDs	MOTA	MOTP
1	94.0%	165	639	11	92.7%	0.265
2	80.0%	151	5240	42	69.5%	0.329
3	88.6%	146	294	14	89.2%	0.205
4	69.5%	378	6268	84	66.6%	0.305
5	76.3%	180	2954	57	69.6%	0.303
overall	79.6%	1020	15395	208	74.0%	0.294

3.5 Comparison with Other Methods

In this section, SOTA approaches including FairMOT [12], SORT [10], Bytetrack [24] and TGraM [15] are compared with the proposed method on the test of VISO, and the experimental results are shown in Table 3. It can be seen from Table 3 that we have achieved the best results on MOTA, IDF1 and other metrics, the proposed HRMOT achieved 74% of MOTA and 79.6% of IDF1, much better than the previous work. Specifically, for the most representative MOT approach-Bytetrack, we can achieve 4.1% higher on MOTA and 1% higher on IDF1. The superiority of the proposed method not only reflected in accuracy but also in speed. Although our tracking method is based on the paradigm of DBT, our tracking part does not have a deep network model and the detection network is simple, with fast inference speed. Our tracking speed can reach over 100 FPS under satellite video, making it suitable for in orbit tracking.

Somevisualization results of the proposed method were shown in Fig. 4, Fig. 5 and Fig. 6. The tracked trajectories comparison between the proposed method and the SOTA multi-object tracking methods were shown in Fig. 4 and Fig. 5. it can be concluded that the trajectories of TGraM, FairMOT, SORT and other methods tend to be fragmented and confused. But in Fig. 4(d) (e) and Fig. 5(d) (e) with our two-step association strategy could perform a stable tracking and the tracklets are tend to be more completed, which demonstrate the availability of the two-step association. Finally, in Fig. 4(f) and Fig. 5(f) with high-resolution feature fusion added, the number of tracklets have been increased, which intuitively demonstrating the effectiveness of our detection network, Yolov5 + HRNet. Key frames of the representative visual experimental results on the challenges of occlusion and intensive targets are shown in Fig. 6. It can be seen in Fig. 6(a) that when two cars meet and block each other, the IDs of the two cars remain unchanged and there is no missed tracking, when a vehicle crosses the bridge, the ID information of the vehicle remains unchanged before and after crossing the bridge, which indicates that our method can effectively handle the problem of background occlusion, and mutual occlusion of targets in complex satellite scenes. It can be seen in Fig. 6(b) that our method can maintain simultaneous and stable tracking of multiple targets in satellite scenes with dense targets. Overall, Facing rather complex tracking challenges, our HRMOT can achieve more stable tracking, forming more complete trajectories with fewer ID conversions and higher accuracy.

Table 3. Comparison of the state-of-the-art methods on VISO test set.

	IDF1	FP	FN	IDs	MOTA	MOTP
FairMOT	22.3%	4014	51454	3036	8.4%	0.523
Baseline1	33.9%	16011	11248	14990	33.9%	0.38
TGraM	32.6%	3240	45621	2548	13.3%	0.475
Yolox + Byte	78.6%	1405	17681	161	69.9%	0.307
Yolov5 + Byte	79.2%	1098	16206	199	72.6%	0.296
HRMOT	79.6%	1020	15395	208	74%	0.294

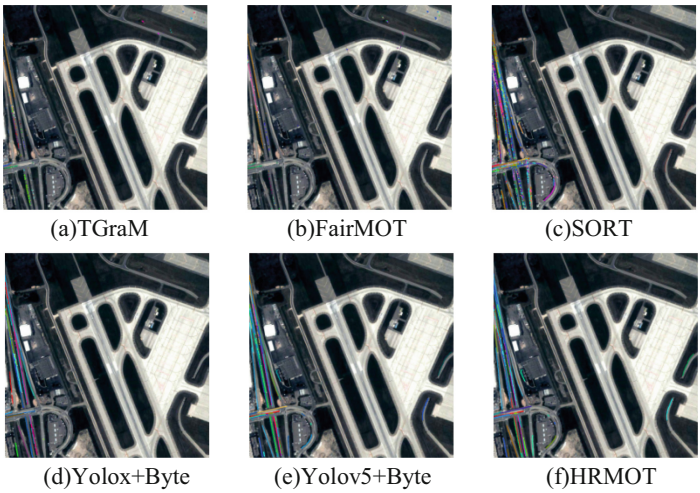


Fig. 4. The tracked trajectories comparison between the proposed method and the SOTA multi-object tracking methods. Different colors represent different trajectories(video1).

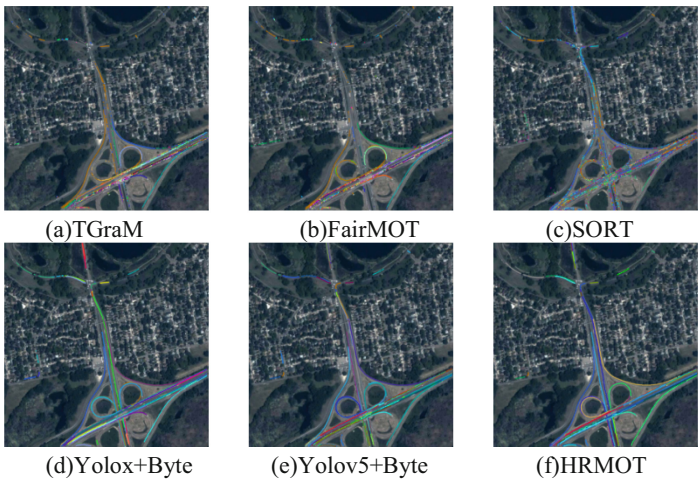
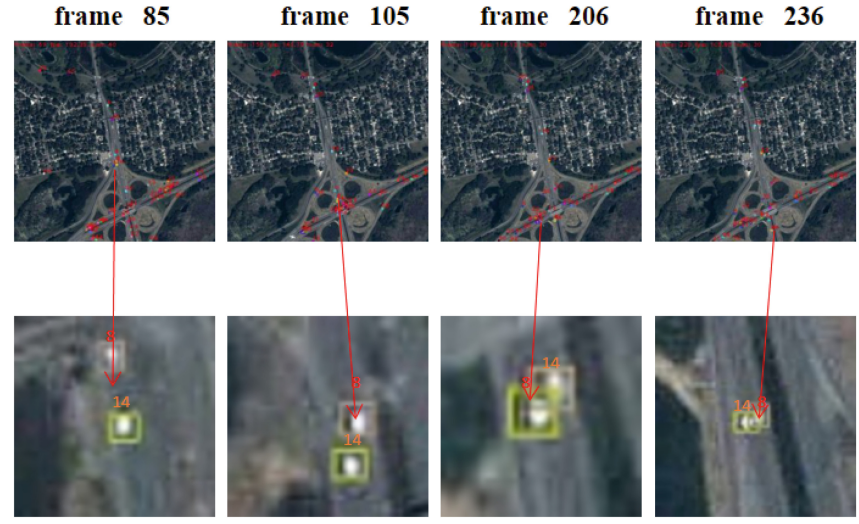
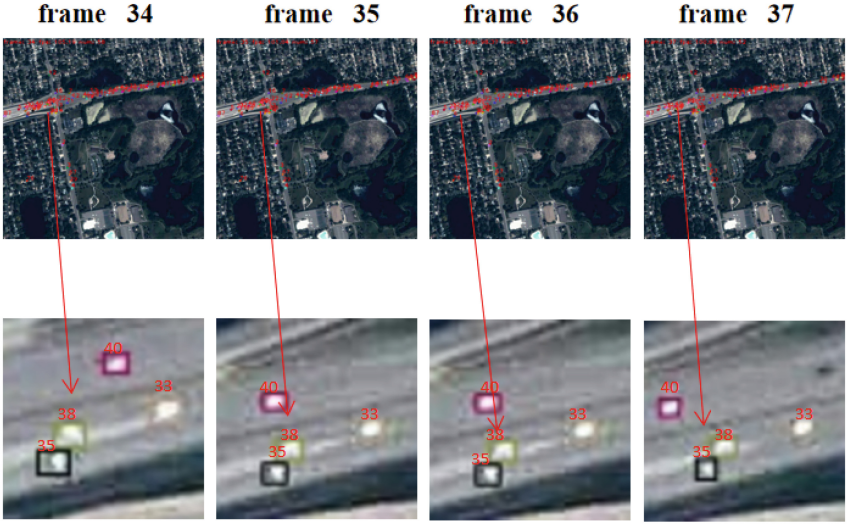


Fig. 5. The tracked trajectories comparison between the proposed method and the SOTA multi-object tracking methods. Different colors represent different trajectories(video2).



（a）occlusion



（b） intensive targets

Fig. 6. Detail experimental results on the challenges of occlusion and intensive targets.

4 Conclusion

In this paper, a novel multi-object tracking approach for satellite videos called HRMOT is proposed. Which consist of a high-resolution detection network for small objects and a two-step association. Compared with the SOTA approaches, our method achieves higher tracking accuracy and a lower frequency on ID switches. Nevertheless, our method is

still inadequate while handling similar objects with cross trajectories. In the future work, we will focus on these problems for tracking performance improvement [25–27].

References

1. Aoran, X., Zhongyuan, W., Lei, W., et al.: Super-resolution for "Jilin-1" satellite video imagery via a convolutional network. Sensors **18**(4), 1194 (2018)
2. Wei, X.U., et al.: Target fast matching recognition of on-board system based on Jilin-1 satellite image. Opt. Precis. Eng. **25**(1), 255–262 (2017)
3. Banks, Adrian P, Dhami, et al.: Normative and Descriptive Models of Military Decisions to Deploy Precision Strike Capabilities. Military Psychology 26(1), 33 (2014)
4. Toth, C.K., Grejner-Brzezinska, D.: Extracting dynamic spatial data from airborne imaging sensors to support traffic flow estimation. ISPRS J. Photogramm. Remote. Sens. **61**(3–4), 137–148 (2016)
5. Dimitropoulos, K., Barmpoutis, P., Grammalidis, N.: Spatio-temporal flame modeling and dynamic texture analysis for automatic video-based fire detection. IEEE Trans. Circuits Syst. Video Technol. **25**, 339–351 (2015)
6. Chen, L., Ai, H., Zhuang, Z., et al.: Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In: IEEE International Conference on Multimedia, pp. 1–6. IEEE Computer Society (2018)
7. Peng, J., et al.: Chained-tracker: chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In: Proceedings of the European Conference on Computer Vision (2020)
8. Yan, Y., Ren, J., Zhao, H., et al.: Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos. Cogn. Comput. **10**, 94–104 (2018)
9. Liu, Q., Ren, J., Wang, Y., et al.: EACOFT: an energy-aware correlation filter for visual tracking. Pattern Recogn. **112**, 0031–3203 (2021)
10. Bewley, A., Ge, Z,. Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: Phoenix, A.Z. (ed.) 2016 IEEE International Conference on Image Processing (ICIP), pp. 3464–3468 (2016)
11. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649 (2017)
12. Zhang, Y., et al.: FairMOT: on the fairness of detection and re-identification in multiple object tracking. Int. J. Comput. Vision **129**, 3069–3087 (2021)
13. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: Proceedings of the European Conference on Computer Vision (2021)
14. Jie Feng, A., 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. **177**, 116–130 (2021)
15. He, Q., Sun, X., Yan, Z., Li, B., Fu, K.: Multi-object tracking in satellite videos with graph-based multitask modeling. IEEE Trans. Geosci. Remote Sens. **60**, 1–13 (2022)
16. Kong, L., Yan, Z., Zhang, Y., Diao, W., Zhu, Z., Wang, L.: CFTracker: multi-object tracking with cross-frame connections in satellite videos. IEEE Trans. Geosci. Remote Sens. **61**, 1–14 (2023)
17. Yin, Q., et al.: Detecting and tracking small and dense moving objects in satellite videos: a benchmark. IEEE Trans. Geosci. Remote Sens. **60**, 1–18 (2022)
18. Mekhalfi, M.L., Nicolò, C., Bazi, Y., Rahhal, M.M.A., Alsharif, N.A., Maghayreh, E.A.: Contrasting YOLOv5, transformer, and efficientdet detectors for crop circle detection in desert. IEEE Geosci. Remote Sens. Lett. **19**, 1–5 (2022)

19. Sun K., Xiao, B., Liu D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, pp. 5686–5696 (2019)
20. Sun, K., Zhao, Y., Jiang, B., et al.: High-resolution representations for labeling pixels and regions. arXiv (2019)
21. Welch, G., Bishop, G.: An Introduction to the Kalman Filter. University of North Carolina at Chapel Hill (1995)
22. Mills-Tettey, A., Stent, A., Dias, M. B.: The Dynamic Hungarian Algorithm for the Assignment Problem with Changing Costs. Carnegie mellon university (2007)
23. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: exceeding YOLO series in 2021. arXiv (2021)
24. Zhang, Y.: ByteTrack: multi-object tracking by associating every detection box. In: Proceedings of the European Conference on Computer Vision (2021)
25. Li, Y., et al.: Cbanet: an end-to-end cross band 2-d attention network for hyperspectral change detection in remote sensing. IEEE Trans. Geosci. Remote Sens. **61** (2023)
26. Luo, F., Zhou, T., Liu, J., Guo, T., Gong, X., Ren, J.: Multiscale diff-changed feature fusion network for hyperspectral image change detection. IEEE Trans. Geosci. Remote Sens. **61**, 1–13 (2023)
27. Liu, Q., Ren, J., Wang, Y., Wu, Y., Sun, H., Zhao, H.: EACOFT: an energy-aware correlation filter for visual tracking. Pattern Recogn. **112**, 107766 (2021)