











Recent Advances in Intelligent Processing of Satellite Video: Challenges, Methods, and Applications

Shengyang Li , Xian Sun , *Senior Member, IEEE*, Yanfeng Gu , *Senior Member, IEEE*, Yixuan Lv ,
Manqi Zhao , Zhuang Zhou , Weilong Guo , Yuhan Sun , Han Wang , and Jian Yang 

Abstract—Intelligent processing of satellite video focuses on extracting specific information of ground objects and scenes from earth observation videos through intelligent image/video processing technology, which has important applications in fields such as traffic monitoring, resource monitoring, and environmental monitoring. The integration of deep learning technology in satellite video processing has led to significant advancements in tasks such as object detection and object tracking, expanding into emerging research areas such as satellite video scene classification and object segmentation. However, there is no comprehensive review and summary in the intelligent processing of satellite video. This article presents a systematic review and quantitative analysis of the results published over the last decade, intending to further promote the development of various intelligent processing tasks for satellite video. It analyzes the current difficulties, challenges, and the methodological system for each task. In addition, it provides an in-depth analysis and summary of publicly available datasets and evaluation benchmarks for each task, as well as classic algorithm performance and application scenarios. Finally, this article summarizes the current research status and looks forward to the future development trend, hoping to inspire researchers in related fields and jointly promote the development of intelligent processing of satellite video.

Index Terms—Deep learning (DL), object detection, object segmentation, object tracking, scene classification, super-resolution, satellite video.

Manuscript received 5 May 2023; revised 25 June 2023; accepted 10 July 2023. Date of publication 18 July 2023; date of current version 28 July 2023. (Corresponding author: Shengyang Li.)

Shengyang Li, Manqi Zhao, Zhuang Zhou, Yuhan Sun, Han Wang, and Jian Yang are with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China, also with the Key Laboratory of Space Utilization, Chinese Academy of Sciences, Beijing 100094, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: shyli@csu.ac.cn; zhaomanqi19@csu.ac.cn; zhouzhuang@csu.ac.cn; sunyuhan21@csu.ac.cn; wanghan221@mails.ucas.ac.cn; yangjian20@csu.ac.cn).

Xian Sun is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, also with the Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: sunxian@aircas.ac.cn).

Yanfeng Gu is with the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: guyf@hit.edu.cn).

Yixuan Lv and Weilong Guo are with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China, and also with the Key Laboratory of Space Utilization, Chinese Academy of Sciences, Beijing 100094, China (e-mail: lvyixuan@csu.ac.cn; guoweilong19@csu.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3296451

I. INTRODUCTION

THE development of aerospace technology makes it possible for satellites to observe the earth by gaze. The emergence of video satellites such as Jilin-1 and the Sky-Sat series has gradually made video satellites an important means of earth observation, attracting widespread attention in various fields. Compared with traditional satellite remote sensing images, video satellite imaging can achieve a wide range of observations. More importantly, it can continuously achieve gaze imaging within the observation area and obtain earth observation dynamic information with high temporal resolution. Thus, video satellite imaging has essential applications in transportation, security, disaster monitoring, resources, and environment. Research on object detection, object tracking, object segmentation, scene classification, and other artificial-intelligence-based tasks using video satellite earth observation data has become a frontier hot spot in remote sensing [1], [2]. Most of the research topics in satellite video before 2013 revolved around satellite video coding, satellite video communication, and satellite video streaming. Subsequently, several earth observation video satellites and satellite constellations were launched. Since 2013, Planet Labs has successively launched the Skysat-1, Skysat-2, and Skysat-C video satellites. Skysat-1 is the first submeter video satellite with a spatial resolution of 1.1 m and a temporal resolution of 30 frame/s (FPS). It can capture high-quality black-and-white visible light images. Urthecast uses International Space Station (ISS) to embark on the world's first spaceflight full-color video camera Iris, with a spatial resolution of 1 m. Changguang Satellite Company also launched the Jilin-1 video 01–08 satellites between 2015 and 2018, the first spaceflight full-color video camera outside North America with a spatial resolution of 1.13 m. Zhuhai 01 and 02 video satellites were launched in 2017 and 2018, respectively. Qilu 04 video satellite was launched in 2021 with a spatial resolution better than 0.7 m. Wuhan University launched the LuoJia-3 video satellite in 2023, a satellite in-orbit real-time processing technology breakthrough.

With the successful launch and in-orbit operation of the above series of video satellites and satellite constellations, more and more high-temporal-resolution satellite video earth observation data can be obtained. The improved temporal resolution improves the timeliness of some traditional remote sensing applications, such as disaster monitoring, ocean monitoring, and ecosystem disturbance monitoring. It makes some applications like traffic condition monitoring a reality where traditional

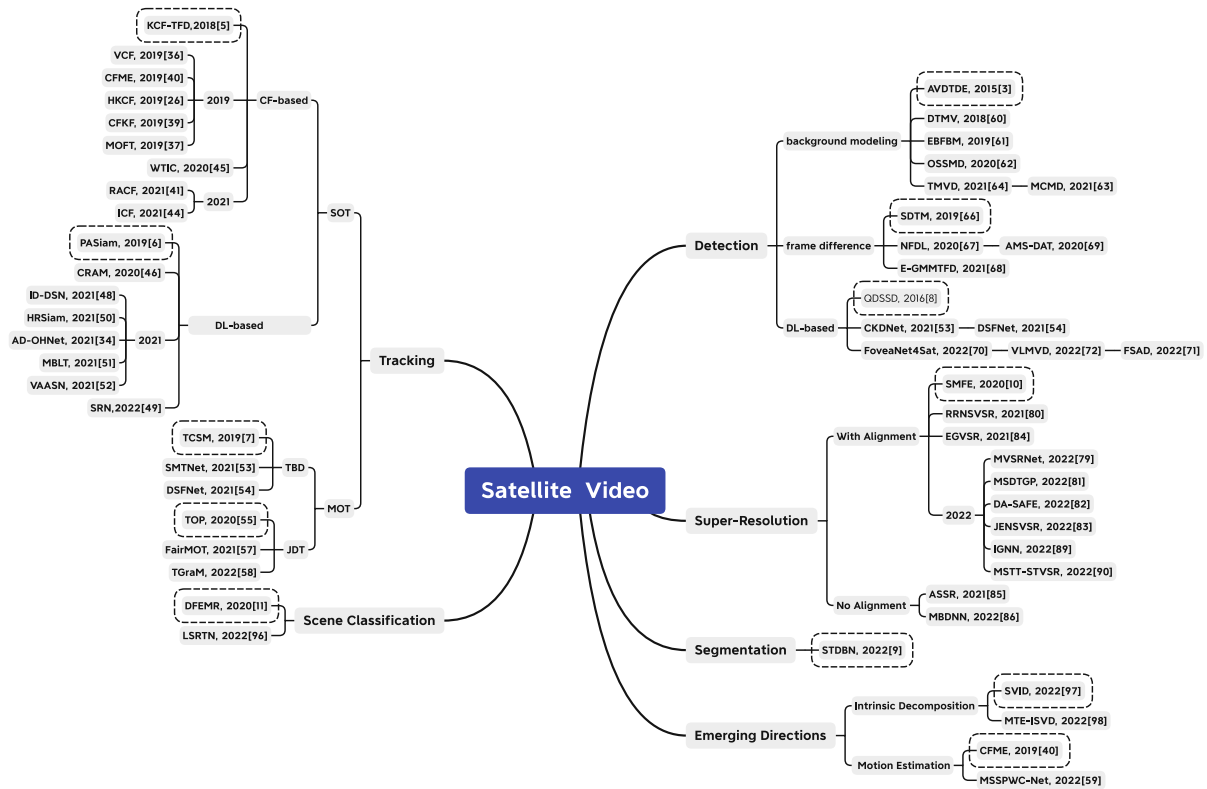


Fig. 1. Summary of the development of satellite video intelligent processing. The dotted boxes mark the jobs that appear for the first time in each direction.

remote sensing cannot perform well [3], [4]. These applications require the support of key technologies such as super-resolution reconstruction and moving object detection, identification, and tracking. Therefore, research into these technologies becomes essential. There is an urgent need for intelligent real-time monitoring of global hot spots and moving objects of interest, using the continuously dynamic information in large amounts of satellite video data.

With the significant increase in artificial intelligence computing power, deep learning (DL) has been rapidly applied in various fields such as computer vision, natural language processing, and satellite remote sensing image processing. The rapid increase in the number of satellite videos available has also made it possible to apply data-driven DL techniques to the intelligent processing of satellite video. In recent, DL algorithms have been rapidly developed for intelligent processing tasks such as object detection, object tracking and motion estimation, and super-resolution of satellite video. Meanwhile, many excellent works have emerged, attracting widespread attention from academia and industry. There are also many emerging research directions, such as satellite video scene classification (SVSC) and object segmentation.

As shown in Fig. 1, research on satellite video object tracking is divided into correlation-filtering-based and DL-based ways. Du et al. [5] first used a correlation filter (CF)-based approach to solve the single-object tracking (SOT) problem in 2018, Shao et al. [6] proposed a DL-based PASiam method for solving SOT tasks in 2019, Ao et al. [7] first proposed a network

named Tracking City-Scale Moving Vehicles From Continuously Moving Satellite (TCSM) in 2020 for solving multiobject tracking (MOT) tasks. For satellite video object detection tasks, there are three main methods: background modeling, interframe differencing, and DL-based methods. Kopsiaftis and Karantzaolos [3] first investigated satellite video object detection in 2015, using background modeling for vehicle detection and further achieving traffic density estimation. In 2018, Liu et al. [8] first applied a single-shot multibox detector (SSD) to effectively achieve satellite video aircraft detection. The spatiotemporal dual-branch network (STDBN), proposed by Zhong et al. [9] in 2022, performs effective segmentation for single aircraft and train in satellite video. Zhang et al. [10] conducted the first DL satellite video super-resolution (VSR) study based on Jinlin-1 data, and more works related to satellite VSR have sprung up by 2022. Gu et al. [11] presented the first SVSC work in 2020. Subsequently, a series of emerging directions, such as satellite video object segmentation (VOS), motion estimation, and intrinsic decomposition, have occurred by 2022. Overall, many aspects of satellite video intelligent processing research are still in their infancy, and there is still much room for exploration and development.

Several issues remain to be resolved in satellite video earth observation. On the one hand, video satellite imaging suffers from large variations in illumination, severe imbalances in foreground and background, significant differences in object scale, and insufficient spatial resolution due to its overhead imaging mode and detector performance. It is a major challenge

to design specific algorithms that are fully integrated with video satellite imaging mechanisms and characteristics to solve the common problems of low accuracy and poor robustness for multiple tasks such as object tracking, detection, super-resolution, and segmentation, as well as the individual issues for different tasks. On the other hand, temporal information is a unique characteristic of satellite video data. While there is large redundancy in satellite video data, how to fully use the temporal dynamic information and background invariance information in satellite video to optimize the model performance is also a major difficulty. To advance the development of intelligent satellite video processing, this article presents a review and multidimensional quantification of the current works on intelligent satellite video processing. In addition, this article also compiles and analyzes the evaluation results on public datasets, the pros and cons of methods, application scenarios, and future research directions. We hope that this article provides researchers in this field with a comprehensive review of the intelligent processing of satellite video. The work in this article can be summarized as follows.

- 1) This survey conducts a comprehensive review of the relevant works on the intelligent processing of satellite video. We perform multidimensional quantitative statistics to analyze the research hot spots and trends.
- 2) We summarize the difficulties and challenges currently faced in intelligent processing of satellite video and the methodological systems for different tasks such as object detection, object tracking, super-resolution, scene classification, and object segmentation.
- 3) This survey collates publicly available datasets, evaluation results, and analyzes benchmark methods' advantages and disadvantages for each satellite video intelligent processing task.
- 4) This survey analyzes the application scenarios and challenges of intelligent processing tasks for satellite video and looks at future research directions.

The rest of this article is organized as follows. Section II presents a statistical and quantitative analysis of the existing published literature and related research results in the field of satellite video to visualize the distribution and development trend of existing research work. Section III provides a detailed analysis of the difficulties and challenges in the field of satellite video. Section IV provides a detailed description of the methodology for specific tasks. Section V investigates the existing public datasets and corresponding experimental results in the field of satellite video. Sections VI and VII introduce typical application scenarios of satellite video and look into future research directions, respectively. Finally, Section VIII concludes this article.

II. QUANTITATIVE ANALYSIS OF ARTICLES

This section is mainly based on the Web of Science (WOS) and the China Knowledge Network (CNKI) to systematically analyze the research trends and hot spots in satellite video intelligence processing. WOS has more than 12 400 authoritative and high-impact international academic journals in three major

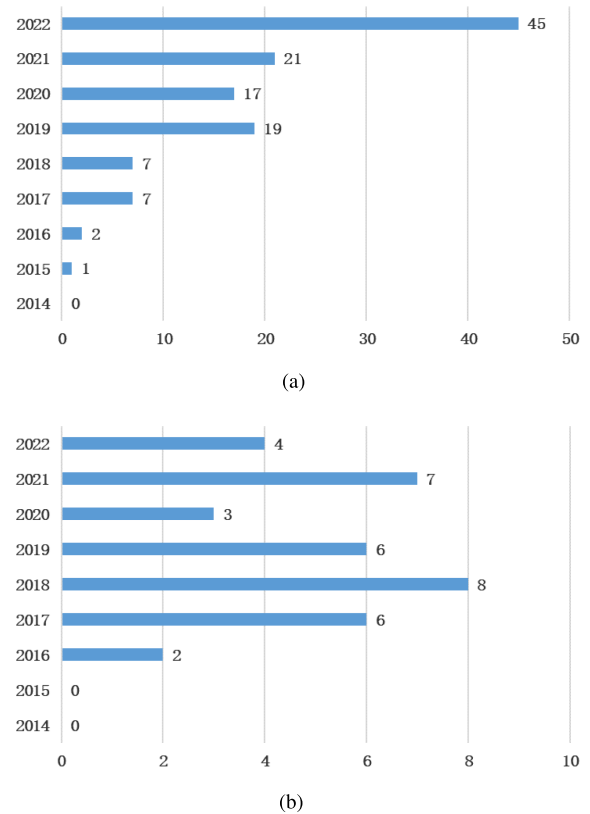


Fig. 2. Number of publications of earth observation satellite video from 2014 to 2022. (a) Data source from WOS. (b) Data source from CNKI.

citation systems (SCIE, SSCI, and A&HCI) covering natural sciences, engineering, social sciences, arts and humanities, and other disciplines. The retrieval condition is set to (title=satellite video) AND (duration=2014–2022), and finally, 119 valid articles on the intelligent processing of satellite video are obtained through manual selection. CNKI contains a database of Chinese journal articles, dissertations, and patents. The retrieval condition is set to (subject=satellite video) AND (duration=2014–2022), obtaining 36 valid articles through manual selection.

Fig. 2 shows a quantitative analysis of published articles. The number of articles shows a gradual increase in overall publications since 2015, with rapid growth in 2022, reaching 50 articles.

Then, we perform statistical analysis on articles published in different journals or conferences based on WOS and CNKI retrieval results. As shown in Table I, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Remote Sensing*, IEEE International Geoscience and Remote Sensing Symposium, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS are the four journals/conferences with the largest number of published articles. Among them, 26 papers are published in the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING AND REMOTE SENSING, accounting for 20.8%. The total number of articles in the four journals is nearly half of the available papers.

Moreover, based on WOS, this section presents a statistical analysis of several main existing research direction for intelligent processing of satellite video during

TABLE I
NUMBER OF ARTICLES IN MAINSTREAM JOURNALS/CONFERENCES

Journals/Conferences	Number	Percentage
IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING	16	10.32%
<i>Remote Sensing</i>	16	10.32%
IEEE International Geoscience and Remote Sensing Symposium	10	6.45%
IEEE GEOSCIENCE AND REMOTE SENSING LETTERS	9	5.81%
IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING	6	3.87%
<i>International Journal of Remote Sensing</i>	4	2.58%
<i>Journal of Applied Remote Sensing</i>	3	1.94%
<i>ISPRS Journal of Photogrammetry and Remote Sensing</i>	2	1.29%
<i>Remote Sensing of Environment</i>	2	1.29%
IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE	1	0.65%
<i>Neurocomputing</i>	1	0.65%
<i>Science China Information Sciences</i>	1	0.65%
International Conference on Pattern Recognition	1	0.65%

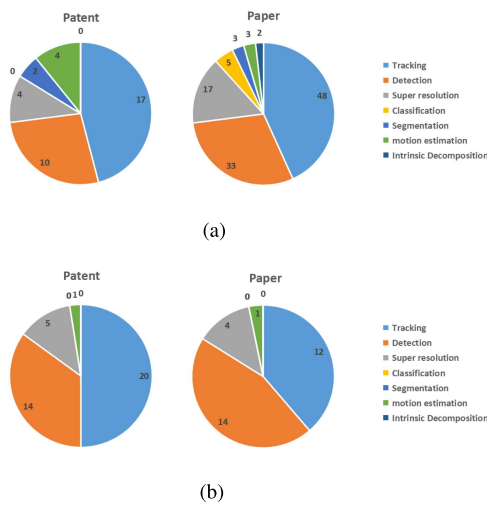


Fig. 3. Statistics on papers and patents in different fields. (a) Data source from WOS. (b) Data source from CNKI.

2014–2022. The retrieval condition is set to ((title=satellite video) AND (title=tracking)), ((title=satellite video) AND (title=detection)), ((title=satellite video) AND (title=segmentation)), ((title=satellite video) AND (title=scene classification)), and ((title=satellite video) AND (title=super resolution)). Fig. 3(a) shows the number of published papers and patents in different directions. Object tracking and object

detection have the highest number of relevant works, with 48 and 35 papers and 17 and 10 patents, respectively. The number of works on the remaining emerging directions is inadequate.

Similarly, based on CNKI, this section also presents a statistical analysis of several main existing research directions for intelligent processing of satellite video during 2014–2022 in Fig. 3(b). The retrieval condition is set to (subject=satellite video). Object tracking and object detection have the highest number of relevant works, with 12 and 14 papers and 20 and 14 patents, respectively.

In addition, based on WOS keyword trend and hot-spot analysis, Fig. 4 visualizes the distribution of research hot spots in the field of satellite video intelligent processing; object tracking is a major hot-spot direction, while DL and feature extraction derived from satellite video topics are also important research hot spots. Super-resolution, object detection, and vehicle detection are the next hot spots. Following closely behind, segmentation, classification, and motion estimation are gradually increasing in hotness. A number of methodological techniques derived from these directions, such as optical flow, Kalman filtering, correlation filtering, and video coding, have also drawn attention.

Fig. 5 shows the trend analysis of WOS-based keywords over the years, with the vertical axis representing the number of occurrences of the term in each year. Satellite video, object tracking, and DL gaining in popularity in 2022.

Finally, this section also researches existing reviews in the field of satellite video; setting the retrieval criteria (title=satellite

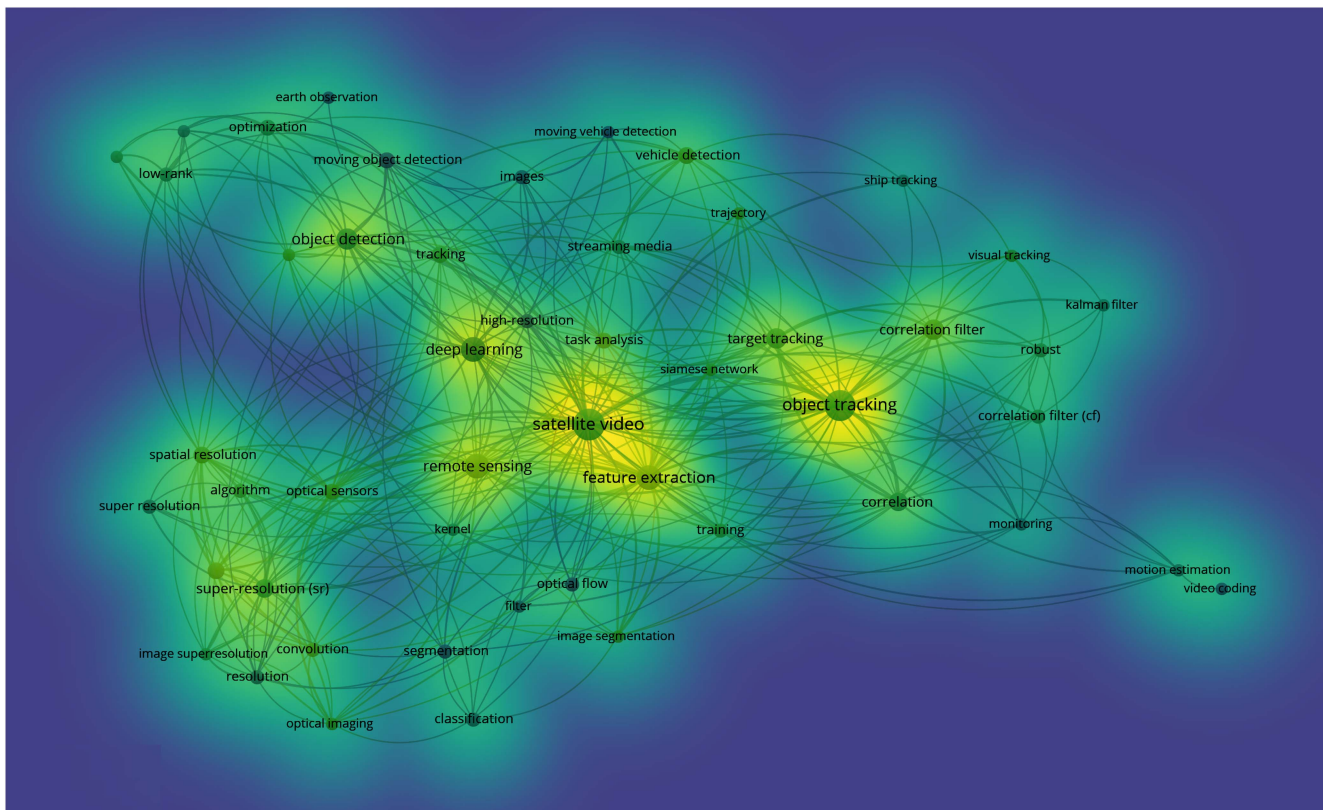


Fig. 4. Distribution of keyword hot spots in the field of intelligent processing of satellite video.

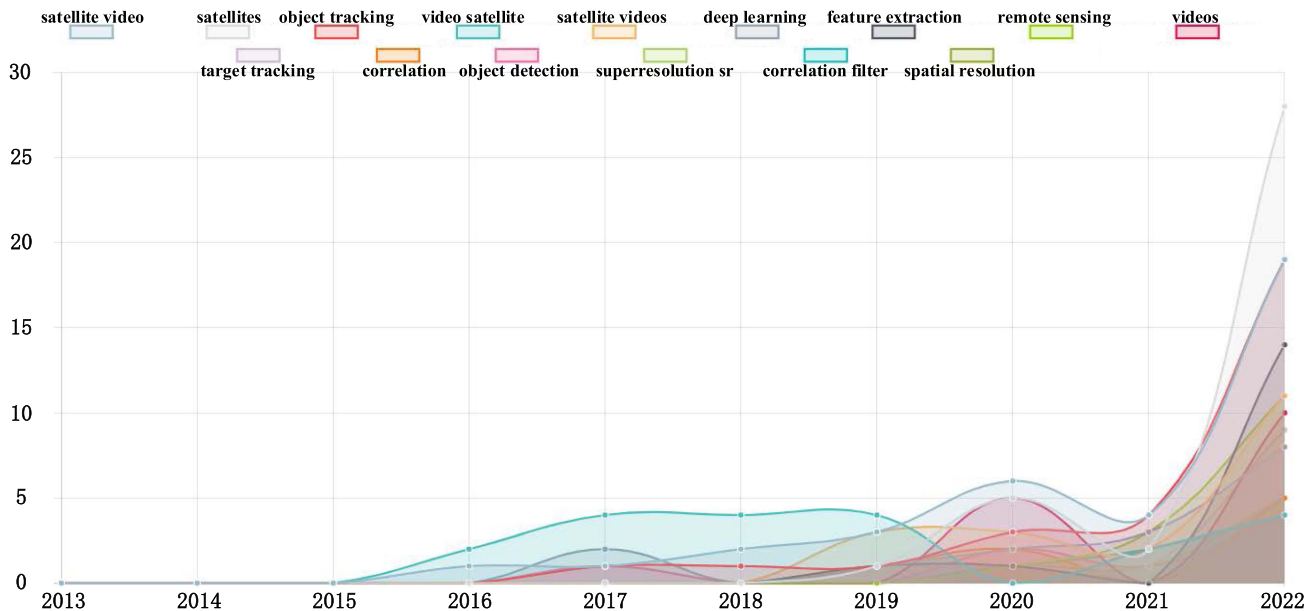


Fig. 5. Keyword trend analysis.

video AND title=survey) OR (title=satellite video AND title=benchmark) OR (title=satellite video AND title=dataset) OR (title=satellite video AND title=review) OR (title=satellite video AND title=research), five reviews are found (see Table II). Paper [12] mainly focuses on object tracking in

satellite videos. Paper [13] proposes a method for monitoring and analyzing urban traffic based on commercial video satellite and intelligent image processing technology and develops the calculation method of the traffic density, speed, and flow based on the video satellite data. Paper [14] briefly summarizes the first

TABLE II
STATISTICS OF AVAILABLE REVIEW ARTICLES

Name	Journal/Conference	Year
Deep-learning-based object tracking in satellite videos: A comprehensive survey with a new dataset [12]	<i>IEEE Geoscience and Remote Sensing Magazine</i>	2022
Research and application of urban traffic survey method based on commercial video satellite remote sensing technology [13]	<i>Resilience and Sustainable Transportation Systems</i>	2020
The 1st challenge on moving object detection and tracking in satellite videos: Methods and results [14]	International Conference on Pattern Recognition (ICPR)	2022
A summary of super-resolution for satellite videos via learning-based methods [15]	Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing(WHISPERS)	2019
Object tracking based on satellite videos: A literature review [16]	<i>Remote Sensing</i>	2022

challenge on moving object detection and tracking in satellite videos, and the top-performing methods and their results in each track are described with details. This challenge establishes a new benchmark for satellite video analysis on moving object detection and tracking in satellite videos. In order to investigate the adaptability to satellite video with low image quality, Liu and Gu [15] mainly focuses on some classical learning-based super-resolution methods, including sparse representation, collaborative representation, and DL methods. The survey [16] systematically investigates current satellite-video-based tracking approaches and benchmark datasets and summarizes the essential aspects of each tracking target (traffic target tracking, ship tracking, typhoon tracking, fire tracking, and ice motion tracking). It can be seen that each of the above reviews mainly addresses a single aspect of satellite video field. However, different from them, this article gives a comprehensive research, analysis, and summary for satellite video multitask intelligent processing and applications, including challenges, methods, and applications for different satellite video tasks.

According to the results of quantitative statistical analysis in this section, we can summarize some conclusions.

- 1) After the relevant video satellites were launched and satellite video data became available, researchers began to initially conduct research on the intelligent processing of satellite video data in 2015. The field has reached a development climax in 2022 in terms of speed and heat.
- 2) Object tracking, object detection, and super-resolution are the three directions that researchers have paid most attention to, and the number of articles related to the three directions is the largest.
- 3) First, the satellite video itself has the dynamic information of the object, which can better focus on the movement of the dynamic object. This enables the rapid development of satellite video object tracking technology. At present, the method represented by the combination of correlation filtering and DL network has attracted more attention and research. Second, the rich target information captured by satellite video has also drawn more attention to the research on satellite video object detection.
- 4) As satellite video gets more and more application and attention, some expanded research directions, such as object segmentation, scene classification, and motion estimation, have gradually attracted the attention and exploration of researchers. This enables the further development of

emerging application directions based on satellite video intelligent processing technology.

Overall, with the development of satellite video technology and satellite video constellation, the research represented by intelligent processing technology will receive more and more attention and research and will play an important role in the applications of traffic detection density estimation, scene monitoring, incident and disaster response, and land space use regulation.

III. DIFFICULTIES AND CHALLENGES

Satellite videos have evolved from single still images to multiframe continuous image sequences. Compared with general videos, satellite videos have the following characteristics, which pose greater difficulties for various processing tasks.

- 1) *Poor data continuity*: The current video satellites usually have continuous imaging times of the 90 and 120 s and are unable to observe the same area for a long time, which leads to poor data continuity. Although the video satellite constellation formed by SkySat, Jilin-1, or other video satellites shortens the reentry cycle, it is difficult to meet the demand for the real-time observation of specific objects [17].
- 2) *Spatial resolution needs to be further improved*: The spatial resolution of satellite videos is usually about 1 m, which is still lower than general videos and high-resolution aerial remote sensing images. Therefore, the typical remote sensing objects in satellite videos, such as vehicles, ships, and trains, have few pixels and small sizes, and the shape and textural features are not salient, resulting in low contrast and difficulty in distinguishing the foreground and background.
- 3) *Global motion due to platform movement*: The video is collected from the sensor on the video satellite. The platform is always in motion, and the imaging sensor needs to continuously adjust the shooting angle and pitch attitude along the direction of travel. The platform movement causes the satellite video background to move continuously and slowly, bringing the global motion of the video content.
- 4) *Large changes in illumination*: Video satellites target specific areas with dynamic imaging to capture dynamic changes in the ground but also introduce changes in

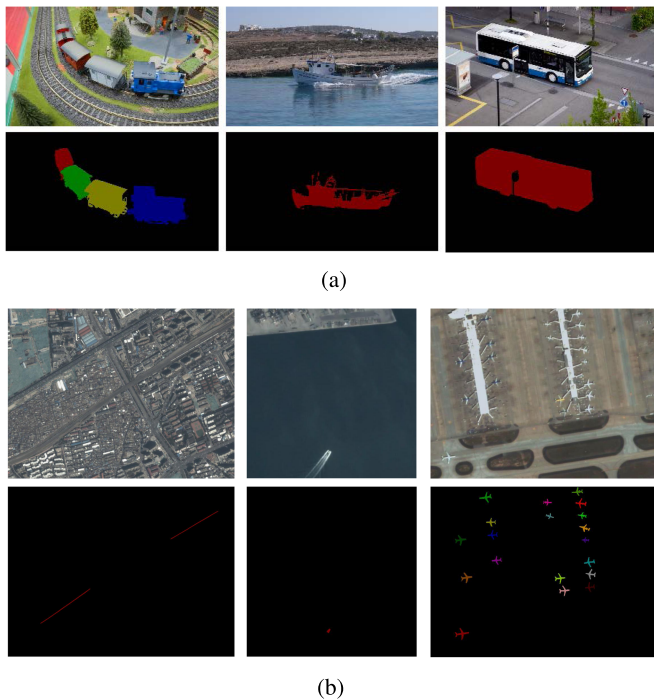


Fig. 6. Comparison of natural video and satellite video scenes. (a) Natural video scenes. (b) Satellite video scenes.

lighting. Illumination changes produce occlusion of objects on the surface and even lead to image distortion, seriously affecting video quality and content integrity.

- 5) *Large redundancy between video frames*: The frame rate of continuous imaging satellite video is usually less than 25 FPS, while due to the long imaging distance, satellite video content changes between adjacent frames are small, and object movement within the field of view is not apparent. There is a lot of redundancy in visual static information.

In terms of the imaging source, the altitude at which the satellite is located causes the scale of the obtained satellite videos to be very different from the natural scene videos. Therefore, the most notable difficulty in this field is that unlike targets in natural videos, which occupy a larger area, the targets (e.g., vehicles, ships, and aircraft) in satellite video often occupy only a few to a few dozen pixels, and they will receive much more interference. Besides, the complex background in the large scene of satellite video also brings more noise interference. Fig. 6 shows the comparison of natural video and satellite video scenes.

In general, the problems inherent in the field of satellite video are challenging, such as large scenes, insignificant features, small object areas, complex scenes, lighting variations, and redundant information between frames. These challenges have different impacts on different tasks.

Specific to each task, for satellite video object tracking, since the size of each object is too small compared to the whole image, and the object and background are very similar, tracking failure is very easy to occur. For satellite video object detection, the size of the moving objects of interest is often very small (e.g., most of the moving vehicles in the Jilin-1 satellite video are smaller than 20 pixels), resulting in a lack of texture and appearance

geometric information, and sometimes, there are motion artifacts caused by nongeostationary satellite imaging platforms. These issues make it difficult to achieve accurate positioning between consecutive frames. For satellite VOS, the main problems are the extreme imbalance in the front background due to small targets and blurred boundaries caused by low resolution and motion artifacts. These factors make high-precision segmentation difficult. For satellite VSR, the lower resolution of satellite video frames leads to the lack of sufficient texture and detail information, which makes feature extraction more difficult. Besides, the huge scene size also makes super-resolution reconstruction inefficient. Details are summarized as follows.

- 1) *Extreme foreground–background imbalance*: As shown in Fig. 6, satellite videos usually have large scenes, with typical moving objects such as vehicles and ships as small as less than 10 pixels, resulting in an extremely imbalanced distribution of positive and negative samples in the scene. Moreover, even with 1-m spatial resolution, small objects may lack shape, texture, and other features, posing significant challenges for algorithms.
- 2) *Complex background environment*: The imaging area of satellite videos is usually hundreds of times larger than that of general videos, leading to a large amount of redundant background information and various background interferences. These include situations where objects blend into the background, are difficult to distinguish, and where there are very similar interfering objects around the object, as well as sudden changes in lighting conditions and shadows.
- 3) *Severe occlusion*: Due to the complex traffic environment, objects such as cars face more severe occlusion problems, which means that algorithms are more likely to lose objects, especially for small and insignificant objects.
- 4) *Huge scene size*: Satellite videos typically have a large imaging width, which increases the computational burden on algorithms and requires longer processing times. Moreover, influenced by spatial resolution, the edge details of typical objects such as buildings, water bodies, and roads in large scenes are more blurred, posing enormous challenges for image quality restoration and feature extraction.
- 5) *Large differences in object scales*: Satellite videos have both spatial and temporal dimensions. For spatial scales, different objects have large differences in size, making it difficult for algorithms to coordinate different feature representations. In the temporal dimension, interframe motion blur and different object speeds make it difficult for algorithms to align on the time scale. Overall, the efficiency of spatiotemporal information fusion is also a challenge for algorithms.

IV. METHODOLOGICAL SYSTEM

A. Characteristics of Satellite Video Observation

Along with the rapid development of remote sensing technology, the ability to acquire earth observation data from space has increased. The imaging temporal resolution of land observation satellites has been decreased, but the revisit of single-satellite mode high-resolution satellites still takes two to five days. Even

TABLE III
INTRODUCTION OF EXISTING HIGH-RESOLUTION VIDEO SATELLITES

Country	Satellite Platform	Sensors	Launch time	Spatial Resolution (m)	Continuous observation time (s)	Covering range (km)
America	–	Skysat-1	2013.11.21	1.1	90	2×1.1
		Skysat-2	2014.07.08			
		Skysat-C	2016.06.22			
America	ISS	Iris	2014.01.27	1.1	90	5×3.4
		01, 02	2015.10.07	1.13		4.6×3.4
China	Jinlin-1	03	2017.01.10	0.92	60	11×4.5
		04, 05, 06	2017.11.21			19×4.5
		07, 08	2018.01.19			
China	Zhuhai-1	OVS-1	2017.06.15	1.98	90	8×6
		A/B				
		OVS-2	2018.04.26	0.9	120	3.6×2.7
China	Tiantuo-2	–	2014.09.08	5	180	–
U.K.	Vivid-i	VividX2	2018.01.12	1	–	5.2×5.2
U.K.	Carbonite-2	VividX2	2018.12.01	1	–	5.2×5.2
China	Foshan-1	–	2021.04.27	0.5-0.7	–	–
China	Qilu-4	–	2021.04.27	0.5-0.7	–	–
China	Luojia-3	01	2023.01.15	0.7	–	5.2×5.2

TABLE IV
INTRODUCTION TO PUBLICLY AVAILABLE DATASETS

Name	Source	Video Number	Size	Category Number	Task	Year	Unit
SatSOT [99]	Jinlin-1	105	12 000×5000	4	Tracking	2022	Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences
VISO [18]	Jinlin-1	47	12 000×5000	4	Detection Single/Multiobject Tracking	2021	National University of Defense Technology
AIR-MOT [58]	Jinlin-1	149	1920×1080	2	Multiobject Tracking	2022	Aerospace Information Research Institute, Chinese Academy of Sciences
Jinlin-189 [89]	Jinlin-1	199	640×640	2	Super-Resolution	2022	Wuhan University
SAT-MTB [100]	Jinlin-1	249	640×640	12	Detection Segmentation Single/Multiobject Tracking	2023	The Space Applications Center of CAS

though the constellation formed by light and small satellites has shortened the reentry cycle, it is still difficult to meet the demand for continuous and long-time observation of specific ground objects for the time being. Satellite video remote sensing earth observation is a new type of remote sensing technology that has been developed in the last decade. The major difference between video imaging satellites and traditional optical earth remote sensing satellites is that the former can continuously observe a certain area and obtain more information about the continuous movement of the object, such as the movement speed and direction of the object. Moreover, video satellites are in an almost gazing manner, which is particularly suitable for the perception of moving objects, thus obtaining dynamic information with a high temporal resolution. Dynamic information is difficult to obtain with conventional ground-based optical remote sensing satellites.

Several video satellites have been launched in recent years. Planet Labs of the United States first launched Skysat-1 in 2013, with a spatial resolution of 1.1 m and an imaging range of $2 \text{ km} \times 1.1 \text{ km}$. Changguang Satellite Company of China launched the first color video satellite Jilin-1 in 2015, with a video resolution of about 1 m and an imaging range of $4.6 \text{ km} \times 3.4 \text{ km}$. Qilu-4 and Luojia-3 video satellites were both launched in January 2023, with a spatial resolution of 0.5–0.7 m. Details of existing video satellites are shown in Table III.

Video images of natural scenes are usually not conducive to studying large-scale moving objects due to the small shooting range and the little information obtained, which is an advantage of satellite video. However, as shown in Fig. 6, compared to the natural scene video captured by the camera, the satellite video

has difficulties such as a low percentage of object foreground, weak and insignificant object features, complex background, blurred image, and low frame rate due to its unique imaging mechanism of a long-range overhead view. Thus, generic intelligent video processing algorithms for natural scenes cannot be directly applied to satellite-video-related tasks. It is necessary to consider the unique characteristics of satellite video for targeted algorithm innovation and improvement.

B. Satellite Video Object Tracking and Motion Estimation

1) *Satellite Video Object Tracking*: Satellite video object tracking aims to track moving objects of interest in satellite videos, such as airplanes, ships, vehicles, and trains, and automatically estimate their states, such as position and size, in the video. Depending on the number of tracking objects, it can be generally divided into two tasks: SOT and MOT, as shown in Fig. 7. Given the state of the object to be tracked in the first frame, SOT algorithms need to locate the object frame by frame in the video and provide the position and bounding box of the object [12]. On the other hand, MOT simultaneously tracks multiple objects of interest in the video of a specified category. It distinguishes them by different labels and temporally associates the objects between frames [18].

a) *Single-object tracking*: Researchers have recently proposed several algorithms for SOT in satellite videos, mainly including generative and discriminative methods. Generative methods extract object features for modeling and find similar objects frame by frame, including mean shift [19], particle filtering [20], Kalman filtering [21], sliding window search [22], and

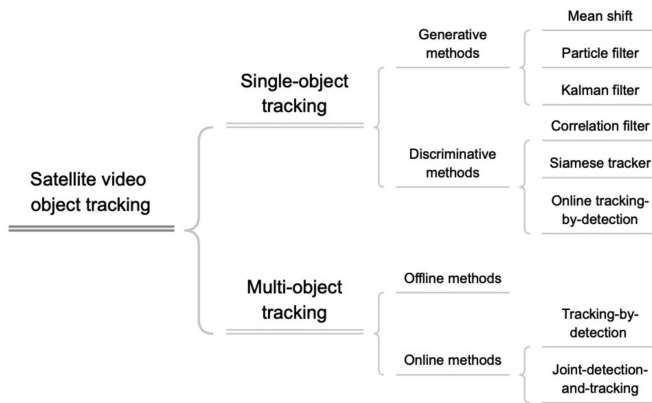


Fig. 7. Overview of satellite video object tracking methods.

other methods. Generative methods ignore background information, and their accuracy significantly decreases when objects undergo significant deformation, similar objects in the background, or shaking in the video. Compared with generative methods, discriminative methods have higher accuracy and faster speed. Typically, an object classifier is trained to classify objects as foreground and enable tracking during the tracking process. Discriminative correlation filtering methods [23], [24], [25], [26] are the most representative. With the development of DL, some DL-based trackers with stronger feature representations have been proposed, including Siamese-based trackers [27], [28], [29], deep discriminative CFs [30], [31], [32], online detection-based trackers [33], reinforcement learning-based trackers [34], etc.

The correlation filtering methods are popular due to their efficiency and accuracy. Some researchers have combined object detection algorithms to improve the performance of satellite video trackers. Du et al. [5] proposed a satellite video tracker by combining the three-frame difference algorithm with the CF tracker. Ahmadi and Mohammadzadeh [35] proposed a method for detecting and tracking vehicles and ships in satellite videos based on background subtraction technology. Some algorithms track objects by extracting motion information. Shao et al. [36] designed a velocity-related filtering algorithm that utilizes velocity features obtained through optical flow and inertial mechanisms. Du et al. [37] constructed a multiframe optical flow tracker that combines optical flow and multiframe difference methods for object tracking in satellite videos. Chen and Sui [38] proposed a spatial mask to promote CF to give different contributions based on spatial distance and then applied a Kalman filter (KF) to predict the position of objects in large and similar background regions. Later, Guo et al. [39] introduced the global motion characteristics of moving vehicles to constrain the tracking process and corrected the trajectories of moving objects by integrating their positions and velocities. Xuan et al. [40] proposed a motion estimation algorithm that combines a KF and a motion trajectory averaging strategy to address occlusion problems in satellite videos. Other methods track objects using their features. Xuan et al. [41] proposed a rotation adaptive CF tracking algorithm to solve the rotation problem of objects in satellite videos. The proposed method maintains the stability of the feature map for object rotation and achieves the ability to estimate changes in bounding box size. Chen et al. [42] decoupled

rotation and translation motion patterns and developed a new rotation adaptive tracker with motion constraints. In addition, Pei and Lu [43] designed a kernel correlation filter (KCF) based on color name features and Kalman prediction. Liu et al. [44] fused different features of the object based on the KCF and introduced KF to compensate for motion position deviation. Wang et al. [45] focused on sample training strategies and sample representation capabilities to enhance object tracking in satellite videos. They established a filtering training mechanism for objects and backgrounds to improve the discriminative ability of tracking algorithms and constructed an object feature model using Gabor filters to enhance the contrast between objects and backgrounds.

With the development of DL and neural networks, some researchers have used deep neural networks to enhance the feature modeling process of trackers. Hu et al. [46] constructed a convolutional regression network for satellite video object tracking that uses a pretrained deep neural network to extract appearance and motion features. Uzkent et al. [47] utilized a convolutional neural network to extract hyperspectral domain features and used the KCF to handle satellite video tracking problems. Due to the significant efficiency advantage of Siamese networks' weight-sharing structure, some algorithms have built Siamese network tracking frameworks. Shao et al. [6] proposed a fully convolutional Siamese (Siamese-FC) network with shallow features to extract fine-grained appearance features for satellite video tracking. The network incorporates a Gaussian mixture model (GMM) and utilizes Kalman filtering to handle tracking occlusion and motion blur issues. Similarly, Zhu et al. [48] proposed a deep Siamese network (DSN) with an interframe difference centroid inertia motion model to alleviate model drift and used a Siamese region proposal network to obtain object location. In addition, Ruan et al. [49] proposed a two-stream Siamese convolutional neural network that combines Siamese networks and motion regression networks to achieve satellite object tracking and further alleviate model drift by using the trajectory fitting motion model based on historical trajectories. Shao et al. [50] designed a high-spatial-resolution lightweight parallel network and proposed a pixel-level refinement model based on online moving object detection and adaptive fusion to enhance the tracking robustness in satellite videos. Zhang et al. [51] learned the motion and background of the object to help the tracker identify the object more accurately. They predict the probability of the object position in each pixel of the next frame using a fully convolutional network and introduce a segmentation method to assign high probabilities to feasible regions of the object in each frame. Bi et al. [52] proposed a satellite video object tracking algorithm based on a variable-angle adaptive Siamese network (VAASN). This method utilizes a multifrequency feature representation method in the feature extraction phase of a Siamese-FC network to reduce the impact of complex backgrounds. It introduces a variable-angle adaptive module to adapt to changes in object rotation during the tracking phase.

b) Multiobject tracking: Compared to SOT, MOT in satellite videos is still in its early stages of research. The methods can be divided into two main trends: detection-based tracking (TBD) methods and joint detection and tracking (JDT) methods. TBD methods treat detection and tracking as two independent

tasks and use external detectors to generate frame-by-frame detection results, followed by applying additional models for the interframe association. JDT methods also design models to perform detection and association simultaneously for more efficient tracking.

In the TBD framework, researchers usually utilize object detectors to discover and detect potential objects in the scene and then perform interframe associations to obtain tracking trajectories. Some studies focus on the research of moving object detection. Ao et al. [7] provided a vehicle detection algorithm based on local noise modeling, which distinguishes potential vehicles from noise patterns using an exponential probability distribution. Feng et al. [53] performed cross-frame keypoint detection through interframe information and constructed a two-branch structure with long short-term memory (LSTM) to effectively detect and track dense vehicles. Xiao et al. [54] proposed a dynamic and static fusion dual-stream network (DSFNet) to detect moving objects in satellite videos by extracting static context information from a single frame and dynamic motion clues from continuous frames.

In the JDT framework, algorithms perform JDT, combining object detection with the temporal association. Zhou et al. [55] proposed a synchronous detection and tracking algorithm that applies keypoint detection models to image sequences and the previous frame's detection results, locating different objects by associating keypoints to complete tracking. Wang et al. [56] and Zhang et al. [57] extracted detection features and identity switch (ID) features simultaneously using a shared network and associated the predicted IDs to complete tracking. He et al. [58] modeled MOT as a graphical information reasoning process from the perspective of multitask learning and proposed a graph-based spatiotemporal reasoning module to explore potential high-order correlations between video frames. These single-stage methods save much inference time but are difficult to detect and associate objects that lack appearance information.

2) *Satellite Video Motion Estimation*: Satellite video motion estimation can give support to object tracking. However, the background of satellite video scenes is complex and noisy, and traditional methods cannot extract dense motions. In addition, traditional methods are always time consuming in computing motions, and it is also difficult to directly apply DL methods. Appropriate features can solve the problem of complex backgrounds, but they are powerless for small objects and noise. Satellite video scenes, especially urban scenes, contain a large number of small and fuzzy objects, and labeling the ground truth of these object motions is challenging.

In summary, two challenges exist in extracting dynamic information in satellite video scenes: 1) how to extract the motion of unlabeled small blurred objects and 2) how to extract the accurate motion of blurred objects from the noisy background.

Xuan et al. [40] proposed the first novel motion estimation algorithm by combining Kalman filtering and motion trajectory averaging. Based on the assumption that the motion of the object is a uniform linear motion in a relatively short time (even if the object is in a turn, sharp stop or acceleration, etc.), the motion trajectory averaging method is used to calculate the motion state of the object before the KF converges. The average of

the displacement of the previous frames is used to estimate the object's velocity in the current frame, and the velocity of the object and position in the previous frame are used to estimate the position of the object in the current frame. After the KF converges, the result of the KF is used as the output of the motion estimation. Wang et al. [59] proposed MSSPWC-Net, which consists of a sparse self-learning network, PWC-Net, and a multiframe framework that uses a sparse warping loss function to improve the sensitivity of small objects to self-learning methods. Satellite video objects are sparse concerning the background, and motion consistency constraints can be used to solve the fuzzy object motion problem. With a multiframe framework, the motions of adjacent frames are successfully fused to estimate the accurate motion of the fuzzy objects. However, MSSPWC-Net can only perform motion estimation based on depth features, so the network must be trained to fine-tune the features to obtain accurate results. In subsequent studies, sparse prior constraints can be used to improve the segmentation results or increase the cost volume to obtain more accurate information about small objects.

C. Satellite Video Object Detection

Compared to the task of image-based object detection, the most significant benefit of video object detection is the inclusion of temporal contextual information, where each frame has a temporal contextual association, correspondence, and similarity. Since there is a subcontextual relationship, the detection results of the adjacent frames can be used to improve the detection accuracy of the current frame. Since the adjacent frames have similar continuity, the redundant information can be used to speed up the detection of each frame. Compared to mainstream object detection based on high-resolution remote sensing imagery, the challenges of satellite video object detection are mainly reflected in the object characteristics and data quality issues such as small object size, low contrast, and poor clarity of video frames.

1) *Satellite Video Object Detection Based on Conventional Methods*: Conventional methods perform object detection by capturing the change region in the sequence image of the satellite video and extracting the moving object from the background. The main techniques include background modeling methods and interframe differencing methods.

Several background-based modeling approaches were proposed [60], [61], [62], [63], [64]. Ao et al. [60] proposed a detection algorithm based on local noise modeling to correct the detection results by distinguishing the latent probability distribution of the vehicle. Lei et al. [64] proposed a satellite video vehicle detection method based on spatiotemporal information, which combines the vital interframe temporal information to optimize the detection. Zhang et al. [61], [62], [63] proposed a set of detection methods based on low-rank structured sparse decomposition for satellite video moving vehicle detection.

Some interframe differencing methods were also designed for satellite video object detection [65], [66], [67], [68], [69]. Zhang et al. [65] segmented the image based on local variable thresholding and combined the correlation between multiframe object motion and satellite pose motion information to detect

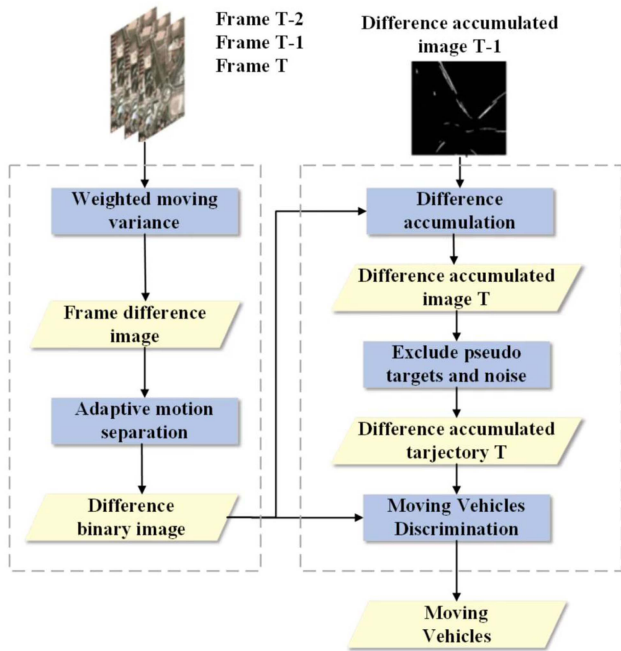


Fig. 8. Technical flow of the adaptive motion separation method for satellite video detection [69].

the object. Li et al. [66] proposed an automatic detection and tracking method for moving ships of various sizes in satellite videos. Shi et al. [67] developed a normalized frame difference tagging method to enable stable satellite video moving aircraft detection. Shu et al. [68] reduced the false detection of vehicles due to illumination changes and background movement by fusing a GMM with three frames of differential detection results. Chen et al. [69] proposed an adaptive motion separation method for vehicle detection by accumulating object trajectories to help separate moving objects from the background. The technical flow of the method is shown in Fig. 8. Conventional satellite video object detection methods do not rely on the object's annotation information to train the model but only on its motion changes. Thus, they belong to the weakly supervised learning type. The conventional methods can only detect moving objects in the satellite video and are unable to distinguish the category of the object, so most methods are currently used to detect moving vehicles. In addition, currently published papers are validated on small or nonpublic datasets, and there is a lack of benchmark evaluation to measure the performance and robustness of the methods.

2) *DL-Based Satellite Video Object Detection*: Due to the lack of large-scale publicly available annotated datasets for object detection, DL-based methods for satellite video are still in their infancy, with less published research than conventional methods in general. Feng et al. [53] proposed a detection and tracking framework for moving vehicles in satellite video, consisting of a cross-frame keypoint-based detection network (CKDNet) and a spatial-motion-information-guided tracking network (SMTNet). Among them, a cross-frame module was

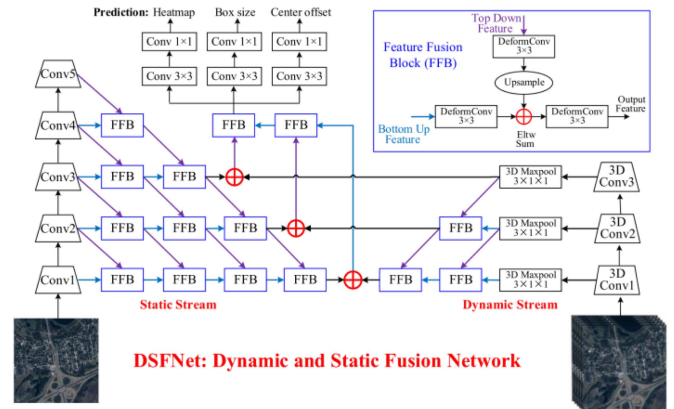


Fig. 9. Satellite video object detection with DSFNet [54].

designed to support keypoint detection, which effectively exploited the interframe complementary information in the detection network CKDNet. It optimized the results by combining size prediction around keypoints and defining invalid match suppression for oversized keypoint pairs. Liu et al. [8] proposed the quality deconvolutional single-shot detector (QDSSD) based on the SSD network for the problem of small-size objects, which enriched the feature information by deconvolution and significantly improved the aircraft detection results, especially the small-size aircraft objects close to each other. Xiao et al. [54] proposed a DSFNet with the network structure shown in Fig. 9, where a 2-D backbone is used to extract static contextual information in each frame, and a 3-D backbone extracts successive dynamic motion cues of the video. Moving vehicle detection in satellite video was efficiently performed by fusing static and dynamic features. Pflugfelder et al. [70] proposed a DL-based satellite video vehicle detection method, using a tight convolutional kernel to extract spatiotemporal feature information, ignored maximum pooling, and uses weak RLUs to improve vehicle detection. Zhou et al. [71] proposed a detection method with feature scale selection and contrastive proposal encoding. By leveraging external remote sensing image datasets to accomplish the network pretraining, the aircraft detection can be achieved by relying on only a small number of satellite video annotated samples. To address the problem of unremarkable vehicle appearance information, Pi et al. [72] designed a feature interframe differential module to obtain neighboring motion information, extracted semantic features, and further introduced Transformer to refine the semantic features to achieve effective vehicle detection.

Unlike conventional methods, which can only detect moving objects in satellite video and cannot distinguish between categories, DL-based methods rely on annotated data to train models that can learn distinguishable features of objects in a supervised learning manner. With the continuous development and expansion of datasets in the field, how to design algorithmic networks for specific characteristics of satellite video is the future research of satellite video object detection.

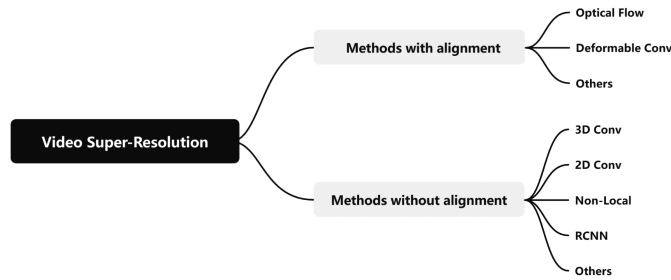


Fig. 10. Overview of satellite VSR methods.

D. Satellite VSR

The task of VSR is an extension of image super-resolution, which aims to reconstruct a high-resolution video from a low-resolution video. VSR has a significant practical value as it can improve the performance of high-level tasks such as object detection, object tracking, and object segmentation. Moreover, VSR can also be used for data compression. However, VSR is more challenging than single-frame image super-resolution (SISR) due to the extra temporal dimension. This dimension makes it difficult for image super-resolution methods to achieve satisfactory results on video. The high-resolution videos produced using these methods often suffer from artifacts that cause video incoherence [73]. The overview of existing satellite VSR methods is shown in Fig. 10

Despite these challenges, video possesses richer information than the image, and exploiting this redundant information can lead to higher upper limits in VSR. To better use interframe information, scholars often include the step of alignment in their methods and expand the length of the input sequence [74], [75]. These steps are not available in image super-resolution. DL has become a popular approach for satellite VSR methods in recent years and has shown excellent performance.

Earlier studies directly applied image super-segmentation methods to satellite video [76], [77], while methods designed specifically for video have only emerged recently. These methods can be broadly divided into alignment and no-alignment methods, with methods with alignment being dominant. Methods with alignment typically include four basic parts: propagation, alignment, aggregation, and upsampling. Alignment is crucial in VSR, and the absence of proper alignment can significantly degrade the results [74]. Alignment can be achieved through image or feature alignment, and the primary means of achieving alignment include optical flow and deformable convolution.

Zhang et al. [10] were among the first to use satellite video interframe information for super-resolution. They employed a combined single-frame and multiframe network. The multiframe network was derived from the classical generic VSR network EDVR [78] and used deformable convolution for feature alignment. In contrast, the approach of He et al. [79] employed optical flow estimation for alignment. Specifically, their method upsamples the images and then passes them through an attention-based residual network to obtain the final high-resolution image. Xiao et al. [80] proposed a recurrent refinement network that

aligns the reference images by the optical flow method and extracts information from them to add to the SISR of the object frame. Another approach by this author, MSTDGP [81], proposed a novel temporal grouping projection fusion strategy and a DCN-based multiscale residual alignment module. Ni et al. [82] also used DCN for alignment and proposed a scale-adaptive feature extraction module, as well as an upsampling module that allows arbitrary magnification. The method proposed by Liu and Gu [83] consists of two subnetworks, one branch predicting high-resolution images and one branch predicting fuzzy kernels, coupled by a cross-task feature fusion module, whose alignment is based on patch matching in the feature space and is more stable than using optical flow. The method proposed by Shen et al. [84] also utilizes dual branches, which adds an edge branch to EDVR that can simultaneously predict high-resolution edge maps and fuses features from both branches at the end of the network.

He et al. [85], [86] also proposed a no-alignment approach to feature extraction and fusion using 3-D convolution directly. He and He [85] proposed a network with arbitrary image magnification implemented by subpixel convolution and Bicubic for upsampling. In [86], they split the objective function of the degraded model into two suboptimization problems. For the first time, they proposed a fusion of DL and model-based methods for the super-resolution of satellite video.

Furthermore, the method in [87] utilized unsupervised learning, consisting of a downsampling network and an upsampling network that did not require low-resolution high-resolution training pairs. This satellite VSR method [88] focused on modeling and super-resolution of aircraft in videos. Graph neural networks have also been applied to the super-resolution of satellite videos [89], and another work by the same authors implemented super-resolution of both time and space in a single network, predicting unknown frames by coupling optical flow and multiscale deformable convolution [90].

E. Satellite VOS

The unique temporal information of satellite videos makes them more suitable for practical applications. Video object and instance segmentation allow further refinement of object processing and analysis, so studying satellite video object and instance segmentation has important application value and significance. However, typical objects in satellite video, such as aircraft, vehicles, and trains, have small sizes and blurred appearance features. VOS requires pixel-level annotation, which is difficult and costly. Besides, satellite videos are expensive to acquire, and it is difficult to collect sufficient samples to support DL model training. Therefore, the current related research has been conducted in a small number of satellite video sequences for training and evaluation [9]. The lack of open-sourced large-scale satellite VOS dataset has seriously restricted the development of satellite VOS.

The existing satellite VOS algorithms follow the definition of semisupervised VOS, where the ground truth of a particular object in the first frame is given in the test phase. The goal is to segment the corresponding object in the whole video [91]. Zhong et al. [9] collected 17 satellite videos from SkySat,

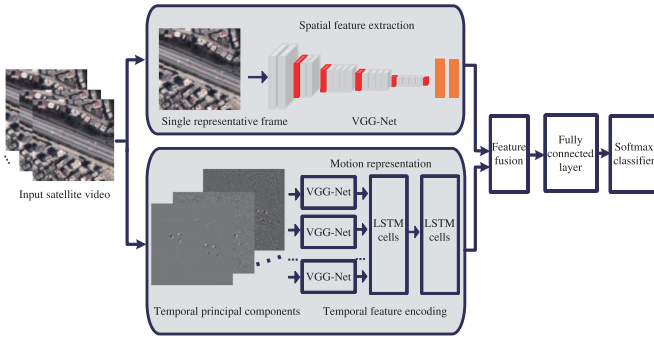


Fig. 11. Framework of the two-stream structure for SVSC [11].

UtherCast, and Jinlin-1 and constructed the DAVOS dataset. They designed spatiotemporal dual-stream branches to learn the spatiotemporal features of the object of interest in satellite videos. They utilized the online learning method One-Shot Video Object Segmentation (OSVOS) [92] for training. The temporal consistency branch was pretrained in the ImageNet Large Scale Visual Recognition Challenge 2015 object detection from video dataset [93], and the spatial segmentation branch was pretrained in the PASCAL VOC 2012 segmentation dataset [94]. Then, the model is trained on the DAVOS dataset and then fine-tuned in the test phase based on the annotation mask in the first frame of the video, ultimately achieving significant region similarity and contour accuracy on aircraft and trains.

VOS of general domains is mainly divided into semisupervised VOS, interactive VOS, and unsupervised VOS. The various relevant tasks will provide new ideas for satellite VOS. Meanwhile, the video instance segmentation task segments all the objects of interest in each video frame and associates inter-frame object ID [95], which will expand the application scenario and practical value of satellite VOS. The existing satellite VOS algorithms are limited by spatial resolution and lead to low contour accuracy. The satellite VSR reconstruction will optimize image quality and improve contour accuracy.

F. Satellite Video Scene Classification

SVSC plays a vital role in the intelligent interpretation of satellite videos, which describes the semantic information of ground contents in satellite video. Different from the remote sensing image scene classification task, SVSC aims to describe both static and dynamic semantic information of ground objects. It can generate an overall description of the local ground area within a certain time. In essence, it is similar to the video classification task in general video understanding and is a future research topic in satellite video intelligence understanding. Existing studies rely on DL techniques and focus on the joint representation of spatial and temporal features in satellite videos to improve classification accuracy. They are mainly based on the two-stream framework. In 2020, Gu et al. [11] first proposed an SVSC method based on the two-stream framework to jointly represent the spatial and temporal features of satellite videos, as shown in Fig. 11. It consists of two stages: the keyframe selection

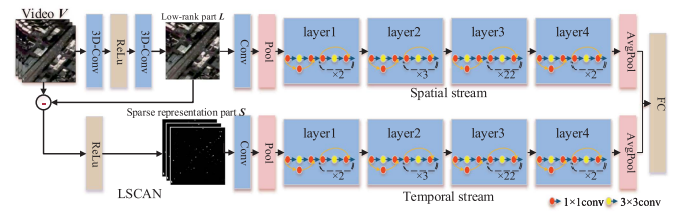


Fig. 12. Mainframe of the LSRTN [96].

and long-term sequence feature encoding. The keyframe is selected based on fuzzy detection and the activity of ground objects in satellite video scenes. Its feature extracted by the pretrained VGGNet is treated as the spatial feature of the satellite video. At the same time, an LSTM network is used to encode frame features extracted by PCA and VGGNet, which is treated as the video-level feature representation of the given satellite video. The proposed method achieves 73.97% overall accuracy (OA) on the proposed SVSC dataset, including 8 static scenes and 7209 videos, from the Jilin-1 video satellite.

To effectively represent the features of small moving objects in satellite videos, Wang et al. [96] proposed a low-rank sparse representation two-stream network (LSRTN) for satellite video single-label scene classification, which consists of two parts: low-rank sparse decomposition and the spatial and temporal features representation in Fig. 12. A low-rank sparse component analysis network (LSCAN) was designed to decompose satellite videos into low-rank background images and sparse moving object sequences. Then, a two-stream structure was applied to obtain the spatial and temporal features based on original frame images and the sparse moving object sequence images, which was used for classification after feature fusion. The LSRTN achieves 81.2% OA on the constructed dataset, demonstrating its effectiveness in representing the features of small moving objects in satellite video scenes.

G. Emerging Directions

In addition to intelligent processing of regions and objects of interest for earth observation by satellite video, researchers have also conducted research around the temporal characteristics of satellite video itself, in which satellite video intrinsic decomposition (SVID), as an auxiliary and enhancement type method for precise object location and identification, provides a new research direction for enhancing the extraction of static and dynamic components by networks.

Establishing SVID can eliminate the effect of light interference on the reflectance component because the light is mainly concentrated on the shadow component rather than the reflectance component. SVID will help to build video algorithms with light interference suppression and improve the effectiveness of the related algorithms. SVID will also help to analyze and extract the static components of satellite videos.

Gao et al. [97] proposed the first SVID algorithm to extract reflectance and shadow information from satellite video scenes, including stable static and sparse dynamic components, respectively. First, the satellite video information is divided into four

components: the intrinsic reflectance image of the static scene, the sparse dynamic reflectance video, the shadow image of the static scene, and the sparse dynamic shadow video. Second, based on the above signal composition, the satellite video is decomposed into intrinsic information for the invariance of the scene and background. Although the algorithm can achieve the intrinsic decomposition of satellite video, it cannot yet achieve real-time processing for remote sensing satellite video with large scenes and cannot extract the intrinsic information of continuous shadow regions; and for some smaller dark objects, the algorithm has limited improvement for the subsequent tracking steps; meanwhile, the algorithm does not have good theoretical processing capability for videos with severe platform vibration. Pan et al. [98] proposed a satellite video intrinsic decomposition model MTE-ISVD with moving object energy constraint to maintain the temporal coherence of reflectivity and improve the performance of moving objects. MTE-ISVD has four reasonable constraints: retinex local constraint, absolute scale constraint, reflectivity time constraint, and moving object energy constraint. Eventually, the SVID becomes a closed-form solution, and its computational speed is relatively improved. However, the experimental results of MTE-ISVD are very dependent on the parameter settings, and the real-time processing of large scenes is still difficult to achieve. According to the retinex theory, the illumination requirements are uniform and slowly changing, and MTE-ISVD has limited improvement on the highlight or shadow regions.

V. PUBLIC DATASETS AND EXPERIMENTAL RESULTS

A. Introduction to the Dataset

The field of satellite video is still in the development stage, and the available public datasets are relatively small and not comprehensive. Introduction to publicly available datasets is shown in Table IV. There are four publicly available datasets, all from Jilin-1, for detection, tracking, and super-resolution missions, respectively. The VISO dataset, proposed by the National University of Defence Technology in 2021, has four categories: aircraft, vehicles, ships, and trains. VISO is derived from 47 video segments, mainly for detection, SOT, and MOT tasks (<https://satvideodt.github.io/>). The SatSOT dataset was proposed in 2022 by the Space Applications Engineering and Technology Centre of the Chinese Academy of Sciences (CAS) for SOT missions, containing four categories, i.e., aircraft, ships, trains, and cars, and derived from 105 video segments. The dataset is publicly available and can be downloaded at http://www.csu.cas.cn/gb/jggk/kybm/sjlyzx/gcxx_sjj/sjj_wxxl/.

The Air-MOT dataset, proposed in 2022 by the Institute of Air and Space Information Innovation, CAS, contains both aircraft and ships and is mainly used for MOT missions (<https://github.com/HeQibin/TGraM>). Jilin-189 is a dataset proposed by Wuhan University in 2022 for super-resolution research. There are no publicly available datasets in the segmentation and scene classification field. Examples of these datasets are shown in Figs. 13–16 (<https://github.com/XY-boy/MSDTGP>).

Recently, the Space Applications Center of CAS has proposed a large-scale multimission satellite video benchmark dataset

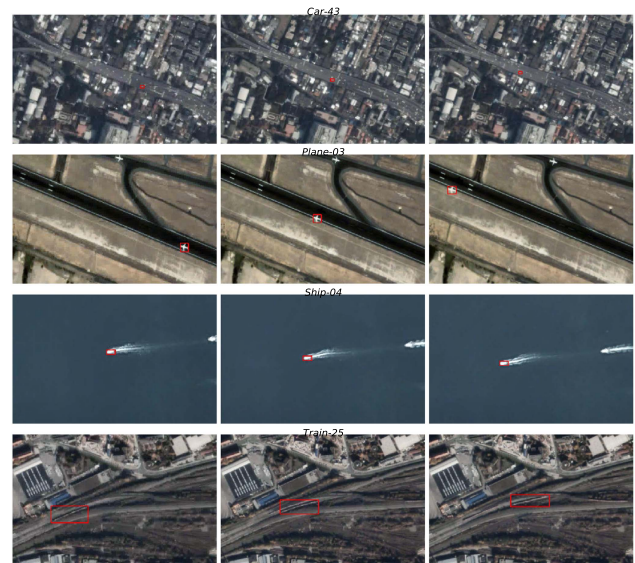


Fig. 13. Example data of SatSOT [99].

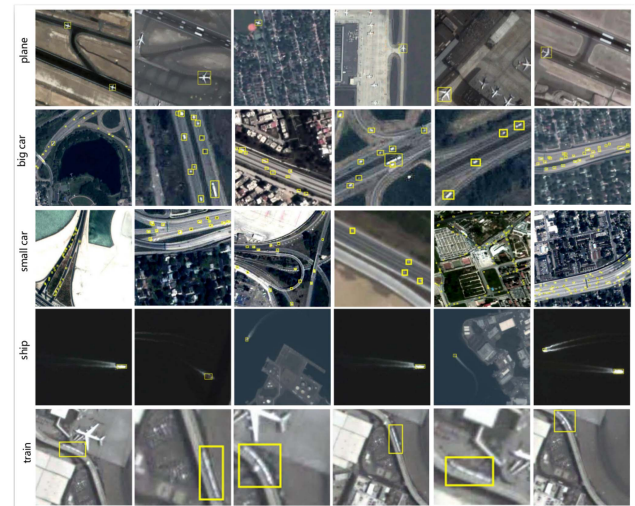


Fig. 14. Example data of VISO [18].



Fig. 15. Example data of AIR-MOT [58].

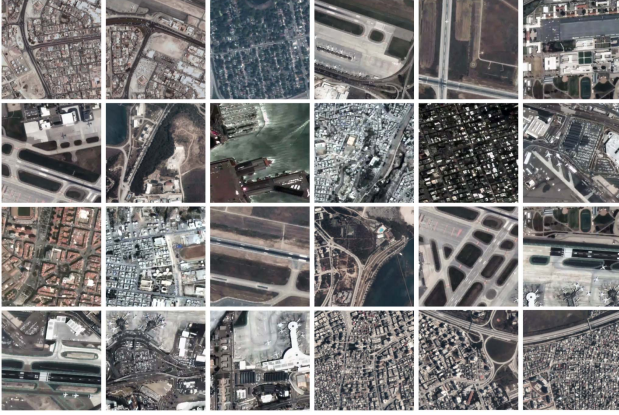


Fig. 16. Example data of Jilin-189 [89].

TABLE V
VOLUME STATISTICS OF SAT-MTB

			HBB	
			Coarse-grained	Fine-grained
Detection	Videos	Frames	3	12
	144	33 228		
			OBB	
			Coarse-grained	Fine-grained
	Videos	Frames	3	12
	106	22 767		
Tracking			HBB	
			Coarse-grained	Fine-grained
	Videos	Frames	4	14
	249	50 046		
			OBB	
			Coarse-grained	Fine-grained
	Videos	Frames	3	12
	106	22 767		
Segmentation			Mask	
			Coarse-grained	Fine-grained
	Videos	Frames	3	12
	144	33 228		

SAT-MTB, which contains 249 content-rich video scenes with 12 fine-grained categories of objects, including aircraft, ships, cars, and trains, covering object tracking, detection, and segmentation tasks [100]. The details of SAT-MTB are shown in Table V. Example data of SAT-MTB is shown in 17. The dataset is publicly available at http://www.csu.cas.cn/gb/kybm/sjlyzx/gcxx_sjj/sjj_wxxl/. Based on the proposed SAT-MTB, Li et al. [100] presented a comprehensive and adequate comparison of benchmark methods on detection, segmentation, and tracking tasks.

B. Evaluation Metrics

This section addresses the evaluation metrics corresponding to satellite video object tracking, object detection, and super-resolution.

1) *Single-Object Tracking*: One-Pass Evaluation (OPE) evaluation criteria include precision plot and success plot. The precision plot shows the percentage of tracking results under a given center location error threshold. The success plot shows the percentage of tracking results with intersection over union (IoU) greater than the given threshold. The tracker is initialized with the given object in the first frame and evaluated on successive frames without resetting.

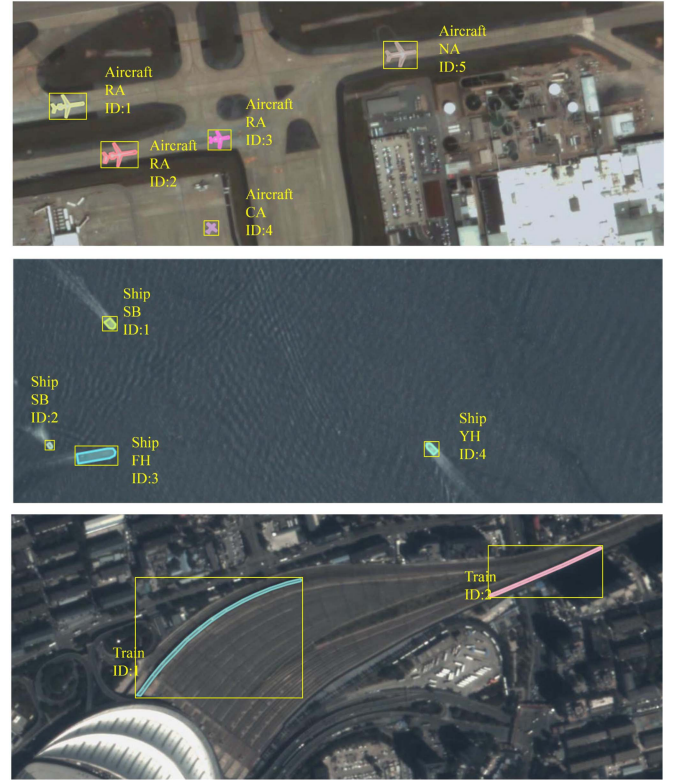


Fig. 17. Example data of SAT-MTB [100].

Expected Average Overlap (EAO) evaluation criteria include accuracy (A), robustness (R), and EAO. Accuracy is the average overlap (AO) between the true value and the predicted bounding box during successful tracking, equivalent to the success score calculated in the OPE metric. Robustness counts the number of times the tracker loses the object during tracking. EAO is the AO estimate for a large number of short-term sequences. During the evaluation, the tracker is reset each time there is no overlap between the predicted bounding box and the true value. And the FPS metric refers to the frame rate at which the algorithm runs and measures the speed of the algorithm.

2) *Multiobject Tracking*: Multiobject tracking accuracy (MOTA) assesses tracking accuracy, which combines false positives (FPs), false negatives (FNs), and IDs. Multiobject tracking precision measures the tracker's precision in estimating object position. Mostly tracked trajectories, partially tracked trajectories, and mostly lost trajectories are used to measure the quality of the tracked trajectory. IDF1 indicates the tracker's recognition performance, which balances identification precision and identification recall by harmonic average. MOTA focuses more on detection performance, while IDF1 is an important measure for long-term tracking. The FPS is also used in MOT to measure the speed.

3) *Object Detection*: Object detection is usually evaluated using the Pascal VOC evaluation criteria: mean average precision (mAP). The mAP is obtained by calculating the average precision (AP) for all the classes and averaging them. The AP represents the area under the precision–recall curve. The larger area means more accurate detection. The recall and precision

are calculated as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

where true positive (TP) represents the number of correctly detected objects, FN represents the number of missed objects, and FP is the number of false predictions. Whether a prediction object is correctly detected is determined by calculating the IoU ratio between the prediction and the ground truth. The prediction objects with greater IoU than the particular threshold is correctly detected, and vice versa. The AP and mAP obtained at different IoU thresholds are also different. For example, the AP and mAP obtained at a threshold of IoU = 0.5 are usually expressed as AP50 and mAP50.

4) *Super-Resolution*: In the field of super-resolution, the most commonly used evaluating metrics with reference images are peak signal-to-noise ratio (PSNR) and structure similarity index measure (SSIM). The PSNR is based on the mean square error (MSE) and measures the ratio of the maximum possible signal power to the noise power. SSIM measures the similarity of the images based on a perceptual model. Both the metrics represent better super-resolution with higher values.

For image X and Y with $m \times n$ size, the MSE and PSNR are calculated as follows:

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [X(i, j) - Y(i, j)]^2 \quad (3)$$

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \quad (4)$$

where MAX is the maximum of image pixel.

The formula of SSIM used in engineering is

$$\text{SSIM} = \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2)}{(\mu_X^2 + \mu_Y^2 + c_1)(\sigma_X^2 + \sigma_Y^2 + c_2)} \quad (5)$$

where μ_X and μ_Y are the average of X and Y . σ_X and σ_Y are the standard deviation of X and Y . σ_{XY} is the covariance of X and Y . c_1 and c_2 are constants. Natural image quality evaluator is more commonly used for blind super-resolution, which does not require a reference image and is more in line with human visual habits. It constructs a series of features that measure image quality and calculates the difference in the distribution of images. The lower the value, the better the image quality.

C. Experimental Results

1) *Single/Multiobject Tracking*: Previous research has shown that several datasets for satellite video object tracking exist to provide a fair and standardized assessment of object tracking algorithms. The SatSOT dataset released in 2022 by the Space Applications Center of CAS [99] is a dataset that focuses on satellite video SOT. It contains 105 video sequences from three commercial satellite sources: Jilin-1, Skybox, and Carbonite-2. The dataset contains aircraft, cars, boats, and

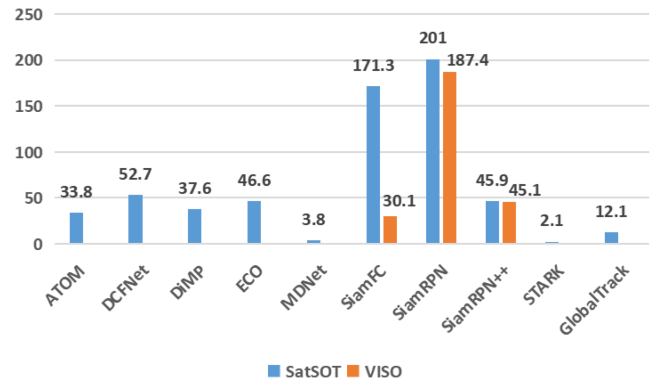


Fig. 18. Tracking speed using GPU (FPS) of each tracker on SatSOT and VISO.

trains with a video frame rate of 10 or 25 FPS and a total of 27 664 frames. The results of the experiments on the classic methods ATOM [31], DCFNet [101], DiMP [32], ECO [30], MDNet [33], SiamFC [29], SiamRPN [102], SiamRPN++ [28], STARK [103], and GlobalTrack [104] are shown in Fig. 18.

The VISO [18] dataset, released by the National University of Defense Technology in 2021, is a large-scale dataset for moving object detection and tracking in satellite video. For the SOT task, the dataset provides 3159 video sequences. For the MOT task, the dataset collected 47 video segments captured by the Jilin-1 satellite, containing 3711 individual example objects, including aircraft, cars, ships, and trains. Each full scene in VISO has a resolution of 12 000 × 5000 pixels and contains a large number of objects at different scales, and the video has a frame rate of 10 FPS. The data organization corresponding to the detection task is also provided in VISO, with 1 646 038 annotated instances. The dataset is only annotated with objects in motion in the video, and over 90% of the instances are vehicles.

Over the past few years, several benchmarks have been developed for satellite video. Generally, DSFNet [54] and CFME [40] are for object detection and SOT, respectively. DeepSORT [105] is generally chosen as the benchmark for MOT. Table VI shows these benchmark experimental results on SatSOT and VISO.

AIR-MOT [58], released in 2022 by the Institute of Air and Space Information of the CAS, contains a total of ten complete scenes and 149 videos collected by the Jilin-1 satellite. The dataset has 5736 instances labeled using axis-aligned bounding boxes and contains aircraft and ships. Each video has a frame rate of 5–10 FPS and a resolution of 1920 × 1080 pixels. Test accuracy of different MOT algorithms (e.g., DeepSORT [105], RAN [106], HOGM [107], DAN [108], Tracktor+CTdet [109], CKDNet+SMTNet [53], TubeTK [110], CTracker [111], JDE [56], UMA [112], CenterTrack [55], GSdT [113], FairMOT [57], TraDeS [114], and TGraM [58]) on AIR-MOT is shown in Fig. 19.

2) *Super-Resolution*: The majority of satellite VSR datasets were constructed from videos captured by the Jilin-1 and OVS-1 satellites. Still, only a few are publicly available, and there is a lack of widely used publicly available datasets.

Xiao et al. [89] provided a publicly available satellite VSR dataset Jilin-189, which consists of ten Jilin-1 videos cropped

TABLE VI
PRECISION AND SUCCESS OF OPE ON SATSOT AND VISO

	GlobalTrack [104]	STARK [103]	SiamRPN++ [28]	MDNet [33]	ATOM [31]	CFME [40]	SiamFC [29]	SiamRPN [102]
Precision of OPE	0.216/–	0.404/–	0.537/0.479	0.597/–	0.538/–	0.555/–	0.488/0.491	0.509/0.437
Success of OPE	0.211/–	0.345/–	0.423/0.199	0.481/–	0.435/–	0.428/–	0.404/0.269	0.393/0.171

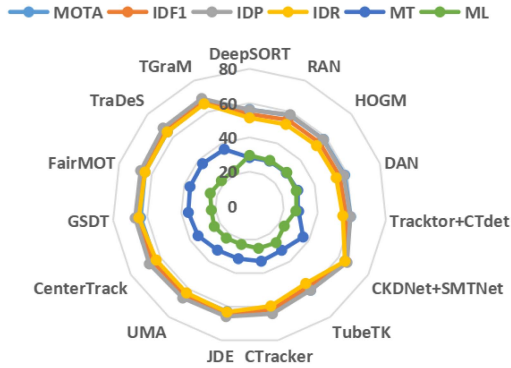


Fig. 19. Test accuracy of different methods on AIR-MOT.

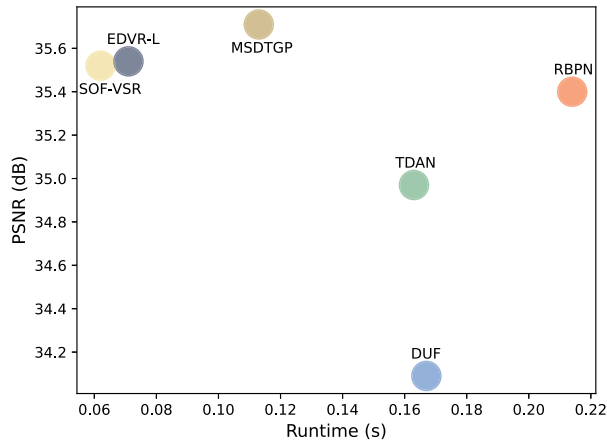


Fig. 20. Speed and performance comparison on the Jilin-189.

and divided into 189 training sets and ten test sets, each with 640×640 resolution and 100 frames in length, and its low-resolution videos are obtained by quadruple downsampling through Bicubic. Fig. 20 shows the performance of some VSR methods on the Jilin-189 dataset, where SOF-VSR [115] and EDVR-L [78] have the fastest speed and RBPN [116] is the slowest. SOF-VSR, EDVR-L, MSDTGP [81], and RBPN all have good accuracy, while TDAN [117] and DUF [118] perform poorly on this task, with MSDTGP being the most balanced, with a PSNR of 35.71 dB, achieving the highest accuracy while having good speed.

VI. TYPICAL APPLICATION SCENARIOS

The development of video satellite technology has led to an increasingly wide range of applications in traffic detection density estimation, scene monitoring, automatic 3-D model construction, global change research, and disaster monitoring. This section introduces the typical applications of satellite video

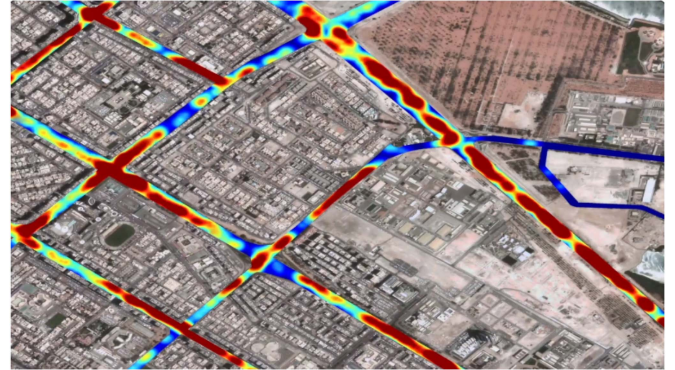


Fig. 21. Jilin-1 satellite video multiobjective dynamic monitoring heat map [119].



Fig. 22. Jilin-1 satellite video traffic flow statistics analysis [120].



Fig. 23. Example of monitoring of Zondervoort Correctional Center [120].

in traffic detection density estimation, scene monitoring, and automatic 3-D model construction.

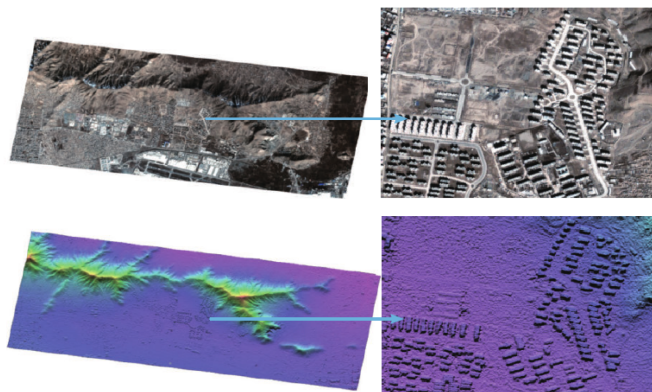


Fig. 24. Left: Jilin-1 video satellite image single frame and digital surface model. Right: Image single frame and digital surface model building partial Automatic construction of 3-D models of Jilin-1 video satellite images [121].



Fig. 25. Oil spill and fire incident monitoring [122].

A. Traffic Detection Density Estimation

Using the high-resolution satellite video, Jilin-1, the motor vehicles driving on the traffic road within the satellite video are detected, and their geographical location and semantic features are obtained. At the same time, multiple motor vehicles driving in the traffic road within the satellite video are tracked, and their dynamic information, such as motion object, direction, trajectory, and speed, is extracted and analyzed to generate a video and heat map. Using the video data returned by the Jilin-1 video satellite in real time, the moving vehicles in the video can be automatically detected, tracked, and located. According to the detection results, the traffic heat map, traffic flow statistics map, and the video analyzing the vehicles' motion postures are generated, thus realizing intelligent analysis of road condition information and traffic flow information, reducing human cost, and realizing the intelligent transportation system. Fig. 21 shows the Jilin-1 satellite video multiobjective dynamic monitoring heat map.

B. Scene Monitoring

By planning the operation sequence and operation time of the satellites, the Carbonite series satellites can achieve multiple visits to a specific earth location during the time of day, allowing for uninterrupted monitoring of regional hot spots, as well as the detection of changes in hot-spot areas using uninterrupted monitoring video synthesized from multiple satellites. Fig. 22 shows the Jilin-1 satellite video traffic flow statistics analysis. Fig. 23 shows an example of monitoring of Zondervoort Correctional Center.

C. Automatic Construction of 3-D Models

Through absolute orientation and image stabilization processing of satellite video, the matching success rate is improved using multiview stereo matching, giving faith to generating dense matching results of the same name image points and forming digital surface models through the rendezvous of images to realize the automatic construction of ground 3-D models. Fig. 24 shows the automatic construction of 3-D models of Jilin-1 video satellite images.

D. Incident and Disaster Response

The time-continuous nature of video satellite observation of the earth makes it useful for many emergency and disaster response applications. When natural disasters such as earthquakes, tsunamis, typhoons, and forest fires occur, satellite video can help locate the disaster's location quickly, support subsequent rescue, and help disaster relief departments make quick decisions. In significant accidents, such as city fires, hazardous materials explosions, offshore oil leaks, and other occurrences, satellite video can not only help determine the level of the accident and assist firefighters to rescue but also provide a solid basis for the post-accident analysis of the cause of the accident and find the party responsible for the accident, while also providing experience in the prevention of similar accidents. The image above is an example of an image sequence taken by the PlanetScope series of satellites, a time series of images taken in 2018 in Balikpapan Bay, Indonesia, monitoring an oil spill fire event in the area, and providing assistance in locating and tracking oil slicks, locating oil spill vessels, and future accident prevention. Fig. 25 shows the oil spill and fire incident monitoring.

E. Land Space Use Regulation

Land space use regulation regulates the sustainable use of natural resource carriers based on spatial use, development, and utilization restrictions determined by land space planning. Geological environment disaster prevention and mitigation, ecological restoration, and law enforcement supervision belong to the land space use supervision field, while the object identification of the above application scenarios mainly relies on manual image interpretation and field investigation.

DL-based object detection technology can accurately identify all the object categories and scenes of interest in satellite images and quickly determine their locations and sizes. It can accurately

identify multiple natural resource objects and scene categories and assist in land space use control, ecological restoration, geological disaster control, and law enforcement inspection by determining the locations and interrelationships of crucial natural resource objects [123].

F. Maritime Vessel Situational Awareness

China has a vast marine area and rich marine resources, and it is of great strategic significance to strengthen the rational use of resources for the development of China. Maritime ship situational awareness is an essential maritime safety and security research direction. It senses the ship itself and the surrounding environmental factors through intelligent analysis technology and then understands and analyzes the sensed situational elements to make predictions on the movement trend of the ship to avoid maritime accidents. Traditional maritime navigation safety assurance usually relies on commanders to make a judgment with the assistance of AIS, radar, and remote sensing images with high labor costs.

As intelligent information processing technology is widely used in navigation safety, the DL-based technology of computer vision scene analysis plays a crucial role in maritime ship situational awareness tasks such as ship detection and heading prediction. The team of Hainan University designed a multitasking panoramic ship situational awareness intelligent model integrating ship detection, sea, land segmentation, and heading prediction, which can predict the driving status and movement trend of ships under different weather conditions. It developed a panoramic marine ship situational awareness system to realize accurate sensing of ship situational and surrounding environment situational and other elements and assist navigation commanders in making more accurate, reasonable, and fast decisions. The actual application of maritime ship situational awareness can achieve an accuracy of not less than 90% at 12 FPS [124].

VII. FUTURE OUTLOOK

It should be emphasized that it is necessary to consider the practical application requirements of remote sensing scenarios and the unique properties of the targets of interest in satellite video. For example, most targets are rigid bodies; problems of deformation, occlusion, and scale transformation are not common; and problems of small target occupancy and sparse temporal information are more common.

Based on the above analysis, this section proposes some unsolved tasks and future possible development directions, hoping to provide some ideas and inspiration for researchers to jointly promote the innovative development of satellite video intelligent processing direction. The details are as follows.

- 1) *Establishing satellite video datasets for multiple tasks and unifying annotation formats*: In the field of satellite video, although some datasets have been constructed for various research tasks, they are oriented to fewer task categories and cannot meet the needs of multiple satellite video tasks. Moreover, the available data sources of satellite video are limited, and constructing a dataset only for a single task

cannot fully utilize the existing satellite video data. At the same time, the existing satellite video public datasets have low category richness, nonuniform annotation format, and a large gap in the number of annotations compared with image-based datasets, so it is crucial to building a large-scale satellite video dataset with rich object categories and uniform annotation format that integrates multiple tasks for future research in various satellite video tasks.

- 2) *Enhancing the robustness of the algorithm and improving the upper limit of the satellite video task in practical applications*: In practical scenarios, due to factors such as lighting changes, cloud occlusion, and different geographical locations of satellite photography, satellite video has problems such as complex background environment and unstable video quality, which seriously affect the accuracy of the satellite video processing algorithm. At the same time, the labeling noise brought by the imaging quality and manual labeling errors also makes the algorithm easily interfered with by irrelevant noise features. Therefore, how to improve the robustness of algorithm learning from coarse labeled samples is an urgent problem for each task algorithm in the satellite video field.
- 3) *Few-sample and zero-sample learning*: The existing algorithms usually define small samples as hundreds of labeled samples. This experimental setup may sometimes be unrealistic. On the one hand, satellite video data are harder to obtain and smaller in amount compared to general-purpose video data. On the other hand, there is a large gap between the number of training samples required by algorithms and practical applications in the existing satellite video datasets, where some categories have only a few videos. Therefore, with the need to reduce the training samples for video understanding, less and zero-sample learning applicable to the satellite video domain is a promising direction.
- 4) *Weakly supervised and unsupervised learning*: In the field of satellite video, the leading algorithms are still supervised algorithms based on a large amount of fully labeled data, which are more effective but require the use of datasets that take a lot of time and manpower to label, and the small size of the objects in satellite video data can cause a great burden for labelers when performing data labeling for tasks with intensive prediction requirements such as segmentation. Meanwhile, the small size of objects in satellite video data can cause a great burden to the annotators when annotating data for tasks with intensive prediction needs, such as segmentation, and affect the accuracy of annotation. At the same time, the fine-grained annotation of objects also requires a lot of reliance on the empirical knowledge of experts. Therefore, weakly supervised learning and unsupervised learning methods that require little manpower for annotation are one of the future directions for each task of satellite video.
- 5) *Using multimodal satellite video data to achieve multimodal fusion of data and cross-modal migration of models*: With the development of in-orbit technology, more

and more video satellites can support other modal video capture besides visible videos, such as SAR video, infrared video, etc. The learning of multimodal data can make up for the deficiencies of models in feature extraction completeness and resistance to noise; different modal data complement each other. Most current satellite video algorithms are only applicable to a single modality and cannot migrate to other modalities, so multimodal data fusion and cross-modal model migration is a research direction for satellite video algorithms for each task, which is important to achieve multidimensional satellite video understanding [122].

VIII. CONCLUSION

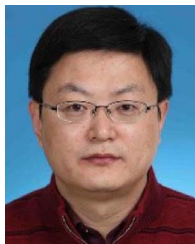
Intelligent processing of satellite video has made rapid development in the past decade. For different demands, satellite video intelligent processing has derived more and more task directions, especially object tracking, object detection, and super-resolution, the three most popular directions. This article introduced and summarized the latest progress in the field of intelligent processing of satellite video, including the existing challenges, existing methods, and relevant application scenarios. First, we quantitatively and statistically analyzed the relevant research results on the topic of satellite video intelligence processing, conducted statistical analysis on the distribution of the year of publication, journal distribution, and task-specific direction distribution of articles on the topic of satellite video, and showed the keyword hot-spot distribution and development trend in this field. Then, this article introduced the research progress and methodological systems for satellite video object tracking and motion estimation, satellite video object detection, satellite VSR, satellite VOS, and scene classification tasks. Next, to make a fair comparison of the performance of existing methods, we investigated the existing public datasets under different tasks and compared the experimental results of different methods on each dataset. Furthermore, this article introduced the typical application scenarios of satellite video intelligent processing in real life. Finally, taking into account the current challenges and practical needs in this field, this article discussed several promising directions that can be explored and studied.

REFERENCES

- [1] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.
- [2] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.
- [3] G. Kopsiaftis and K. Karantzas, "Vehicle detection and traffic density monitoring from very high resolution satellite video data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 1881–1884.
- [4] L. Yile, L. Yanping, and W. Yan, "Research on satellite video traffic flow parameter extraction based on optical flow method," *Comput. Eng. Appl.*, vol. 54, no. 10, pp. 204–207, 2018.
- [5] 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.
- [6] 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.*, 2019, pp. 1504–1509.
- [7] 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, 2020.
- [8] G. Liu, S. Li, and Y. Shao, "Low-quality and multi-target detection in RSIs," 2016.
- [9] Y. Zhong, M. Shu, Z. Liu, and X. Lu, "Spatio-temporal dual-branch network with predictive feature learning for satellite video object segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5523215.
- [10] S. Zhang, Q. Yuan, and J. Li, "Video satellite imagery super resolution for 'Jilin-1' via a single-and-multi frame ensemble framework," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 2731–2734.
- [11] Y. Gu, H. Liu, T. Wang, S. Li, and G. Gao, "Deep feature extraction and motion representation for satellite video scene classification," *Sci. China Inf. Sci.*, vol. 63, pp. 1–15, 2020.
- [12] Y. Li et al., "Deep learning-based object tracking in satellite videos: A comprehensive survey with a new dataset," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 181–212, Dec. 2022.
- [13] Z.-P. Luo, J.-Z. Yang, Z.-P. Xue, and M. Li, "Research and application of urban traffic survey method based on commercial video satellite remote sensing technology," in *Proc. 13th Asia Pacific Transp. Develop. Conf.*, 2020, pp. 10–18.
- [14] Y. Guo et al., "The first challenge on moving object detection and tracking in satellite videos: Methods and results," in *Proc. 26th Int. Conf. Pattern Recognit.*, 2022, pp. 4981–4988.
- [15] H. Liu and Y. Gu, "A summary of super-resolution for satellite videos via learning-based methods," in *Proc. 10th Workshop Hyperspectral Imag. Signal Process.: Evol. Remote Sens.*, 2019, pp. 1–4.
- [16] Z. Zhang, C. Wang, J. Song, and Y. Xu, "Object tracking based on satellite videos: A literature review," *Remote Sens.*, vol. 14, no. 15, 2022, Art. no. 3674.
- [17] C. Gómez, J. C. White, and M. A. Wulder, "Optical remotely sensed time series data for land cover classification: A review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 116, pp. 55–72, 2016.
- [18] Q. Yin et al., "Detecting and tracking small and dense moving objects in satellite videos: A benchmark," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5612518.
- [19] T. Vojir, J. Noskova, and J. Matas, "Robust scale-adaptive mean-shift for tracking," *Pattern Recognit. Lett.*, vol. 49, pp. 250–258, 2014.
- [20] K. Nummiaro, E. Koller-Meier, and L. Van Gool, "Object tracking with an adaptive color-based particle filter," in *Proc. Joint Pattern Recognit. Symp.*, 2002, pp. 353–360.
- [21] R. E. Kalman, "A new approach to linear filtering and prediction theory," *ASME J. Basic Eng., Ser. D*, vol. 82, no. 1, pp. 35–45, 1961.
- [22] H. Possegger, T. Mauthner, and H. Bischof, "In defense of color-based model-free tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2113–2120.
- [23] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2544–2550.
- [24] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.
- [25] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4310–4318.
- [26] 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, Nov. 2019.
- [27] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 850–865.
- [28] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4282–4291.
- [29] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, "SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 12549–12556.
- [30] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6638–6646.

- [31] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4660–4669.
- [32] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6182–6191.
- [33] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4293–4302.
- [34] Y. Cui, B. Hou, Q. Wu, B. Ren, S. Wang, and L. Jiao, "Remote sensing object tracking with deep reinforcement learning under occlusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5605213.
- [35] S. A. Ahmadi and A. Mohammadzadeh, "A simple method for detecting and tracking vehicles and vessels from high resolution spaceborne videos," in *Proc. Conf. Joint Urban Remote Sens. Event*, 2017, pp. 1–4.
- [36] J. Shao, B. Du, C. Wu, and L. Zhang, "Tracking objects from satellite videos: A velocity feature based correlation filter," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7860–7871, Oct. 2019.
- [37] 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.
- [38] X. Chen and H. Sui, "Real-time tracking in satellite videos via joint discrimination and pose estimation," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 23–29, 2019.
- [39] Y. Guo, D. Yang, and Z. Chen, "Object tracking on satellite videos: A correlation filter-based tracking method with trajectory correction by Kalman filter," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3538–3551, Sep. 2019.
- [40] 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.
- [41] S. Xuan et al., "Rotation adaptive correlation filter for moving object tracking in satellite videos," *Neurocomputing*, vol. 438, pp. 94–106, 2021.
- [42] Y. Chen, Y. Tang, T. Han, Y. Zhang, B. Zou, and H. Feng, "RAMC: A rotation adaptive tracker with motion constraint for satellite video single-object tracking," *Remote Sens.*, vol. 14, no. 13, 2022, Art. no. 3108.
- [43] W. Pei and X. Lu, "Moving object tracking in satellite videos by kernelized correlation filter based on color-name features and Kalman prediction," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–16, 2022.
- [44] Y. Liu, Y. Liao, C. Lin, Z. Li, X. Yang, and A. Zhang, "Object tracking in satellite videos based on improved correlation filters," in *Proc. 13th Int. Conf. Commun. Softw. Netw.*, 2021, pp. 323–331.
- [45] 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.
- [46] Z. Hu, D. Yang, K. Zhang, and Z. Chen, "Object tracking in satellite videos based on convolutional regression network with appearance and motion features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 783–793, 2020.
- [47] B. Uzkenet, A. Rangnekar, and M. J. Hoffman, "Tracking in aerial hyperspectral videos using deep kernelized correlation filters," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 449–461, Jan. 2019.
- [48] K. Zhu et al., "Single object tracking in satellite videos: Deep siamese network incorporating an interframe difference centroid inertia motion model," *Remote Sens.*, vol. 13, no. 7, 2021, Art. no. 1298.
- [49] L. Ruan, Y. Guo, D. Yang, and Z. Chen, "Deep siamese network with motion fitting for object tracking in satellite videos," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6508005.
- [50] 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.
- [51] W. Zhang, L. Jiao, F. Liu, L. Li, X. Liu, and J. Liu, "MBLT: Learning motion and background for vehicle tracking in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4703315.
- [52] F. Bi, J. Sun, J. Han, Y. Wang, and M. Bian, "Remote sensing target tracking in satellite videos based on a variable-angle-adaptive siamese network," *IET Image Process.*, vol. 15, no. 9, pp. 1987–1997, 2021.
- [53] 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. Photogrammetry Remote Sens.*, vol. 177, pp. 116–130, 2021.
- [54] C. Xiao et al., "DSFNet: Dynamic and static fusion network for moving object detection in satellite videos," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 3510405.
- [55] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 474–490.
- [56] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 107–122.
- [57] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 3069–3087, 2021.
- [58] Q. He, X. Sun, Z. Yan, B. Li, and K. Fu, "Multi-object tracking in satellite videos with graph-based multitask modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619513.
- [59] T. Wang, Y. F. Gu, and S. Li, "A multi-frame sparse self-learning PWC-Net for motion estimation in satellite video scenes," *Sci. China Inf. Sci.*, vol. 60, pp. 1–13, 2022.
- [60] W. Ao, Y. Fu, and F. Xu, "Detecting tiny moving vehicles in satellite videos," 2018, *arXiv:1807.01864*.
- [61] J. Zhang, X. Jia, and J. Hu, "Error bounded foreground and background modeling for moving object detection in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2659–2669, Apr. 2020.
- [62] J. Zhang, X. Jia, J. Hu, and J. Chanussot, "Online structured sparsity-based moving-object detection from satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6420–6433, Sep. 2020.
- [63] J. Zhang, X. Jia, J. Hu, and K. Tan, "Moving vehicle detection for remote sensing video surveillance with nonstationary satellite platform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5185–5198, Sep. 2022.
- [64] J. Lei, Y. Dong, and H. Sui, "Tiny moving vehicle detection in satellite video with constraints of multiple prior information," *Int. J. Remote Sens.*, vol. 42, no. 11, pp. 4110–4125, 2021.
- [65] X. Zhang, J. Xiang, and Y. Zhang, "Space object detection in video satellite images using motion information," *Int. J. Aerosp. Eng.*, vol. 2017, 2017, Art. no. 1024529.
- [66] H. Li, L. Chen, F. Li, and M. Huang, "Ship detection and tracking method for satellite video based on multiscale saliency and surrounding contrast analysis," *J. Appl. Remote Sens.*, vol. 13, no. 2, 2019, Art. no. 026511.
- [67] F. Shi, F. Qiu, X. Li, R. Zhong, C. Yang, and Y. Tang, "Detecting and tracking moving airplanes from space based on normalized frame difference labeling and improved similarity measures," *Remote Sens.*, vol. 12, no. 21, 2020, Art. no. 3589.
- [68] M. Shu, Y. Zhong, and P. Lv, "Small moving vehicle detection via local enhancement fusion for satellite video," *Int. J. Remote Sens.*, vol. 42, no. 19, pp. 7189–7214, 2021.
- [69] X. Chen, H. Sui, J. Fang, M. Zhou, and C. Wu, "A novel AMS-DAT algorithm for moving vehicle detection in a satellite video," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 3501505.
- [70] R. Pflugfelder, A. Weissenfeld, and J. Wagner, "Deep vehicle detection in satellite video," 2022, *arXiv:2204.06828*.
- [71] Z. Zhou, S. Li, W. Guo, and Y. Gu, "Few-shot aircraft detection in satellite videos based on feature scale selection pyramid and proposal contrastive learning," *Remote Sens.*, vol. 14, no. 18, 2022, Art. no. 4581.
- [72] Z. Pi et al., "Very low-resolution moving vehicle detection in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5624517.
- [73] H. Liu et al., "Video super-resolution based on deep learning: A comprehensive survey," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 5981–6035, 2022.
- [74] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The search for essential components in video super-resolution and beyond," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4947–4956.
- [75] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "BasicVSR : Improving video super-resolution with enhanced propagation and alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5972–5981.
- [76] Y. Luo, L. Zhou, S. Wang, and Z. Wang, "Video satellite imagery super resolution via convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2398–2402, Dec. 2017.
- [77] A. Xiao, Z. Wang, L. Wang, and Y. Ren, "Super-resolution for "Jilin-1" satellite video imagery via a convolutional network," *Sensors*, vol. 18, no. 4, 2018, Art. no. 1194.
- [78] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1954–1963.

- [79] Z. He, J. Li, L. Liu, D. He, and M. Xiao, "MultiFrame video satellite image super-resolution via attention-based residual learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5605015.
- [80] Y. Xiao, X. Su, and Q. Yuan, "A recurrent refinement network for satellite video super-resolution," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 3865–3868.
- [81] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, "Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610819.
- [82] N. Ni, H. Wu, and L. Zhang, "Deformable alignment and scale-adaptive feature extraction network for continuous-scale satellite video super-resolution," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 2746–2750.
- [83] H. Liu and Y. Gu, "Deep joint estimation network for satellite video super-resolution with multiple degradations," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621015.
- [84] H. Shen, Z. Qiu, L. Yue, and L. Zhang, "Deep-learning-based super-resolution of video satellite imagery by the coupling of multiframe and single-frame models," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5612114.
- [85] Z. He and D. He, "A unified network for arbitrary scale super-resolution of video satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8812–8825, Oct. 2021.
- [86] Z. He, X. Li, and R. Qu, "Video satellite imagery super-resolution via model-based deep neural networks," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 749.
- [87] Z. He, D. He, X. Li, and J. Xu, "Unsupervised video satellite super-resolution by using only a single video," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6000905.
- [88] D.-L. Chen, L. Zhang, and H. Huang, "Robust extraction and super-resolution of low-resolution flying airplane from satellite video," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4700916.
- [89] Y. Xiao, X. Su, and Q. Yuan, "Learning an intrinsic graph neural network for satellite video super-resolution," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 3751–3753.
- [90] Y. Xiao et al., "Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial-temporal transformer," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, 2022, Art. no. 102731.
- [91] V. Badrinarayanan, I. Budvytis, and R. Cipolla, "Semi-supervised video segmentation using tree structured graphical models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2751–2764, Nov. 2013.
- [92] S. Caellès, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 221–230.
- [93] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015.
- [94] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, pp. 98–136, 2015.
- [95] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5188–5197.
- [96] T. Wang, Y. Gu, and G. Gao, "Satellite video scene classification using low-rank sparse representation two-stream networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622012.
- [97] G. Gao, Y. Gu, and S. Li, "Satellite video intrinsic decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2022, Art. no. 5631413.
- [98] J. Pan, Y. Gu, S. Li, G. Gao, and S. Wu, "Intrinsic satellite video decomposition with motion target energy constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5633913.
- [99] M. Zhao, S. Li, S. Xuan, L. Kou, S. Gong, and Z. Zhou, "SatSOT: A benchmark dataset for satellite video single object tracking," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617611.
- [100] S. Li et al., "A multitask benchmark dataset for satellite video: Object detection, tracking, and segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5611021.
- [101] Q. Wang, J. Gao, J. Xing, M. Zhang, and W. Hu, "DCFNet: Discriminant correlation filters network for visual tracking," 2017, *arXiv:1704.04057*.
- [102] 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.*, 2018, pp. 8971–8980.
- [103] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4904–4913.
- [104] L. Huang, X. Zhao, and K. Huang, "Globaltrack: A simple and strong baseline for long-term tracking," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 11037–11044.
- [105] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3645–3649.
- [106] 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.*, 2018, pp. 466–475.
- [107] 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.*, 2018, pp. 1809–1814.
- [108] S. Sun, N. Akhtar, H. Song, A. Mian, and M. Shah, "Deep affinity network for multiple object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 104–119, Jan. 2021.
- [109] P. Bergmann, T. Meinhardt, and L. Leal-Taixé, "Tracking without bells and whistles," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 941–951.
- [110] B. Pang, Y. Li, Y. Zhang, M. Li, and C. Lu, "TubeTK: Adopting tubes to track multi-object in a one-step training model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6308–6318.
- [111] J. Peng et al., "Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 145–161.
- [112] J. Yin, W. Wang, Q. Meng, R. Yang, and J. Shen, "A unified object motion and affinity model for online multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6768–6777.
- [113] Y. Wang, K. Kitani, and X. Weng, "Joint object detection and multi-object tracking with graph neural networks," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2021, pp. 13708–13715.
- [114] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, and J. Yuan, "Track to detect and segment: An online multi-object tracker," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12352–12361.
- [115] L. Wang, Y. Guo, Z. Lin, X. Deng, and W. An, "Learning for video super-resolution through HR optical flow estimation," in *Proc. 14th Asian Conf. Comput. Vis.*, 2019, pp. 514–529.
- [116] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3897–3906.
- [117] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally-deformable alignment network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3360–3369.
- [118] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3224–3232.
- [119] *Jilin-1 Satellite Constellation Eco-open Store*, Jilin-1. Accessed: Mar. 29, 2023. [Online]. Available: <https://www.jl1mall.com/SatelliteImagery/SatVideo>
- [120] *Carbonite Video Demonstration Missions of SSTL on Microsatellites*, eoPortal.org. Accessed: Mar. 29, 2023. [Online]. Available: <https://www.eoportal.org/satellite-missions/carbonite>
- [121] L. Beibei, B. Han, T. Tian, R. Zhu, and Y. Bai, "Application status and future development of Jilin no.1 video satellite," *Satell. Appl.*, vol. 03, pp. 23–27, 2018.
- [122] *Pipeline Failure Cause of Fatal Oil Spill in Indonesia*, SkyTruth, Shepherdstown, WV, USA. Accessed: Jun. 14, 2013. [Online]. Available: <https://skytruth.org/2018/04/pipeline-failure-cause-of-fatal-oil-spill-in-indonesia/>
- [123] F. Minglu, "Application research of target detection technology based on AI+ remote sensing in natural resource monitoring." Accessed: Jun. 14, 2013. [Online]. Available: https://www.elecfans.com/application/Military_avionics/2020/0805/1265380.html
- [124] J. Wang, "Research on ship situation awareness in remote sensing image based on multi-task learning," M.S. thesis, Hainan Univ., Haikou, China, 2022.



Shengyang Li received the Ph.D. degree in computer application technology from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2006.

He is currently a Professor with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences. His research interests include machine learning in remote sensing image interpretation, deep learning in satellite videos processing and analysis, intelligent image processing, analysis and understanding for space utilization, and space

scientific big data modeling and analysis.



Xian Sun (Senior Member, IEEE) received the B.Sc. degree in electronic information engineering from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees in signal and information processing from the Institute of Electronics, Chinese Academy of Sciences (CAS), Beijing, in 2009.

In 2013, he was a Visiting Scholar with the Karlsruhe Institut für Technologie, Karlsruhe, Germany. He is currently a Professor with the Aerospace Information Research Institute, CAS. His research inter-

ests include computer vision, geospatial data mining, and remote sensing image understanding.

Dr. Sun was a recipient of the Outstanding Science and Technology Achievement Prize of the CAS in 2016 and the First Prize for the State Scientific and Technological Progress of China in 2019. He is an Associate Editor for IEEE ACCESS and a Guest Editor for the special issue of IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING and other journals.



Yanfeng Gu (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2005.

He joined the School of Electronics and Information Engineering, HIT, as a Lecturer and became an Associate Professor in 2006. He was enrolled in the First Outstanding Young Teacher Training Program of HIT. From 2011 to 2012, he was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, CA, USA. He is currently a Professor with the Department of Information Engineering, HIT. He has authored more than 60 peer-reviewed articles and four book chapters. He holds seven patents. His research interests include image processing in remote sensing, machine learning, pattern analysis, and multiscale geometric analysis.

Dr. Gu is an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Science China Technological Sciences*, and *Neurocomputing*.



Yixuan Lv received the B.Sc. degree in electronic information engineering from Xidian University, Xi'an, China, in 2019, and the M.Sc. degree in signal and information processing from the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2022.

She is currently an Assistant Engineer with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences. Her research interests include computer vision and deep learning, especially in satellite video object segmentation and

synthetic aperture radar object detection.



Manqi Zhao received the B.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, in 2019. He is currently working toward the Ph.D. degree in computer applied technology with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing.

His research interests include satellite video, unmanned aerial vehicle video, and conventional video analysis, with focus on object tracking.



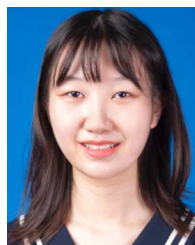
Zhuang Zhou received the B.Eng. degree in electrical engineering and automation from the China University of Mining and Technology, Xuzhou, China, in 2013, and the M.S. degree in cartography and geography information system from Beijing Normal University, Beijing, China, in 2016.

He is currently an Engineer with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing. His research interests include satellite video object detection.



Weilong Guo received the B.Eng. degree in software engineering from Jilin University, Jilin, China, in 2018, and the M.Eng. degree in computer applied technology from the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, in 2022.

He is currently an Assistant Engineer with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing. His research interests include intelligent analysis and understanding of image and video.



Yuhan Sun received the B.S. degree in automation and computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 2021. She is currently working toward the Ph.D. degree in computer-applied technology with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China.

Her research interests include satellite video and conventional video analysis, with a focus on object detection and segmentation.



Han Wang received the B.E. degree in electrical engineering from Chongqing University, Chongqing, China, in 2022. He is currently working toward the Ph.D. degree in computer applied technology with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China.

His research interests include computer vision and its application to remote sensing image super-resolution and object detection.



Jian Yang received the B.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, in 2020. He is currently working toward the Ph.D. degree in computer applied technology with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing.

His research interests include images/videos understanding and analysis.