ISPRS Journal of Photogrammetry and Remote Sensing 210 (2024) 212–240

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Satellite video single object tracking: A systematic review and an oriented object tracking benchmark

Yuzeng Chen [a], Yuqi Tang [b,*], Yi Xiao [a], Qiangqiang Yuan [a,c], Yuwei Zhang [b], Fengqing Liu [b], Jiang He [a], Liangpei Zhang [d]

[a] School of Geodesy and Geomatics, Wuhan University, Wuhan, Hubei 430079, China
[b] School of Geosciences and Info-Physics, Central South University, Changsha, China
[c] Hubei Luojia Laboratory, Wuhan, Hubei 430079, China
[d] State Key Laboratory of Information Engineering, Survey Mapping and Remote Sensing, Wuhan University, Wuhan, China

ARTICLE INFO

Keywords:
Satellite video
Deep learning
Correlation filter
Single object tracking
Benchmark

ABSTRACT

Single object tracking (SOT) in satellite video (SV) enables the continuous acquisition of position and range information of an arbitrary object, showing promising value in remote sensing applications. However, existing trackers and datasets rarely focus on the SOT of oriented objects in SV. To bridge this gap, this article presents a comprehensive review of various tracking paradigms and frameworks covering both the general video and satellite video domains and subsequently proposes the oriented object tracking benchmark (OOTB) to advance the field of visual tracking. OOTB contains 29,890 frames from 110 video sequences, covering common satellite video object categories including car, ship, plane, and train. All frames are manually annotated with oriented bounding boxes, and each sequence is labeled with 12 fine-grained attributes. Additionally, a high-precision evaluation protocol is proposed for comprehensive and fair comparisons of trackers. To validate the existing trackers and explore frameworks suitable for SV tracking, we benchmark 33 state-of-the-art trackers totaling 58 models with different features, backbones, and tracker tags. Finally, extensive experiments and insightful thoughts are also provided to help understand their performance and offer baseline results for future research. OOTB is available at https://github.com/YZCU/OOTB.

1. Introduction

Single object tracking (SOT) is one of the most essential tasks in computer vision, which allows the establishment of object correspondences in video sequences (Javed et al., 2022). Given the initial state, SOT aims to determine subsequent states of an arbitrary object. SOT can be applied to a variety of fields such as autonomous driving, intelligent surveillance, robotics, and augmented reality. Tracking technology has received a lot of attention, and many advanced trackers have been proposed to solve realistic challenges such as scale variation, deformation, similar appearance, and illumination changes (Chen et al., 2023b). With the advancement of trackers, the tracking benchmark plays a fundamental role in performance evaluation (Wang et al., 2022a). Several widely used benchmarks such as LaSOT (Fan et al., 2019), TrackingNet (Muller et al., 2018), and LasHeR (Li et al., 2022a) have been released for evaluating trackers and promoting the development of visual tracking.

Satellite video (SV) is a valuable surface observation data that provides a wealth of static and dynamic information on specific areas (Feng et al., 2021). In 2013, the SkySat-1 (SS) satellite captured a panchromatic video with a ground sample distance (GSD) of 1.1 m and a frame rate of 30 frames per second (FPS). In 2016, the International Space Station (ISS) captured a 3 FPS red–green–blue (RGB) video with a GSD of 1.0 m. The Jilin-1 (JL) satellite constellation can capture 30 FPS RGB video sequences with a 0.92 m GSD. Recently, the Luojia-3–01 satellite was launched, which has the capability of multi-mode optical imaging, intelligent processing in orbit, and real-time transmission in star-to-earth and star-to-star communication. Table 1 presents the detailed configurations and parameters of some video satellites, and Fig. 1 shows the sample frames corresponding to these satellites. The emergence of SV data enhances remote sensing observation capabilities and facilitates the visual tracking community (Wu et al., 2022). SOT in SV has promising applications in intelligent traffic surveillance and analysis (Du et al., 2018), etc. As mentioned above, remarkable advances have been

* Corresponding author.
E-mail address: yqtang@csu.edu.cn (Y. Tang).

https://doi.org/10.1016/j.isprsjprs.2024.03.013
Received 6 October 2023; Received in revised form 29 January 2024; Accepted 15 March 2024
Available online 25 March 2024

Table 1
Detailed configurations and parameters of several video satellites.

Satellite platform	Early launch year	GSD	Imaging color	Duration	FPS	Imaging area
SkySat-1	2013	1.1 m	Pan	90 s	30	2 km × 1.1 km
ISS	2014	1 m	RGB	60 s	3	≦4.1 km × 2.2 km
SkySat-2	2014	1.1 m	Pan	90 s	30	2 km × 1.1 km
TT-2	2014	≧5m	Pan	≦180	25	3.1 km × 2.4 km
Jilin-1	2015	≧0.92 m	RGB	120	≦20	11 km × 4.6 km
OVS-1	2017	1.98 m	RGB	≦90	20	8.1 km × 6.1 km
Carbonite-2	2018	1.2 m	RGB	120	≦25	5.9 km × 5.9 km
OVS-2	2018	0.9 m	RGB	≦120	25	4.5 km × 2.7 km
Luojia-3–01	2023	0.7 m	RGB	—	—	—

RGB denotes the red–green–blue. Pan denotes the panchromatic. GSD denotes the ground sample distance. FPS denotes the frames per second.

made in the SOT of generic video (GV). GV can be captured by commonly used devices such as closed-circuit televisions and unmanned aerial vehicles (UAVs) (Wang et al., 2020). In contrast, progress in SV object tracking still lags far behind that of GV due to the lack of well-annotated benchmark datasets and evaluation protocols. It is also difficult to achieve accurate and robust tracking due to the following challenges:

- When it comes to SOT in SV, existing high-quality and public datasets and benchmarks are insufficient. There are rarely available datasets with oriented bounding box (OBB) annotations for single object tracking, which are essential for accurately tracking oriented objects. Additionally, it is fundamental to measure the performance of trackers comprehensively and fairly, particularly for OBB annotations with various sizes and uneven aspect ratios.
- SV typically contains three bands (i.e., red, green, and blue), which results in limited spectral features of objects, as shown in Fig. 1. Furthermore, moving objects are often small and occupy a few pixels resulting in limited spatial features such as context and texture. This can make it difficult to accurately estimate the object state, as demonstrated in Fig. 2.
- SV is photographed by the high-speed moving satellite platform. Accompanied by non-stationary and complex backgrounds, small objects are susceptible to abnormal interferences such as similar appearance, partial occlusion, motion blur, and background clutters, as shown in Fig. 3.

This article establishes the first available oriented object tracking benchmark (OOTB) for SOT in SV. OOTB includes 110 sequences with a total of 29,890 frames, covering common object categories. Moreover, a high-precision evaluation protocol is proposed to achieve comprehensive and fair comparisons of trackers. We also benchmark 33 state-of-the-art (SOTA) trackers with a total of 58 models covering different features, backbones, and tracker tags to help understand their performance and offer baseline results for future research. Extensive comparisons and analysis demonstrate that SV object tracking remains challenging in the visual tracking community. The major contributions are summarized as follows:

- We provide a comprehensive and detailed review of various tracking paradigms and frameworks, covering both the general video and satellite video domains. We also present relevant single object tracking benchmarks for generic and specific applications.
- We construct an oriented object tracking benchmark OOTB. To the best of our knowledge, OOTB is the first available oriented benchmark dedicated to SOT in SV. It consists of 110 sequences totaling 29,890 frames and covers common object categories including car, ship, plane, and train. All sequences are manually annotated with high-quality oriented bounding boxes and labeled with 12 fine-grained attributes, making them an invaluable resource for relevant research. In addition, we propose a high-precision evaluation protocol for fair comparisons between trackers.
- We benchmark 33 SOTA trackers with a total of 58 models, covering various tracking paradigms and application scenarios. Moreover, the in-depth comparison and analysis are conducted to provide baseline results for further research. In light of the advances in trackers, several insightful thoughts are also drawn to point out promising prospects in the SV tracking domain.

The rest of this article is organized as follows. Sections 2-4 give a comprehensive review including GV trackers, SV trackers, and benchmark datasets. The proposed OOTB and compared trackers are introduced in Section 5. Experimental results and analysis are presented in Section 6. In Section 7, we provide several thoughts and insights for future research. Finally, we conclude this article and summarize the contributions.

2. Review on generic video trackers

Depending on the data acquisition platform, SOT can be divided into GV and SV tracking domains. In this article, we provide a comprehensive overview of trackers and techniques including advancements, challenges, and limitations in GV and SV object tracking. We also outline relevant benchmarks that are commonly used to evaluate the performance of SOT trackers.

Typically, SOT trackers can be classified into two categories: generative paradigm and discriminative paradigm (You et al., 2019). The former constructs a model to represent the object and finds an object region that is similar to the description of the generative model by classifying the signal and minimizing the objective loss. Object representation models such as Gaussian mixed model (Jepson et al., 2003), kernel trick (Han et al., 2008), and sparse representation (Wright et al., 2010) can directly affect the accuracy and speed of tracking methods in the generative paradigm. On the other hand, the discriminative paradigm jointly trains foreground and background regions to discriminate the object, which improves the tracking robustness. Its simplicity and strong performance have made it a fundamental paradigm for tracking in recent decades (You et al., 2019). Discriminative correlation filter (DCF) and Siamese neural network (SNN), two of the best-performing discriminative paradigms, have proven their advancement and are dominating the SOT domain (Javed et al., 2022).

Moreover, other paradigms, such as Transformer, recurrent neural network (RNN), generative adversarial network (GAN), and traditional convolutional neural network (CNN), have also achieved satisfactory results in the tracking community (Marvasti-Zadeh et al., 2022). Next, we provide a comprehensive overview of these paradigms.

2.1. DCF for SOT

Over the last decade, DCFs have proved their high performance and efficiency on various benchmarks (Javed et al., 2022). The DCF learns a filter by minimizing a least-squares error to determine the object's position and updates the model to adapt to object changes during tracking. In the following, we review the DCF in terms of discriminative object representations, adaptive scale estimation, and handling boundary effects.

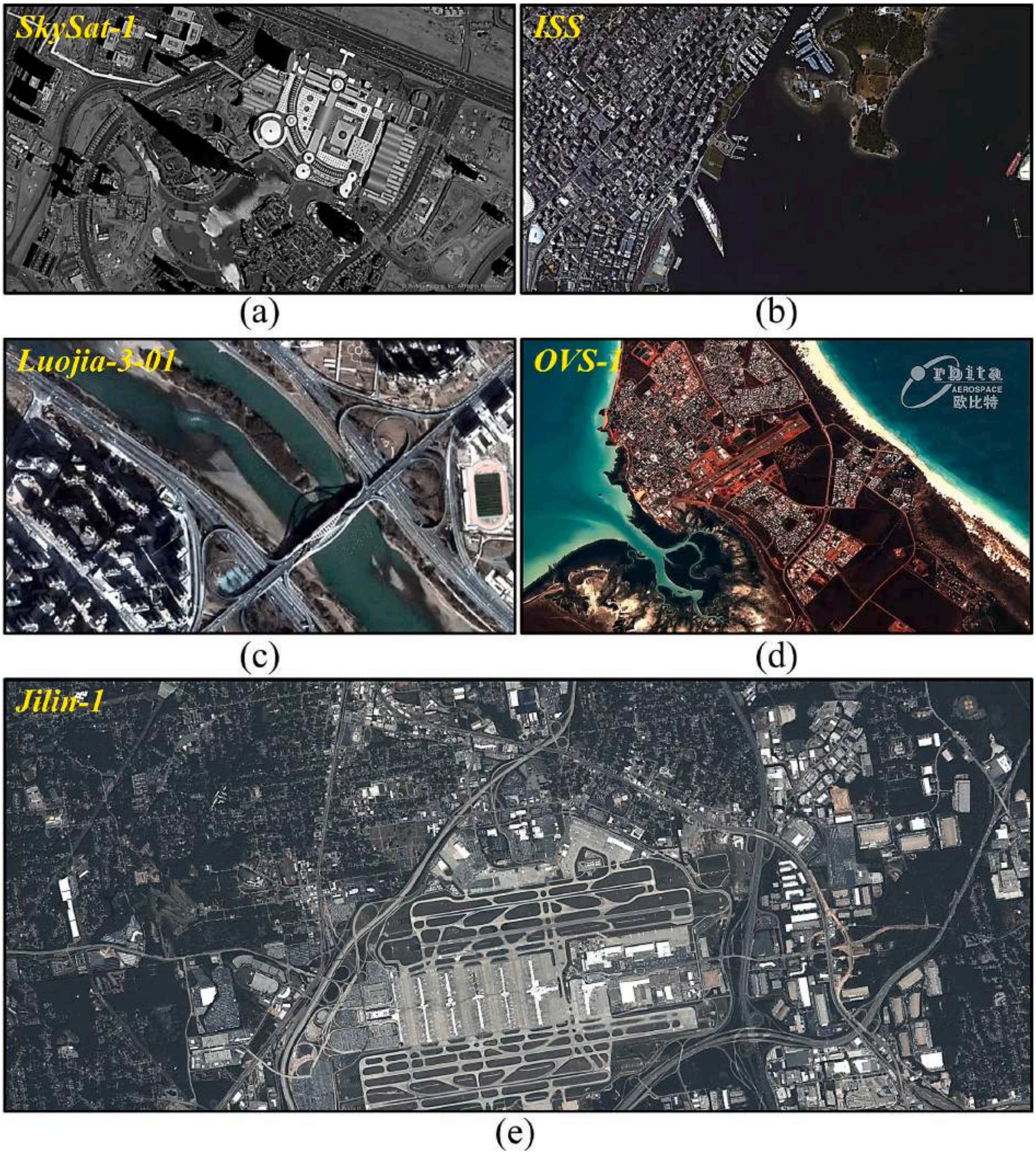

Fig. 1. Sample frames from video satellites. (a), (b), (c), (d), and (e) are captured by the SkySat-1, ISS, Luojia-3–01, OVS-1, and Jilin-1, respectively.

2.1.1. Discriminative object representations

DCF-based trackers have been a highlight since the introduction of MOSSE (Bolme et al., 2010). Subsequently, CSK (Henriques et al., 2012) modeled after MOSSE introduces the circular matrix and kernel trick to improve tracking performance. Both MOSSE and CSK use intensity or raw-pixel features to represent objects.

Considering the limitations of intensity or raw pixel features, several color-based features, e.g., local color, color histogram (CH), and color name have also been explored for enhanced object representations. Representative DCF-based trackers are LCT (Ma et al., 2015b), CN (Danelljan et al., 2014), DAT (Possegger et al., 2015), and Staple (Bertinetto et al., 2016a). The CN tracker extends the training mechanism of CSK to multi-channel color name features and proposes an adaptive dimensionality reduction method, which proves the significance of seriously selecting color transformations. Color name features have been employed in lots of DCF-based trackers such as CSR-DCF (Lukezic et al., 2018), ECO-HC (Danelljan et al., 2017a), ARCF (Huang et al., 2019), GFS-DCF (Xu et al., 2019), and AutoTrack (Li et al., 2020). Another powerful hand-crafted feature is the histogram of oriented gradients (HOG) that is initially designed for pedestrian detection (Dalal and Triggs, 2005). Due to its advantages in capturing contours and remaining intrinsic illumination invariance, KCF (Henriques et al., 2015) extends CSK with multi-channel HOG features and introduces multiple kernel functions to train more discriminative classifiers.

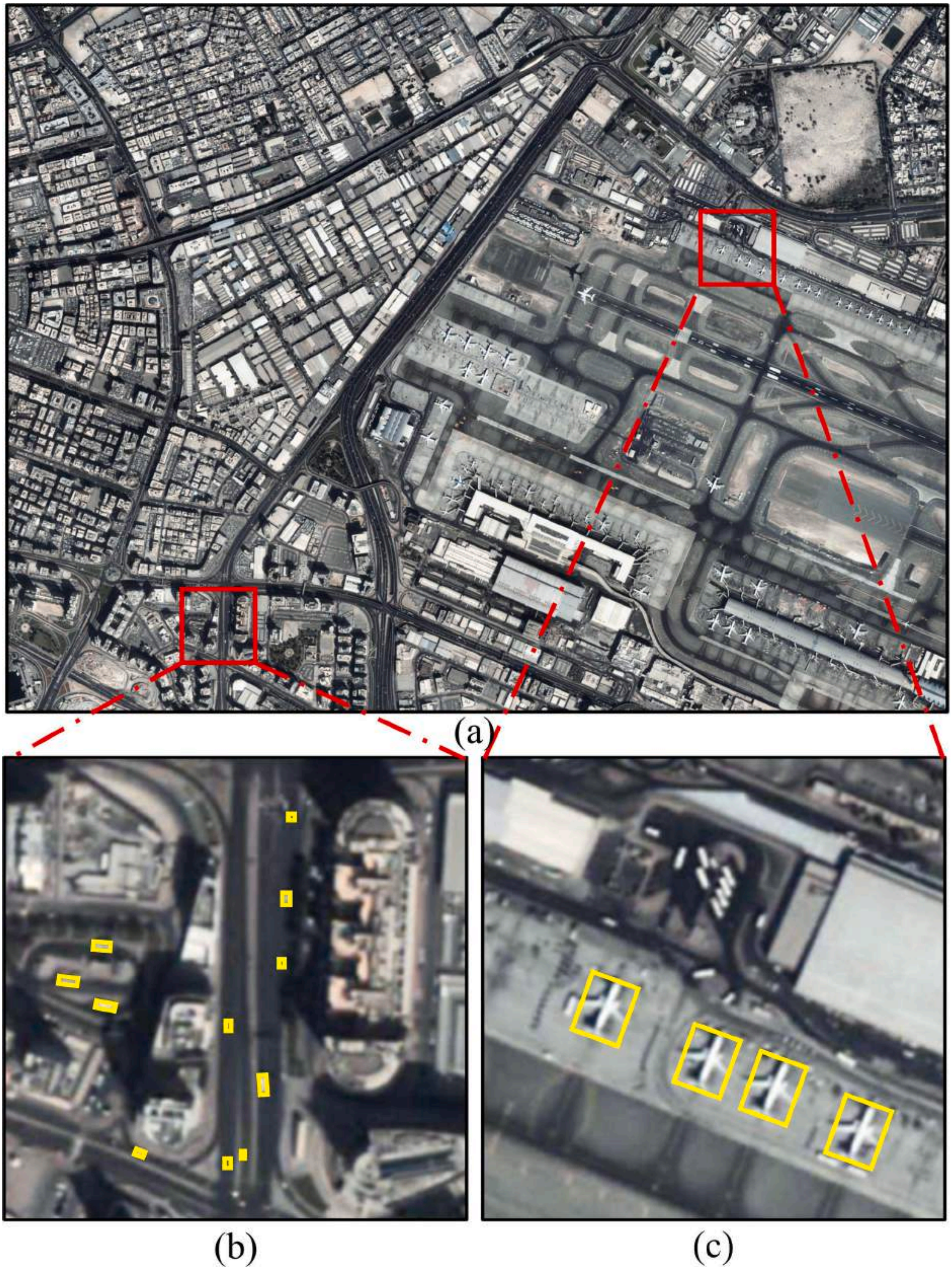

Fig. 2. Visual examples of SV objects. (a) shows the original frame, while (b) and (c) are two zoomed-in regions. (b) and (c) display the car and plane categories, marked by yellow boxes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

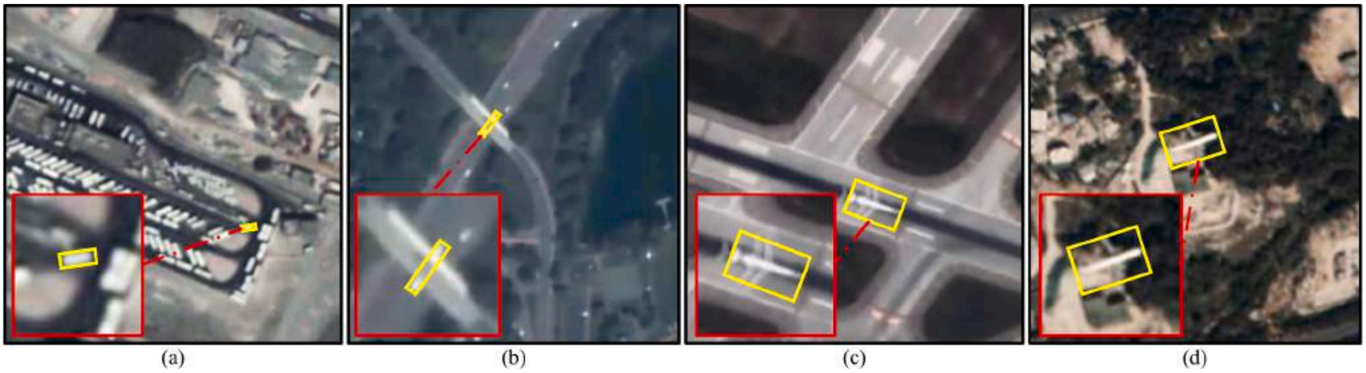

Fig. 3. Visualization of several abnormal interferences. In each example, an object is marked with the yellow OBB. (a) reflects the similar appearance, where the tracked object is surrounded by many similar objects. (b) shows the partial occlusion, where the object is partially occluded by the bridge. (c) shows the motion blur, where the object has a residual shadow. (d) reflects the background clutters, where the background has a similar color or texture to that of the object. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Inspired by KCF, HOG features have also been integrated into many DCF-based trackers for enhanced object representation, such as CFLB (Galoogahi et al., 2015), BACF (Galoogahi et al., 2017b), SAMF (Li and Zhu, 2015), SRDCF (Danelljan et al., 2015), CSR-DCF (Lukezic et al., 2018), AutoTrack (Li et al., 2020), and GFS-DCF (Xu et al., 2019). However, HOG is sensitive to deformation as it relies heavily on the spatial layout of the object (Dalal and Triggs, 2005). Therefore, combining complementary features to cope with multiple challenges is emerging. For instance, Staple combines HOG and CH with a ridge regression framework, achieving robust tracking under color change, illumination variation, and deformation (Bertinetto et al., 2016a). Multiple feature fusion strategy has been validated and adopted by several DCF-based trackers such as SAMF (Li and Zhu, 2015), Staple (Bertinetto et al., 2016a), STRCF (Li et al., 2018b), MCCT (Wang et al., 2018), AutoTrack (Li et al., 2020), GFS-DCF (Xu et al., 2019), and LDES (Li et al., 2019c).

Encouraged by recent advances in deep learning, an increasing number of DCF-based trackers utilize deep convolutional neural networks (CNNs) that are suitable for image processing tasks (Ye et al., 2022). Shallow CNN features comprise low-level information with high spatial resolution, suitable for accurate object localization, while deep features encode high-level semantic information with low resolution, inherently invariant to appropriate object changes. HCF (Ma et al., 2015a) tracker, as one of the earliest DCF-based trackers to use CNN features, explores features of different dimensions to represent objects and trains a multi-resolution filter to locate the object in a coarse-to-fine fashion. Other DCF-based trackers such as CFNet (Valmadre et al., 2017), DeepSRDCF (Danelljan et al., 2015a), ECO (Danelljan et al., 2017a), ASRCF (Dai et al., 2019), ARCF (Huang et al., 2019), ATOM (Danelljan et al., 2019b), DiMP (Bhat et al., 2019), and PrDiMP (Danelljan et al., 2020) have also demonstrated the effectiveness of CNN features on various benchmarks, paving the way for exploring more sophisticated trackers.

2.1.2. Adaptive scale estimation

Blooming features greatly contribute to the accuracy and robustness of DCF-based trackers. However, the tracked object usually suffers from position and scale changes. Standard DCF-based trackers use a fixed-size template and are unable to handle scale changes, leading to severe tracking drifts (Javed et al., 2021). Towards this end, several strategies have been investigated for accurate scale estimation. SAMF (Li and Zhu, 2015) defines a scaling pool that acquires multi-resolution scaling patches to estimate the position and scale of objects. Nevertheless, this method is computationally expensive. Considering small and moderate scale variations in neighboring frames, DSST (Danelljan et al., 2017) first estimates the object position using a two-dimensional filter and then uses a one-dimensional filter for scale estimation. Due to its efficiency and effectiveness, this strategy has been utilized in various trackers, e.g., BACF (Galoogahi et al., 2017b), LCT (Ma et al., 2015b), Staple (Bertinetto et al., 2016a), CACF (Mueller et al., 2017), CSR-DCF (Lukezic et al., 2018), and MCCT (Wang et al., 2018). In recent SOTA trackers, the deep bounding box regression approach has shown appealing results without manually setting the scale estimation parameters. It has become a universal component in DCF-based trackers such as DiMP (Bhat et al., 2019), ATOM (Danelljan et al., 2019b), PrDiMP (Danelljan et al., 2020), KYS (Bhat et al., 2020), and KeepTrack (Mayer et al., 2021).

2.1.3. Handling boundary effects

In the evolution of DCF-based trackers, the boundary effect caused by the periodic assumption of training samples is a stubborn stumbling block that severely limits the search region and degrades the discrimination capability of models (Wang et al., 2022a). Several solutions have been proposed to overcome this issue in numerous DCF-based trackers such as CFLB (Galoogahi et al., 2015), SRDCF (Danelljan et al., 2015), BACF (Galoogahi et al., 2017b), CSR-DCF (Lukezic et al., 2018), STRCF (Li et al., 2018b), ARCF (Huang et al., 2019), ASRCF (Dai et al., 2019), AutoTrack (Li et al., 2020), GFS-DCF (Xu et al., 2019), DRCF (Fu et al., 2020), ATOM (Danelljan et al., 2019b), and DiMP (Bhat et al., 2019). For example, CFLB (Galoogahi et al., 2015) trains filters with few samples to attenuate boundary effects. SRDCF (Danelljan et al., 2015) introduces a spatial regularization function that penalizes filter coefficients so that filters can be trained for large regions. For efficiency, BACF successfully trains a background-aware filter from real negative samples densely sampled from backgrounds. CSR-DCF (Lukezic et al., 2018) applies spatial-domain constraints to the filter to weaken the influence of boundary effects. The aforementioned trackers have made great progress in addressing boundary effects and advancing DCF development. With continuous improvements, SOTA DCF-based trackers such as ATOM (Danelljan et al., 2019b) and DiMP (Bhat et al., 2019) can circumvent the issue by directly learning a filter in the spatial domain.

2.2. SNN for SOT

Conventionally, SNN-based trackers consist of two branches: the template branch and the candidate branch. The template branch takes as input the image patch of the first frame or previous frames, while the candidate branch receives the image patches of the subsequent frames. Both branches share a CNN trained from massive sample pairs to ensure that the same transformation is imposed on these two branches (Javed et al., 2022). Due to its superior performance and efficiency, SNN has aroused extensive attention in the visual tracking community. Given a large number of training sample pairs, the SNN-based tracker is capable of learning the general relationship between object appearance and motion and also locating unseen objects in the training set. The primary objective of SNN-based trackers is to overcome the limitations of pre-trained deep neural networks and fully leverage end-to-end training for real-time applications (Marvasti-Zadeh et al., 2022). In this section, we review the evolution of SNN-based trackers from discriminative object representation, adaptive scale estimation, and balancing training data.

2.2.1. Discriminative object representation

Robust object representations are fundamental for reliable tracking, and discriminative object models rely heavily on the backbone network. One of the pioneers of SNN-based trackers, SiamFC (Bertinetto et al., 2016), has fine-tuned the pre-trained AlexNet (Krizhevsky et al., 2017) parameters for visual tracking, and experimental results have shown its superiority over the DCF-based trackers at that time. Many SNN-based trackers, e.g., GOTURN (Held et al., 2016), SINT (Tao et al., 2016), SiamRPN (Li et al., 2018a), DaSiamRPN (Zhu et al., 2018), C-RPN (Fan et al., 2019b), and SA-Siam (He et al., 2018), also integrate AlexNet as a feature extractor. To achieve better results, SiamRPN introduces the regional proposal network (RPN) (Ren et al., 2017) for proposal generation. However, the AlexNet structure is relatively shallow, making it difficult to extract stronger and more powerful features when compared to deeper networks such as VGGNet (Chatfield et al., 2014; Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2015), ResNet (He et al., 2016), and ResNeXt (Xie et al., 2017). Therefore, exploring how to exploit a deeper and wider network as a backbone is crucial for enhancing discrimination. SiamRPN++ (Li et al., 2019a) driven by ResNet breaks the restriction of translational invariance, enabling more accurate and robust tracking. SiamDW (Zhang, 2019b) uses deeper and wider backbones including VGGNet, ResNet, and Inception. Both SiamRPN++ and SiamDW have proven their superiority on various benchmarks. Building on previous work, SOTA trackers such as SiamMask (Wang et al., 2019c), D3S (Lukezic et al., 2020), SiamGAT (Guo et al., 2021), TransT (Chen et al., 2021), KeepTrack (Mayer et al., 2021), Stark (Yan et al., 2021a), SiamCAR (Cui et al., 2022a), and AiATrack (Gao et al., 2022) are continuously exploring the potential of deeper and wider backbones. However, these SOTA trackers are always dependent on hand-crafted models. To address the issue, LightTrack (Yan et al., 2021) embeds automatically designed lightweight models using the neural architecture search (NAS) (Pham et al., 2018) therefore performing effectively and efficiently. Such customized models can bridge the gap between academia and industry and are expected to identify forward-looking directions in the coming years.

2.2.2. Adaptive scale estimation

Similar to DCF-based trackers, SNN-based trackers also encounter object scale variations. In the early years, SNN-based trackers use common multiple-resolution scale search approaches to handle scale variations. For example, SiamFC searches for multiple scales in the forward pass by integrating a mini-batch of scaled patches. Due to its simplicity, this approach has been utilized in several SNN-based trackers such as SA-Siam (He et al., 2018), TADT (Li et al., 2019b), UDT (Wang et al., 2019b), and FlowTrack (Zhu et al., 2018b). However, the multiple-resolution scale search approach suffers from expensive computational costs. Inspired by scale estimation in object detection, Siamese trackers introduce RPN to predict region proposals with relatively well-defined object scales and aspect ratios (Ren et al., 2017). The RPN consists of a classification network for estimating foreground-background and a regression network for refining anchor boxes. SiamRPN attaches an RPN to the Siamese network to extract region proposals, discarding the time-consuming multi-scale search. Experimental results have shown the superiority of RPN in accuracy and efficiency. RPN has become a fundamental anchor-based bounding box regression approach among various SNN-based trackers, e.g., SiamRPN++ (Li et al., 2019a), DaSiamRPN, SiamMask (Wang et al., 2019c), SiamDW, SPM (Wang et al., 2019a), C-RPN (Fan et al., 2019b), and SwinTrack (Lin et al., 2022). The anchor-free bounding box regression approach is popular in the detection community due to its simplified structure and lack of dependence on hyperparameters (Cui et al., 2022a). Typically, there are two types of bounding box solutions for anchor-free methods, i.e., center-based (Tian et al., 2019) and keypoint-based (Law and Deng, 2020) algorithms. The former directly estimates the object's center and the distance from the center to the boundary. The latter (i.e., the keypoint-based approach) detects the top-left and bottom-right corner positions to form a bounding box. Motivated by the center-based detection strategy, Ocean (Zhang et al., 2020) implements the prediction of the object position and scale in an anchor-free fashion. SNN-based trackers such as SiamBAN (Chen et al., 2020), SiamCAR (Cui et al., 2022a), SiamFC++ (Xu et al., 2020), Stark (Yan et al., 2021a), ODTrack (Zheng et al., 2024), and MixFormer (Cui et al., 2024) have also inherited anchor-free bounding box regression, which is expected to be a popular alternative.

2.2.3. Balancing training data

Training data is crucial for improving model robustness. Some large-scale datasets such as ALOV300++ (Smeulders et al., 2014), MSCOCO (Lin et al., 2014), ILSVRC-DET (Russakovsky et al., 2015), ILSVRC-VID (Russakovsky et al., 2015), NUS-PRO (Li et al., 2016), UAV123 (Mueller et al., 2016), YouTube-VOS (Xu et al., 2018), TrackingNet (Muller et al., 2018), GOT-10k (Huang et al., 2021), and LaSOT (Fan et al., 2019) have been used to train SNN-based trackers offline. However, there are far fewer positive samples than negative samples in offline training. The imbalanced distribution of training samples may seriously affect the discriminative ability of the model. To this end, several strategies have been investigated in Siamese trackers. DaSiamRPN integrates hard negative sampling to introduce more semantic negative sample pairs from the same and different categories. This strategy allows DaSiamRPN to focus on fine-grained object representations, attenuating tracking drifts. C-RPN (Fan et al., 2019b) cascades a sequence of RPNs to stimulate hard negative sampling and progressively refine bounding boxes. Other trackers, such as ATOM (Danelljan et al., 2019b), TADT (Li et al., 2019b), and UDT (Wang et al., 2019b), inherited the correlation filter in the Siamese structure, strive to balance the training data and achieve competitive performance.

2.3. Transformer for SOT

Transformer (Vaswani et al., 2017) is an architecture that transforms one sequence into another using attention-based encoders and decoders. Recently, Transformer trackers have made remarkable progress, which can be classified into two categories including CNN-Transformer trackers and Fully-Transformer trackers (Kugarajeevan et al., 2023).

The former inherits the SNN paradigm and partially uses the Transformer architecture. More concretely, the CNN-Transformer tracker utilizes CNN, such as AlexNet (Krizhevsky et al., 2017), ResNet (He et al., 2016), and ShuffleNetV2 (Ma et al., 2018), to extract deep features of the template and candidate regions, followed by using the Transformer mechanism to achieve feature interactions. Finally, the prediction head receives features generated by the Transformer for localization. Benefiting from the Transformer architecture, CNN-transformer trackers can capture the non-linear interactions between the template and candidate regions, resulting in superior tracking performance, as demonstrated in TransT (Chen et al., 2021), Stark (Yan et al., 2021a), TrDiMP (Wang et al., 2021), AiATrack (Gao et al., 2022), SiamTPN (Xing et al., 2022), ToMP (Mayer et al., 2022), etc.

However, they still rely on CNN for feature extraction, which uses the local convolution kernel to capture features. Therefore, it is difficult for CNN-Transformer trackers to capture global feature representations. To address this issue, the latter (i.e., the Fully-Transformer tracker) has been developed, which can be further categorized into the two-stream two-stage paradigm and one-stream one-stage paradigm. The two-

Table 2
Characteristics and experiments of some SV trackers. These trackers are listed in chronological order.

Tracker	Exploited features @ Tracker prototype	RTFO	PC (CPU, RAM, Nvidia GPU)	Data source	Tracked object			Tracking performance							
					Category	NoSO	Output	Benchmark	AO/AUC_S (%)	AUC_P (%)	EAO	A	R	FPS	
														CPU	GPU
KCF_TFD (Du et al., 2018)	HOG + MFD @ KCF	—	Intel I5 2.8 GHz CPU, 8 GB RAM	JL, ISS	C, T	3	HBB	OTB	56.0 %	76.0 %	—	—	—	9.0	—
MOFT (Du et al., 2019)	OF @ II	—	Intel I7-3770 3.4 GHz CPU, 32 GB RAM	JL, ISS	C, P, T	5	HBB	OTB	52.0–86.8 %	90.6–96.2 %	—	—	—	19.3–66.7	—
CFKF (Guo et al., 2019)	HOG + I + PM @ DSST	APCE + KF	3.50 GHz CPU	SS, JL	C	31	OBB	VOT	70.53 %	—	0.7205	0.71	0	1094.7	—
PASiam (Shao et al., 2019a)	DA + BS + PM @ SiamFC	GMM + KF	—	JL, ISS	C, T	3	HBB	OTB	85.6 %	94.6 %	—	—	—	—	54.8
HKCF (Shao et al., 2019b)	HOG + OF @ KCF	—	—	JL, ISS	C, P, T	6	HBB	OTB	80.9 %	95.2 %	—	—	—	138.3	—
VCF (Shao et al., 2019c)	OF @ KCF	—	—	JL, ISS	P, T	3	HBB	OTB	80.2 %	94.1 %	—	—	—	122.8	—
CRAM (Hu et al., 2020)	DA + DM @ (Danelljan et al., 2019)	—	3.5 GHz CPU, GTX 1080 GPU	JL, SS	C	31	HBB	VOT	70.0 %	—	0.7286	0.70	0	—	17.5
WTIC (Wang et al., 2020)	G + PM @ CSK	TCMI + MS	Intel I5 CPU, 16 GB RAM	JL	C	9	HBB	OTB	61.1 % (0.5)	95.4 % (3)	—	—	—	—	—
CFME (Xuan et al., 2020)	HOG + PM @ KCF	PV + KF + MTA	Intel Xeon-E5-2620v3 2.4 GHz CPU	JL	C, P	13	HBB	OTB	69.3–72.9 %	66.2–96.4 % (5)	—	—	—	102.0–123.0	—
VAASN (Bi et al., 2021)	DA @ SiamFC	—	Intel I7-7800X 3.5 GHz CPU, Titan V GPU	JL, SS,	C, S, P	80	OBB	OTB	55.6 % (0.5)	68.8 % (20)	—	—	—		75.0
CF_FFMC (Liu et al., 2021)	HOG + LBP + I + PM @ KCF	APCE + KF	Intel Xeon E3-1240v5 3.50 GHz CPU	JL	C, S, P	>10	HBB	OTB	73.8–84.2 %	87.4–99.8 % (4)	—	—	—	—	—
HRSiam (Shao et al., 2021)	DA + BS @ SiamRPN	—	GTX 1080Ti GPU	JL, ISS	C, P, T	6	HBB	OTB	79.8 %	95.2 %	—	—	—	—	31.1
RACF (Xuan et al., 2021)	HOG @ KCF	—	Intel I7-6700 k 3.4 GHz CPU	JL	C, P	6	HBB	OTB	70.71 %	99.84 % (5)	—	—	—	34.0	—
ID-DSN (Zhu et al., 2021)	DA + PM @ SiamRPN	—	GTX 1070Ti GPU, 8 GB RAM	JL, ISS	C, S, P, T	6	HBB	OTB	69.4 %	92.7 %	—	—	—	—	32.1
HMTS (Chen et al., 2022b)	I + CN + PM @ KCF	PV + KF	2.4 GHz CPU	JL, SS*, CB*	C	65	HBB	OTB	43.44 %	72.53 % (5)	—	—	—	—	—
DF (Chen. et al., 2022)	HOG + CN + GCS + PM @ Staple	SAI + KF + NR	Intel Xeon 3.2 GHz CPU, RTX 2080Ti GPU	JL	C	14	HBB	OTB	70.0 %	91.2 %	—	—	—	155.2	—
RAMC (Chen et al., 2022c)	HOG + OF @ KCF	—	Intel Xeon 3.2 GHz CPU, RTX 2080Ti GPU	JL, ISS	C, S, P, T	8	OBB	OTB	78.5 %	94.6 %	—	—	—	42.5	—
STRCF_IMM (Li and Bian, 2022)	HOG + CN + PM @ STRCF	PV + APCE + IMM	Intel I7-10875H 2.3 GHz CPU	JL, SS*, CB*	C, S, P	25	HBB	OTB	56.5 %	87.9 % (20)	—	—	—	35.1	—
CPKF (Li et al., 2022b)	HOG + CN + GCS + PM @ STRCF	PV + APCE + KF + CPF	Intel I7-10875H 2.3 GHz CPU	JL, SS, CB	C, S, P, T	105	HBB	OTB	47.1 %	68.1 % (20)	—	—	—	25.4	—
CF_MFMC (Liu et al., 2022)	DA + HOG + PM @ KCF	SDM + AKF	Intel Xeon E3-1240v5 3.5 GHz CPU, GTX 1080Ti GPU	JL	C, P	8	HBB	OTB	69.6–79.8 %	82.0–99.0 %	—	—	—	14.0–28.0	—
SRN-TFM (Ruan et al., 2022)	DA + DM + PM @ CRAM	PSR + TFM	Tesla K80 GPU	JL	C, S, P	74	HBB	VOT	60.0 %	—	0.5513	0.623	0.06	—	10.2
JSANet (Song et al., 2022)	DA @ SiamRPN	—	Tesla V100 GPUs	JL, ISS	C	309	HBB	OTB	50.91–55.24 % (0.5)	81.50–85.07 % (20)	—	—	—	—	33.1

(continued on next page)

Table 2 (*continued*)

Tracker	Exploited features @ Tracker prototype	RTFO	PC (CPU, RAM, Nvidia GPU)	Data source	Tracked object		Output	Tracking performance						FPS	
					Category	NoSO		Benchmark	AO/AUC_S (%)	AUC_P (%)	EAO	A	R	CPU	GPU
MBLT (Zhang et al., 2022)	HOG + CN + DM @ DCF	TCMI + FCN	Intel Xeon-E5-2630v3 2.4 GHz CPU, 64 GB RAM, Titan X GPU	JL, ISS	C	40	HBB	OTB	44.01–69.51 %	75.60–96.30 %	—	—	—	42.7–46.3	—
AD-OHNet (Cui et al., 2022b)	DA, DM @ ADNet	CS + DRL	32 RAM, GTX 1080Ti GPU	JL	C, S, P	—	HBB	OTB	64.5 %	92.6 % (20)	—	—	—	—	3.0–5.5
SiamMDM (Yang et al., 2023)	DA, PM @ SiamFC	PV + MTA	RTX 3060 GPU	JL, CB	C, S, P, T	353	HBB	OTB	45.6–47.5 %	61.1–72.5 %	—	—	—	—	61.5
ThickSiam (Zhang et al., 2023)	DA + PM @ SiamFC	—	GTX1070Ti GPU	JL, ISS	C, S, P, T	12	HBB	OTB	99.1 %	75.5 %	—	—	—	—	56.8

For the Exploited features @ Tracker prototype, DA = deep appearance feature, DM = deep motion feature, PM = physical motion feature, MFD = multi-frame differencing, BS = background subtraction, I = intensity, G = gabor, CN = color name, HOG = histogram of oriented gradients, GCS = global color statistics, OF = optical flow, LBP = local binary patterns, and II = integral image technology. For the Data source, SS = SkySat-1, JL = Jilin-1, ISS = International Space Station, CB = Carbonite-2, and the symbol * denotes the possible data sources. For the RTFO (recognition and treatment of full occlusion), APCE = average peak correlation energy, KF = Kalman filter, GMM = Gaussian mixture model, TCMI = tracking status monitoring indicator, MS = motion smoothness, PV = peak value, MTA = motion trajectory averaging, SAI = state aware indicator, NR = nonlinear regression, IMM = interacting multiple model, CPF = correlation particle filter, SDM = state discriminant method, AKF = adaptive Kalman filter, PSR = peak to sidelobe ratio, TFM = trajectory fitting motion, FCN = fully convolutional network, CS = confidence score, and DRL = deep reinforcement learning. For the PC (CPU, RAM, Nvidia GPU), CPU = Central Processing Unit, RAM = Random Access Memory, and GPU = Graphics Processing Unit. For the Category, C = car, S = ship, P = plane, and T = train. NoSO = number of sequences or objects. For the AO/AUC_S (%) and AUC_P (%), the value in parentheses denotes the distance or IOU (intersection over union) threshold. EAO = expected average overlap.

stream two-stage paradigm first integrates the Transformer architecture to extract features from the template and candidate regions. Another Transformer is then used to interact and enhance the extracted features for subsequent object localization. Representative SOTAs, such as DualTFR (Xie et al., 2021), SparseTT (Fu et al., 2022), and SwinTrack (Lin et al., 2022), are all categorized as the two-stream two-stage paradigm. In contrast, the one-stream one-stage paradigm follows a single and simplified pipeline where the feature extraction, interaction, and enhancement are implemented in a unified Transformer architecture, resulting in a simpler framework with more powerful learning ability. Therefore, this paradigm is widely inherited in many SOTA trackers, e.g., OSTrack (Ye et al., 2022a), SimTrack (Chen et al., 2022a), SeqTrack (Chen et al., 2023a), ARTrack (Wei et al., 2023), VideoTrack (Xie et al., 2023), and MixFormer (Cui et al., 2024). Experimental results have validated their excellent performance on many benchmarks.

2.4. *RNN for SOT*

In addition to the above paradigms, several other architectures are also explored to improve the robustness and efficiency of visual tracking. Among them, RNN has brought new inspiration to researchers since visual tracking is strongly related to spatio-temporal flow information. RNN is good at processing sequential data, e.g., audio signals, temporal series, and texts (Fan et al., 2014; Marchi et al., 2014; Sundermeyer et al., 2012). To enhance the object modeling process and prevent online fine-tuning, RNN spends lots of time in digging out additional information, easily causing the over-fitting of models. Currently, there are few RNN-based trackers due to the complex training process accompanied by large amounts of parameters. Existing trackers mainly focus on spatio-temporal information (Chen et al., 2019; Yang et al., 2017), object-aware using multi-level attention mechanisms (Chen et al., 2019), handling background clusters using texture and structure (Ma et al., 2018a), and cooperating with the long short-term memory (LSTM), to model the object appearance in sequential frames (Yang et al., 2017).

2.5. *GAN for SOT*

GAN has unique advantages in capturing data distribution features and generating expected training samples without dense labeling, which are useful for improving tracking performance. Therefore, researchers have applied GAN to the SOT domain. For instance, VITAL (Song et al., 2018) uses GAN to address the imbalanced distribution of training samples, achieving robust effects. TGGAN (Guo et al., 2018) learns a general appearance distribution model to obtain reliable online adaptive templates. Additionally, ADT (Zhao et al., 2019) optimizes the regression and classification networks through adversarial learning, leading to appealing performance. However, GAN is relatively difficult to interpret and train and requires proper synchronization between the generator and discriminator.

2.6. *Other CNN structures for SOT*

In addition to DCF, SNN, Transformer, RNN, and GAN, trackers based on the graph neural network (GNN), e.g., SiamGAT (Guo et al., 2021) and GCT (Gao et al., 2019), and traditional CNN-based trackers, e.g., TCNN (Nam et al., 2016), MDNet (Nam and Han, 2016), and RT-MDNet (Jung et al., 2018), have also been developed for SOT. It is worth noting that the tracking paradigms described above are not standalone but draw on the strengths of multiple paradigms and model structures to enhance their capabilities. For instance, the Transformer tracker AiATrack is prototyped on the SNN for achieving a remarkable performance. In general, SOT in GV has established a blossoming prospect in recent years. However, when these SOTAs encounter SV objects with limited features and complex backgrounds, they hardly obtain the same competitive results as GV objects. Further exploration of SV trackers is essential to broaden the visual tracking community (Li et al., 2022c).

3. Review on satellite video trackers

Several trackers have been developed for SOT in SV, achieving superior results on their home-grown datasets. Table 2 outlines the characteristics of notable trackers. A comprehensive review and analysis of SV trackers is presented across six aspects including tracker prototype, exploited features, recognition and treatment of full occlusion (RTFO), rotation estimation, data source and tracked object, and evaluation benchmark.

3.1. Tracker prototype

Table 2 reveals that many trackers inherit the tracking paradigms of GV, such as DCF, SNN, CNN, and RNN. For example, some trackers (KCF_TFD (Du et al., 2018), HKCF (Shao et al., 2019b), VCF (Shao et al., 2019c), CFME (Xuan et al., 2020), CF_FFMC (Liu et al., 2021), RACF (Xuan et al., 2021), HMTS (Chen et al., 2022b), RAMC (Chen et al., 2022c), and CF_MFMC (Liu et al., 2022)) are based on the KCF framework, while WTIC (Wang et al., 2020), CFKF (Guo et al., 2019), DF (Chen. et al., 2022), and CPKF (Li et al., 2022b) are modeled on the CSK, DSST (Danelljan et al., 2017), Staple, and STRCF, respectively, to achieve competitive speed. Some trackers, such as PASiam (Shao et al., 2019a), VAASN (Bi et al., 2021), and ThickSiam (Zhang et al., 2023), inherit the SiamFC, while HRSiam (Shao et al., 2021), ID-DSN (Zhu et al., 2021), and JSANet (Song et al., 2022) inherit the SiamRPN. These SNN-based trackers strike a balance between tracking accuracy and speed. AD-OHNet (Cui et al., 2022b) is modeled on ADNet (Yun et al., 2018), which uses CNN to extract discriminative features of objects, while CRAM (Hu et al., 2020) uses convolutional regression networks to resolve the regression problem and applies gradient descent in an end-to-end learning fashion. It is observed that most trackers inherit DCF and exploit hand-crafted features for SV object tracking, which may lead to unsatisfactory accuracy. In contrast, SNN, CNN, RNN, and Transformer-based trackers could emerge as the mainstream direction in SV tracking domain.

3.2. Exploited features

The features play a critical role in SV object tracking. Table 2 provides an overview of exploited features. These features can be broadly categorized into two types: spatial features and temporal features. Spatial features are primarily concerned with representing the appearance information by using both hand-crafted or deep features. Hand-crafted features such as HOG, color name, intensity, Gabor, CH, and local binary patterns (LBP) are commonly used to describe spatial texture and structure information. Deep appearance feature (DA) is one of the most common features, with the shallow layer containing low-level information with high spatial resolution, suitable for accurate localization. In contrast, the deep layer encodes high-level semantic information and is invariant to appropriate object changes. Hierarchical features of DA have been used in many trackers such as HRSiam (Shao et al., 2021), CRAM (Hu et al., 2020), ID-DSN (Zhu et al., 2021), and JSANet (Song et al., 2022), with excellent results.

Temporal features, on the other hand, focus on extracting inter-frame dynamic information using techniques such as multi-frame differencing (MFD), background subtraction (BS), optical flow (OF), deep motion feature (DM), and physical motion feature (PM). KCF_TFD (Du et al., 2018) fuses KCF and three-frame difference for tracking SV objects. Both PASiam (Shao et al., 2019a) and HRSiam (Shao et al., 2021) use BS features to assist with the tracking task. The OF is capable of representing inter-frame motion information and has been explored in many SV trackers such as MOFT (Du et al., 2019), HKCF (Shao et al., 2019b), VCF (Shao et al., 2019c), and RAMC (Chen et al., 2022c). Deep OF, as a common DM, has also yielded excellent tracking results in CRAM (Hu et al., 2020) and SRN-TFM (Ruan et al., 2022). Considering the relatively stable motion state of SV objects, the PM has been widely applied to most trackers such as CFKF (Guo et al., 2019), PASiam (Shao et al., 2019a), WTIC (Wang et al., 2020), CFME (Xuan et al., 2020), CF_FFMC (Liu et al., 2021), ID-DSN (Zhu et al., 2021), HMTS (Chen et al., 2022b), DF (Chen. et al., 2022), STRCF_IMM (Li and Bian, 2022), CPKF (Li et al., 2022b), CF_MFMC (Liu et al., 2022), and SRN-TFM (Ruan et al., 2022). For the PM, some methods such as the Kalman filter (KF) (Kalman, 1960), motion smoothness (MS) (Wang et al., 2020), and motion trajectory averaging (MTA) (Xuan et al., 2020), are embedded to analyze the motion trajectory. Overall, combining the spatial feature with temporal feature can be an effective way to cope with challenging attributes, and its effectiveness has been validated.

3.3. Recognition and treatment of full occlusion

In SV, full occlusion (FO) is a common and challenging attribute due to the nadir view. To correctly track the object under FO, the tracker needs to solve three sub-problems (Xuan et al., 2020):

- Occlusion awareness: A tracker needs to be aware of the occurrence of object occlusion.
- Occlusion handling: A tracker is expected to overcome full occlusion without losing the object.
- End of occlusion awareness: A tracker needs to be aware of the end of the occlusion.

As shown in Table 2, existing trackers typically recognize and handle occlusions by comparing indicators with thresholds to solve the first and third sub-problems, Common indicators used in these trackers include average peak correlation energy (APCE) (Guo et al., 2019), peak value (PV) (Li et al., 2022b), tracking status monitoring indicator (TCMI) (Wang et al., 2020), peak to sidelobe ratio (PSR) (Ruan et al., 2022), and state aware indicator (SAI) (Chen. et al., 2022). These metrics enable analysis of the tracking confidence and thus awareness of the FO. Experimental results demonstrate their effects. FO is usually accompanied by object disappearance. In this case, most trackers analyze historical motion information using traditional methods such as KF, MS, MTA, NR, interacting multiple model (IMM) (Li and Bian, 2022), and trajectory fitting motion (TFM) (Ruan et al., 2022) to predict the object state in the current frame. Additionally, some trackers use CNN methods such as fully convolutional network (FCN) (Zhang et al., 2022) and deep reinforcement learning (DRL) (Cui et al., 2022b) to analyze historical motion information and predict object states. While these methods have shown remarkable performance. However, FO remains a major challenge due to the diverse motion states of objects, such as straight-ahead movement, turning, lane changes, and overtaking. Therefore, more effective approaches are anticipated to address this issue in future studies.

3.4. Rotation estimation

Object rotation is a common phenomenon in SV, which can lead to accuracy degradation (Xuan et al., 2021). This problem has been successfully addressed by some excellent trackers, which can be categorized into two groups based on their outputs. Trackers with horizontal bounding box (HBB) outputs often experience scale changes due to object rotation. To solve this problem, some trackers such as WTIC (Wang et al., 2020), CF_FFMC (Liu et al., 2021), CF_MFMC (Liu et al., 2022), CPKF (Li et al., 2022b), DF (Chen. et al., 2022), and SRN-TFM (Ruan et al., 2022) apply rotation invariance features to represent the tracked objects. For trackers that output OBB, a series of rotation patches with specific angle pools are listed to achieve a better match with the template. This approach allows the tracker to detect angle changes between adjacent frames and obtain a more accurate semantic representation.

Table 3
Several generic and specific datasets.

Dataset	Venue	NoSO	NoF	NoCC	NoA	OD	ToA	Attribute
OTB50 (Wu et al., 2013)	CVPR 2013	51	29 K	10	11	VOT, OTB100, TC128	HBB	SV, DEF, OCC, MB, FM, OV, BC, LR, IV, IPR, OPR
OTB100 (Wu et al., 2015)	TPAMI 2015	100	58.61 K	22	11	OTB50, VOT, TC128	HBB	SV, DEF, OCC, MB, FM, OV, BC, LR, IV, IPR, OPR
NFS (Galoogahi et al., 2017a)	ICCV 2017	100	383 K	17	9	YouTube	HBB	SV, DEF, OOC, OV, FM, BC, LR, VC, IV
VOT2018-ST (Kristan et al., 2018)	ECCV 2018	60	21.35 K	24	12	OTB100, NUS-PRO, ALOV++, TC128, UAV123	OBB	OCO, SCO, ARC, AM, MOC, CM, BC, DEF, MB, SV, OCC, IV
LaSOT (Fan et al., 2019)	CVPR 2019	1400	3.52 M	70	14	YouTube, ImageNet	HBB	SV, DEF, MB, VC, FM, OV, ARC, BC, FOC, LR, CM, POC, ROT, IV
TrackingNet (Muller et al., 2018)	ECCV 2018	30.643 K	14.43 M	27	15	YouTube-BB	HBB	SV, ARC, DEF, BC, MB, FM, OV, IPR, OPR, LR, CM, FOC, POC, SIB, IV
GOT-10k (Huang et al., 2021)	TPAMI 2019	10 K	1.5 M	563	6	VOT, WordNet, ImageNet	HBB	OCC, LO, FM, ARC, SV, IV
UAV123 (Mueller et al., 2016)	ECCV 2016	123	113.476 K	9	12	VOT	HBB	FOC, BC, CM, OV, FM, LR, POC, ARC, SOB, SV, VC, IV
TOTB (Fan et al., 2021)	ICCV 2021	225	86 K	15	12	YouTube	HBB	DEF, BC, MB, ROT, SV, POC, FOC, LR, FM, OV, ARC, IV
WATB (Wang et al., 2022a)	IJCV 2022	206	203 K	8	13	Internet	HBB	OPR, LR, POC, DEF, FOC, MB, FM, CM,, SO, IPR, IV, OV, SV
LasHeR (Li et al., 2022a)	TIP 2022	1224	734.8 K	32	19	—	HBB	NO, PO, TO, HO, OV, LI, HI, AIV, LR, DEF, BC, SA, TC, MB, CM, FL, FM, SV, ARC
VISO (Yin et al., 2022)	TGRS 2022	3159	1.12 M	4	8	—	HBB	BC, CC, LR, OV, POC, FOC, SOB, MB
SatSOT (Zhao et al., 2022)	TGRS 2022	105	27.664 K	4	11	—	HBB	BC, IV, LQ, ROT, POC, FOC, TO, SOB, BJT, ARC, DEF
SV248S (Li et al., 2022c)	GRSM 2022	248	163.234 K	3	10	—	Polygon	BCH, BCL, CO, DS, IPR, IV, STO, LTO, ND, SM
XDU-BDSTU (Zhang et al., 2022)	TGRS 2022	20	13.745 K	1	7	—	HBB	BC, IV, LR, LF, SO, OCC, IPR
ThickSiam_D (Zhang et al., 2023)	GIS-RS 2023	12	5.55 K	4	6	—	HBB	SS, PO, PoV, PTBD, SD, PGFI
OOTB	Ours	110	29.89 K	4	12	SatSOT, VISO, AIR-MOT	OBB	DEF, IPR, PO, FO, IV, MB, BC, OON, SA, LT, IM, AM

NoSO = number of sequences or objects. NoF = number of frames. NoCC = number of categories or classes. NoA = number of attributes. OD = overlapped datasets. ToA = types of annotation.

RACF (Xuan et al., 2021) employs a similar strategy to address the rotation issue and proposes a method to estimate scale changes even with HBB outputs. In RAMC (Chen et al., 2022c), the rotation is decomposed into a translation solution to achieve adaptive rotation estimation of SV objects. Going forward, the spatial structure differences of sequence frames should be further explored to address unwanted rotation issue in SV tracking domain.

3.5. Data source and tracked object

Currently, video satellites are still in the developmental stage and are limited in number. As shown in Table 2, SVs are mainly provided by SS, JL, ISS, and Carbonite-2 (CB). Developed by Surrey Satellite Technology (SSTL), the CB delivers 1.2 m GSD RGB video and is capable of capturing video lasting approximately 120 s. Table 1 presents the detailed configurations and parameters of some video satellites. As shown in Table 2, most trackers use JL SVs due to their high quality. Tracked objects are mainly cars, ships, and planes because these objects are common and have a moderate aspect ratio. In contrast, train objects have a larger aspect ratio, which increases the difficulty of tracking. Therefore, the tracking of trains is very challenging and requires more attention in future research.

3.6. Evaluation benchmark

Table 2 shows that only a few trackers such as CFKF (Guo et al., 2019), CRAM (Hu et al., 2020), and SRN-TFM (Ruan et al., 2022) are evaluated via the Visual-Object-Tracking Challenge (VOT) benchmark (Kristan et al., 2018), whereas the others are evaluated via the one-pass evaluation (OPE) of the Object Tracking Benchmark (OTB) (Wu et al., 2015). The recent tracker REPS (https://github.com/YZCU/REPS) (Chen et al., 2024) also inherits the OTB. This is because the reset mechanism of the VOT may not be suitable for relatively short-term SV tracking tasks, especially in the case of frequent occlusions, dense objects, background clutters, and so on. In contrast, the OPE avoids the reset mechanism by initializing the tracker in the first frame and letting it estimate the object throughout the video. However, the OTB cannot accurately assess OBB results, and the precision score is susceptible to different objects.

4. Related datasets

Benchmark datasets are essential for the fair and standardized evaluation of trackers (Jiao et al., 2023). These benchmarks are typically categorized according to generic and specific applications (Fan et al., 2021). Table 3 provides details of some generic and specific benchmark datasets.

4.1. Generic benchmark datasets

Generic benchmark datasets usually contain a variety of objects gathered from natural scenes, such as vehicles, animals, balls, and human parts, As shown in Table 3, OTB50 (Wu et al., 2013) comprises 50 sequences with manually annotated HBBs for each frame. These videos include both color and gray sequences and are classified into 11 challenge attributes. Subsequently, OTB50 is extended to OTB100 (Wu et al., 2015), which consists of 100 videos with the same 11 attributes. NFS (Galoogahi et al., 2017a) is composed of 100 sequences with a frame rate of 240 FPS, which focuses on testing the impact of a high frame rate on tracking performance. Each sequence is labeled with nine attributes. VOT2018 (Kristan et al., 2018) contains short-term and long-term challenge splits. The VOT2018 Short-Term (VOT2018-ST) dataset

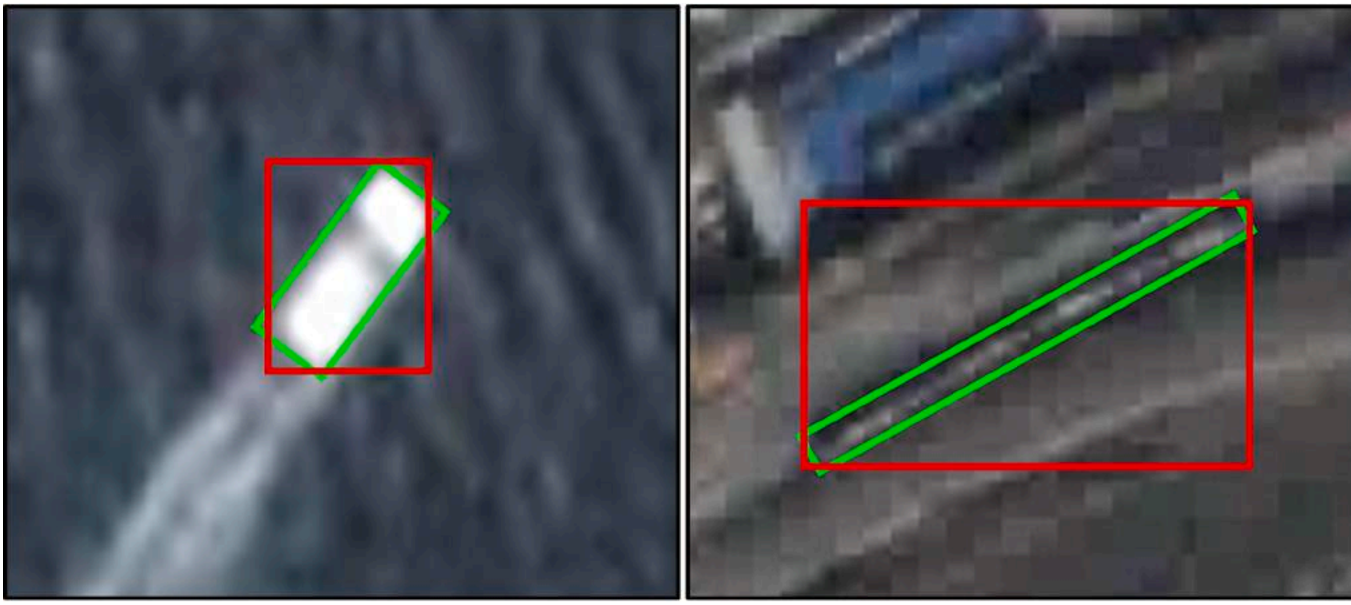

Fig. 4. Visualization of the HBB (red) and OBB (green). (a) Ship. (b) Train. Compared to HBB, OBB is more compact and suppresses the interference from the background, especially for objects with large aspect ratios and angles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

comprises 60 sequences across 24 categories. The large-scale LaSOT consists of 1,120 training sequences and 280 testing sequences, which are annotated with HBB and categorized based on 14 attributes. TrackingNet (Muller et al., 2018) provides 60,643 sequences with 30,643 for training and 511 for testing, respectively. It has over 14 million HBB annotations covering 27 different object categories. While each sequence is represented by 15 attributes. GOT-10k (Huang et al., 2021) contains about 10,000 sequences, including 9,340 for training, 420 for testing, and 180 for validation. It populates 563 moving object categories, six attributes, and 87 motion patterns.

4.2. Specific benchmark datasets

Specific benchmark datasets are used to evaluate trackers under specific applications. As shown in Table 3, UAV123 (Mueller et al., 2016) contains 123 short sequences of nine object categories filmed by professional-grade UAVs, which are similar to SV sceneries. LasHeR (Li et al., 2022a) is a large-scale and high-diversity benchmark for RGB-thermal (RGBT) tracking. It consists of 1,224 pairs of visible and thermal infrared sequences of 32 object categories with over 730 K frames, and each sequence is annotated by 19 attributes. TOTB (Fan et al., 2021) offers 225 sequences aimed at diagnosing trackers under transparent objects. VISO (Yin et al., 2022) is a large-scale dataset with a wide range of HBB annotations for various SV tasks including moving object detection, SOT, and multiple object tracking. Among them, the SOT dataset offers 3,159 tracklets with about 1.12 M frames. SatSOT (Zhao et al., 2022) pays special attention to SOT in SVs and includes 105 sequences with 27,664 HBB annotations, 11 attributes, and four categories of typical objects (i.e., car, ship, plane, and train). SV248S (Li et al., 2022c) provides 248 objects from six SVs captured by JL, with 10 attributes and three categories of objects (i.e., car, plane, and ship). It uses the tight polygon to label the object, which is particularly effective in representing plane objects with relatively complex contours. XDU-BDSTU (Zhang et al., 2022) contains 11 attributes and 20 objects from nine JL SVs, which is specially designed for vehicle tracking in SV. The object is labeled with HBB. ThickSiam_D (Zhang et al., 2023) includes 12 objects derived from eight SVs with a total of 5.55 K frames with HBB annotations. In addition, SAT-MTB (Li et al., 2023), a recent multi-task benchmark dataset, has been proposed for SV object detection, tracking, and segmentation. The proposed OOTB consists of 110 sequences covering typical object categories, such as car, ship, plane, and train, with 12 challenging attributes and a total of 29,890 frames. It is a specific dataset tailored for SOT in SV and includes a small portion of the data from (He et al., 2022; Yin et al., 2022; Zhao et al., 2022). Notably, it is the first benchmark to apply fine OBB annotations to ensure the accuracy of object scale, center, orientation, and motion direction as much as possible.

5. OOTB

5.1. Multi-platform data collection

Currently, available SVs are limited. Datasets collected by a single platform have facilitated the development of SOT. However, this can result in similar characteristics in terms of the spatial resolution, frame rate, and spectral features, thus limiting the diversity of the SV dataset. Towards this end, the OOTB dataset is sampled from multiple satellite platforms, such as JL, SS, and ISS. Moreover, we have also included part of the SV data from (He et al., 2022; Yin et al., 2022; Zhao et al., 2022). The multi-platform data would satisfy the need for dataset diversity and allow for better representation and generalization.

5.2. High-quality annotation with OBBs

A tracking dataset should be equipped with high-quality annotations. In SV, objects are typically shown in rectangles with orientation, making OBB the preferred format. Compared to HBB used in (He et al., 2022; Yin et al., 2022; Zhang et al., 2022; Zhang et al., 2023; Zhao et al., 2022), OBB provides more accurate representations, such as position, size, and orientation. Additionally, OBB is more compact and less susceptible to background interferences, especially for objects with large aspect ratios and angles, as shown in Fig. 4. Therefore, we aim to represent objects in a compatible metadata format by using OBBs, which can be easily transformed into the HBB format through batch procedures.

Specifically, the OBB description includes coordinates of the four corners. We use the roLabelImg software and zoom in 10 times for ac-

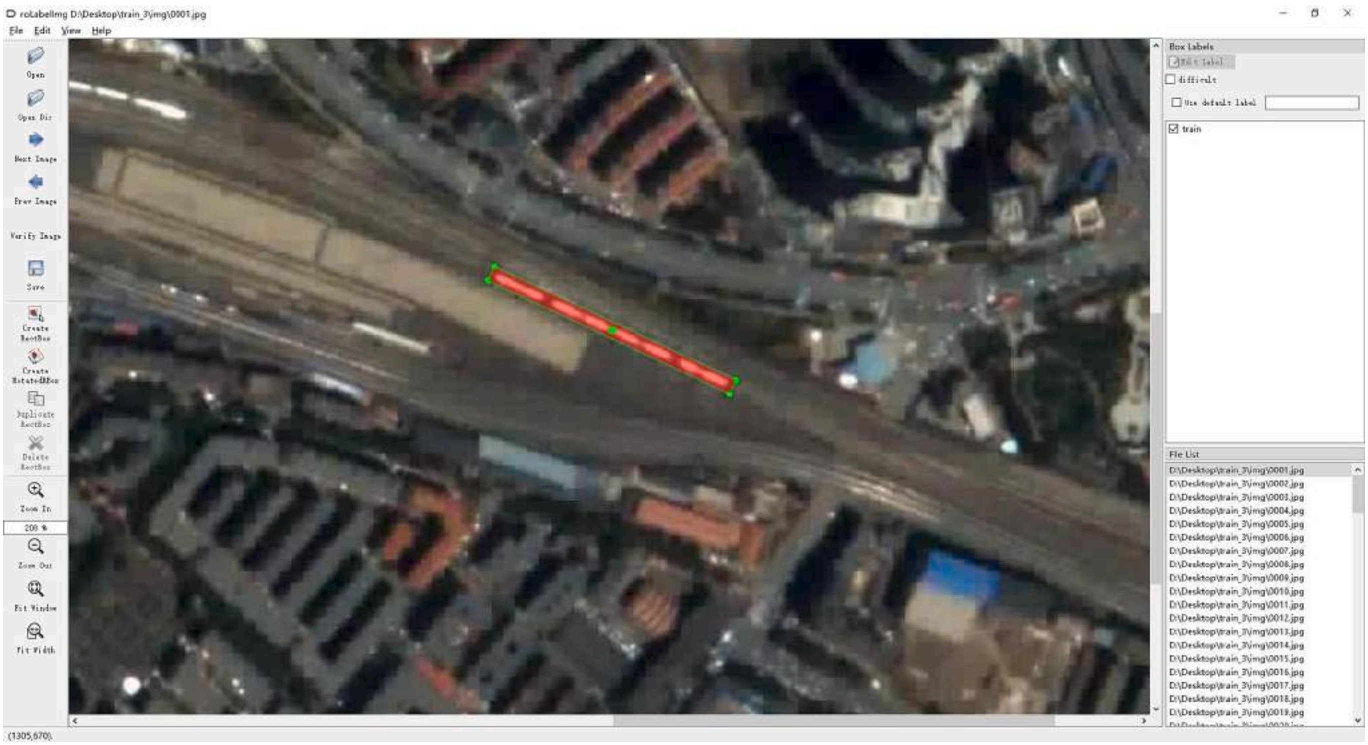

Fig. 5. Interface of the roLabelImg software.

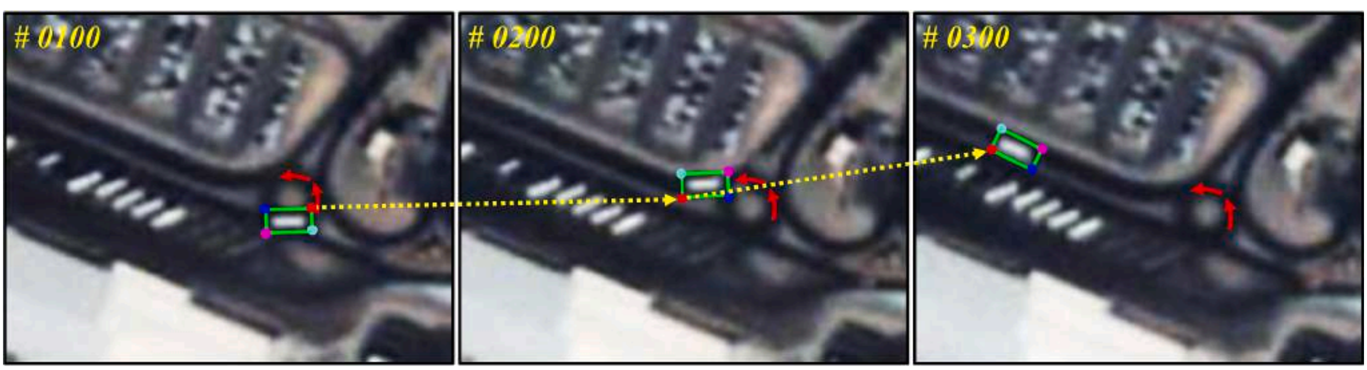

Fig. 6. Visual examples of the annotation consistency constraint. Each OBB is individually annotated because the object rotation can be seen as a linear motion of multiple points. The red arrow represents the turning trajectory. The current frame is shown in the upper left corner. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

curate annotation. The annotation example of a train is shown in Fig. 5. The labeling format of roLabelImg is $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{h}, \boldsymbol{\theta})$, where $(\boldsymbol{x}, \boldsymbol{y})$ represents the center, and $\boldsymbol{w}$, $\boldsymbol{h}$, and $\boldsymbol{\theta}$ represent the width, height, and rotation angle of the bounding box, respectively. To conform to the generic description, we transform the annotation format into float type with four decimal places for precision representation. To ensure annotation consistency, we impose annotation consistency constraints on the total annotation process, as shown in Fig. 6. In the evaluation process, we do not directly consider the direction angles, rather preferring to reflect the direction angle deviation of trackers to the metrics of the evaluation protocol. During tracking, the tracker usually combines the foreground and background information for tracking rather than just relying on that of the object itself (Javed et al., 2022). In the satellite video, the object movement is slow, and the angle change of adjacent frames is relatively slight due to the high frame rates (e.g., 25 FPS) and the long-distance satellite platform. Therefore, it is difficult for trackers to yield a large angle deviation. Each sequence is cropped and labeled by the same person to ensure a uniform annotation protocol and an expert would check and fine-tune the OBB if necessary. Each sequence is refined by at least three people. With these supervision strategies, we can guarantee high-quality OBB annotations.

5.3. Data statistics

Benefiting from the multi-platform data, we could collect prosperous SVs with different regions, spectral features, spatial resolutions, and attributes. In the following, we present detailed statistics and analysis of the OOTB.

5.3.1. Scenery type

The complexity of the scenarios contributes to the diversity of the dataset. As shown in Fig. 7, satellites can observe large areas and also produce dynamic and diverse scenes, which pose a great challenge to trackers. On the one hand, different categories of objects are usually captured in different scenes. For instance, in the first row of Fig. 7, a car drives on a crowded road, a ship sails on the sea, a plane parks at an

Fig. 7. Visualization of the scenery diversity. Each row shows four object categories, while each column shows the same category in different scenarios.

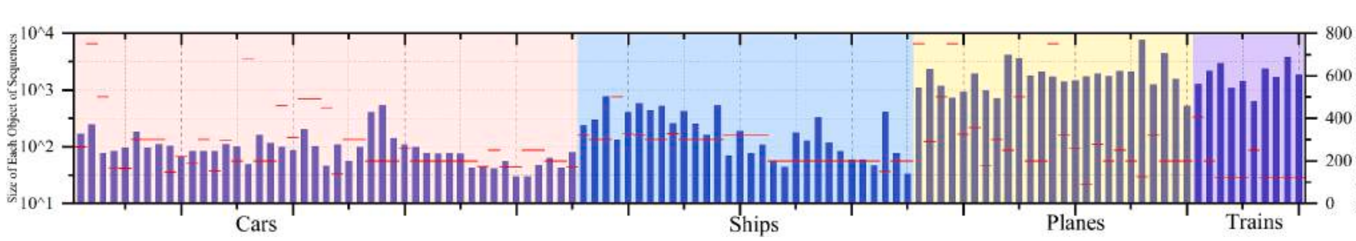

Fig. 8. Overview of the OOTB dataset. It has four object categories including 45 cars, 30 ships, 25 planes, and 10 trains. The blue bar represents the size of the object. The red line indicates the number of frames. The average sizes are 109.7, 238.7, 2075.3, and 1949.0 pixels, respectively. The average frame length of the dataset is 271.7. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

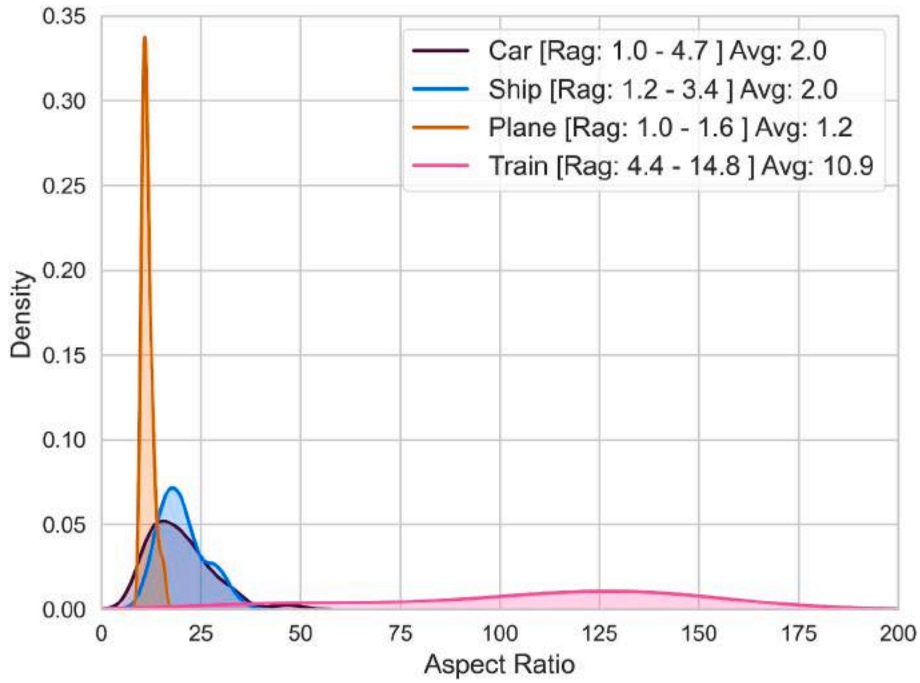

Fig. 9. The aspect ratios of the OOTB dataset. The average aspect ratio of the car and ship is 2.0, while the average aspect ratio of the plane is 1.2. The train has a aspect ratio ranging from 4.4 to 14.8. Rag denotes the range of the aspect ratio. Avg denotes the average aspect ratio.

airport, and a train moves towards the city. On the other hand, similar objects may appear in distinct scenarios. As shown in the third column of Fig. 7, the plane flies over diverse backgrounds and has different motion characteristics (e.g., takeoff, cruising, and landing). In particular, the diversity of satellite platforms results in differences in video spatial–temporal resolution, spectral range, and imaging views. This can help test the robustness of trackers and facilitate designing a reasonable tracking scheme.

5.3.2. *Specific category*

As illustrated in Fig. 8, the OOTB dataset consists of 110 sequences including 45 cars, 30 ships, 25 planes, and 10 trains, totaling 29,890 frames. The average, shortest, and longest videos contain 271.7, 90, and 750 frames, respectively. Car and train are the most challenging categories, while ship and plane are relatively easier to track. This is because cars are typically smaller in size and have more complex backgrounds, while trains have larger aspect ratios and more frequent non-rigid deformations. Consequently, more cars are included in the OOTB, while trains are relatively less common due to their infrequent appearance in typical scenes.

5.3.3. *Size and aspect ratio*

Object size provides two essential pieces of information: 1) it helps determine the search region of the tracker so that computational resources can be allocated appropriately, and 2) it serves as a measure of tracking difficulty, with smaller objects being more challenging to track. Fig. 8 illustrates the object area distribution of the OOTB. The average object areas for cars, ships, planes, and trains are 109.7, 238.7, 2075.3, and 1949.0 pixels, respectively. More than 75 % of the sequences have an object area smaller than 1113 pixels. As complementary information, the aspect ratio can finely describe the shape of an object. It reflects the relationship between the aspect ratio and area. Fig. 9 shows the Gaussian kernel density estimate for the object aspect ratio. The average aspect ratios for cars, ships, planes, and trains are 2.0, 2.0, 1.2, and 10.9, respectively, and their maximum aspect ratios are 4.7, 3.4, 1.6, and 14.8, respectively. The larger the aspect ratio, the more likely to encounter background interference and rotation issues, especially for trackers that only output HBB.

Table 4
Attributes and their definitions.

Attribute	Description
DEF	Deformation – non-rigid deformation of an object.
IPR	In-Plane Rotation – the object rotates in the image plane.
PO	Partial Occlusion – the object appears partially occluded in satellite video.
FO	Full Occlusion – the object appears fully occluded in satellite video.
IV	Illumination Variation – the illumination around the object is significantly changed.
MB	Motion Blur – the object region is blurred due to the motion of the object or satellite platform.
BC	Background Clutters – the background near the object has a similar texture or color as the object.
OON	Out-of-Normal – the aspect ratio of the bounding box is outside the range $[0.3, 3]$ in a video.
SA	Similar Appearance - there are objects with similar appearance near the tracked object.
LT	Less Textures – the texture information of the target is less leading to extreme difficulty to discriminate
IM	Isotropic Motion – there are objects with similar moving in magnitude and direction near the tracked object.
AM	Anisotropic Motion – there are objects with similar magnitude of motion but in opposite directions near the tracked object.

5.4. *Attribute*

GV is usually captured by a variety of optical or infrared devices such as handheld cameras, mobile surveillance devices, UAVs, and infrared cameras, which causes multiple challenges (e.g., fast motion, out-of-view, aspect ratio change, and thermal crossover). However, SV is significantly different from GV in terms of imaging devices, observation means, imaging regions, atmospheric environment, etc. The main challenges are summarized below.

Table 5
Attribute distribution of the OOTB dataset.

OOTB	DEF	IPR	PO	FO	IV	MB	BC	OON	SA	LT	IM	AM
DEF	**16**	9	4	0	6	4	7	10	9	7	2	2
IPR	9	**42**	6	2	16	12	28	11	20	13	5	8
PO	4	6	**17**	4	6	3	9	6	11	9	6	8
FO	0	2	4	**8**	3	0	2	4	4	3	2	0
IV	6	16	6	3	**62**	30	43	6	23	27	7	5
MB	4	12	3	0	30	**45**	37	4	19	24	3	7
BC	7	28	9	2	43	37	**81**	13	35	40	8	14
OON	10	11	6	4	6	4	13	**20**	14	7	5	4
SA	9	20	11	4	23	19	35	14	**46**	26	8	14
LT	7	13	9	3	27	24	40	7	26	**46**	6	11
IM	2	5	6	2	7	3	8	5	8	6	**12**	2
AM	2	8	8	0	5	7	14	4	14	11	2	**17**

The diagonal data corresponds to the distribution in the overall dataset, and each row or column represents the distribution of the attribute subset.

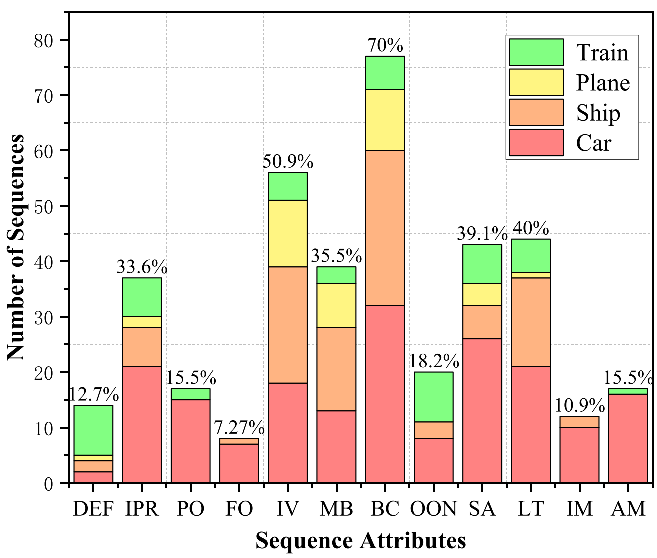

Fig. 10. Attribute distribution for each type of object in the OOTB.

- The object is small, and the image size is large (Shao et al., 2021). Meanwhile, the image resolution is lower and the object is more blurred than that of GV.
- The tracked object is usually surrounded by multiple similar disturbances, which can interfere with the tracking process (Zhang et al., 2022).
- The object blends into the background with serious mutual interference and weak distinguishability (Song et al., 2022).
- Similar objects tend to share homogeneous motion characteristics (e.g., magnitudes and directions), which can increase the difficulty of the motion feature recognition.

Considering the above problems, it is necessary to construct a set of attributes suitable for SV. As listed in Table 4, we define 12 attributes that better reflect the characteristics of SV. Each sequence is annotated with these fine-grained attributes, and the attribute distribution is reported in Table 5. By analyzing sequences with the same and different attributes, it is possible to better understand the properties of the tracker. Furthermore, Fig. 10 presents the attribute distribution for each category of objects. In summary, fine-grained attributes allow for a more detailed and comprehensive evaluation of trackers.

5.5. Evaluation protocol

A high-precision evaluation protocol is proposed for fair comparisons. Firstly, we perform the OPE evaluation and obtain the precision plot of all trackers. In addition, the normalized precision plot is used for evaluation to avoid the effect of object size. Finally, the success plot is included, and an FPS metric is used to measure the tracking speed.

5.5.1. Precision plot

The precision plot records the percentage of frames where the center location error (CLE) is smaller than the predefined threshold. As shown in Fig. 11, the CLE is determined by computing the average Euclidean distance between the center of the predicted bounding box (i.e., HBB or OBB) $(\boldsymbol{x}, \boldsymbol{y})$ and the ground truth (i.e., OBB) $(\boldsymbol{X}, \boldsymbol{Y})$. Let $\boldsymbol{d_c}$ denote the CLE, defined as

$$\boldsymbol{d_c} = \sqrt{(\boldsymbol{x} - \boldsymbol{X})^2 + (\boldsymbol{y} - \boldsymbol{Y})^2} \tag{1}$$

To account for the low spatial resolution of SVs and the small size of objects, we use the threshold varied from 1 to 30 for the precision plot and measure the tracking performance by using the precision rate (PR) at a threshold of 5 pixels. This contrasts with other benchmarks such as UAV123 (Mueller et al., 2016), OTB (Wu et al., 2013, 2015), TrackingNet (Muller et al., 2018), and LaSOT, which use a threshold varied from 1 to 50 and a threshold of 20 pixels. The reason for this difference is that SV objects are typically smaller than GV objects, and a tracking drift of 5 pixels in SV (i.e., Fig. 12(a)) is almost equivalent to that of 20 pixels in GV (i.e., Fig. 12(b)). Notably, the precision plot is sensitive to the image resolution and object size (Muller et al., 2018). As shown in Fig. 12(a) and Fig. 12(c), their resolution is the same. However, the drift magnitudes of the former are more significant than that of the latter.

5.5.2. Normalized precision plot

To compensate for the precision plot, we further propose the normalized precision plot (Muller et al., 2018) for evaluating SV trackers. The normalized precision plot shows the percentage of frames for which the normalized CLE is smaller than the predefined threshold varied from 0 to 1. Let $\boldsymbol{d_n}$ denote the normalized CLE, defined as

$$\boldsymbol{d_n} = \sqrt{((\boldsymbol{x} - \boldsymbol{X})/\boldsymbol{W})^2 + ((\boldsymbol{y} - \boldsymbol{Y})/\boldsymbol{H})^2} \tag{2}$$

where $\boldsymbol{W}$ and $\boldsymbol{H}$ are the width and height of the ground truth.

Considering that most trackers can only produce HBB, computing the normalized CLE directly using the ground truth (i.e., OBB) and the predicted result (i.e., HBB) may lead to inconsistent representations. This is because there is no explicit correspondence between their widths and heights. To this end, we propose a strategy for adaptively solving the $\boldsymbol{W}$ and $\boldsymbol{H}$. Specifically, the ground truth is first converted to an external HBB format. Subsequently, its width and length are applied for evaluation. The strategy is also embedded in the initialization process for trackers that can only receive the HBB format. For trackers that can predict the OBB format, we naturally employ the height and width of the OBB to compute the normalized CLE. In OOTB, we use the area under the curve (AUC) of the normalized precision plot, i.e., normalized pre-

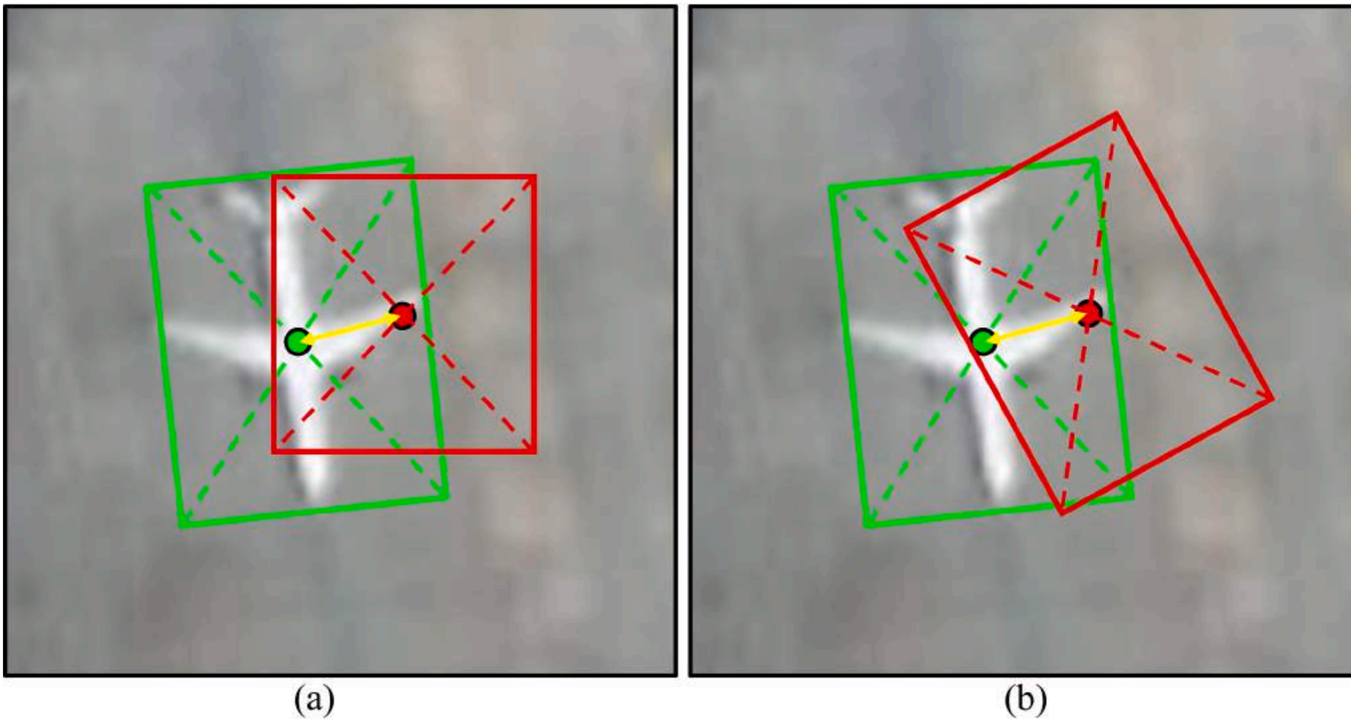

Fig. 11. Visualization of the CLE. Ground truth and predicted result are marked with green and red boxes, respectively. (a) and (b) show the CLE with HBB and OBB predictions, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

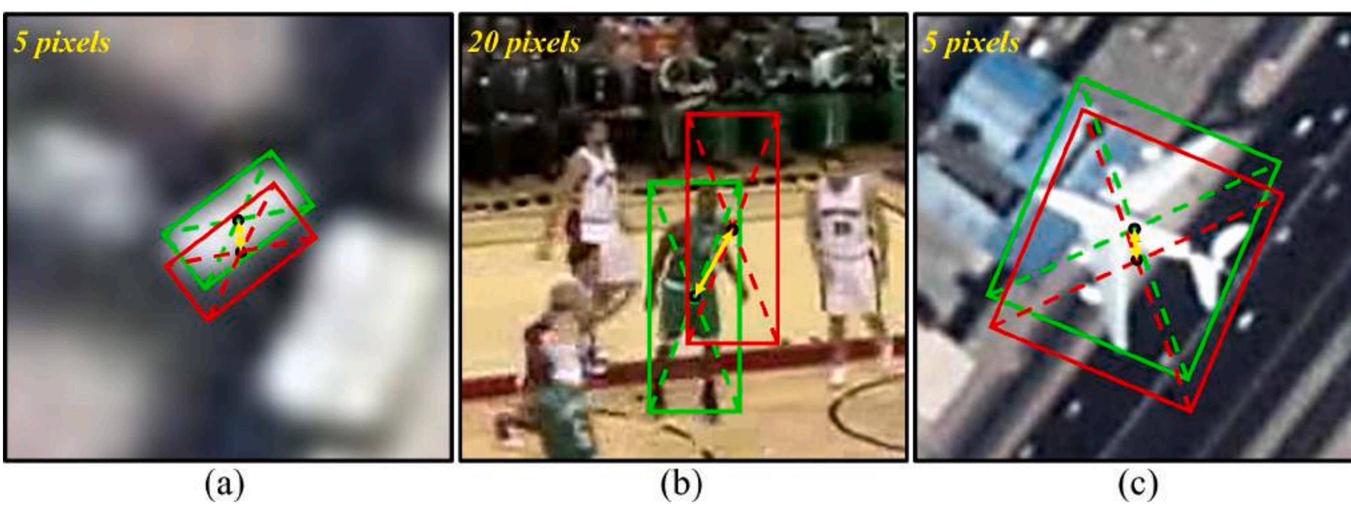

Fig. 12. Comparison of tracking drift for different videos and objects. Drift pixels are displayed in the upper left corner. Ground truth and predicted result are marked with green and red boxes, respectively. (a) and (c) show vehicle and aircraft objects from SVs, respectively. (b) shows an object from the GV. (a) and (b) suffer from approximate drift magnitudes, even though (a) drifts by 5 pixels while (b) drifts by 20 pixels. Both (a) and (c) drift by 5 pixels, but their drift magnitudes vary greatly due to the size of the object. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cision rate (NPR), to rank trackers and avoid unfair comparisons due to specific thresholds.

5.5.3. Success plot

In the success plot, the success rate (SR) aims to calculate the percentage of successful frames where the overlap surpasses the threshold varied from 0 to 1. Given the predicted bounding box r_p and ground truth r_g, the overlap score s is obtained by

$$s = |r_p \cap r_g| / |r_p \cup r_g| \tag{2}$$

where $\cap$ and $\cup$ represent intersection and union, respectively, and $|\bullet|$ denotes the number of pixels in the given region.

As discussed above, OOTB is annotated with OBB format, which is different from previous SV datasets annotated with HBB format, such as VISO (Yin et al., 2022), SatSOT (Zhao et al., 2022), XDU-BDSTU (Zhang et al., 2022), ThickSiam_D (Zhang et al., 2023), and AIR-MOT (He et al., 2022). For fair assessment, two methods for resolving the overlap score are proposed. The first method is designed to assess the tracker with HBB output. Concretely, the ground truth (i.e., OBB) and predicted bounding box (i.e., HBB) are converted into corresponding external HBBs. The intersection and union regions of these two HBBs are then obtained to calculate the overlap, as shown in Fig. 13(a). While the second method is designed to assess the tracker with OBB output. In particular, we directly

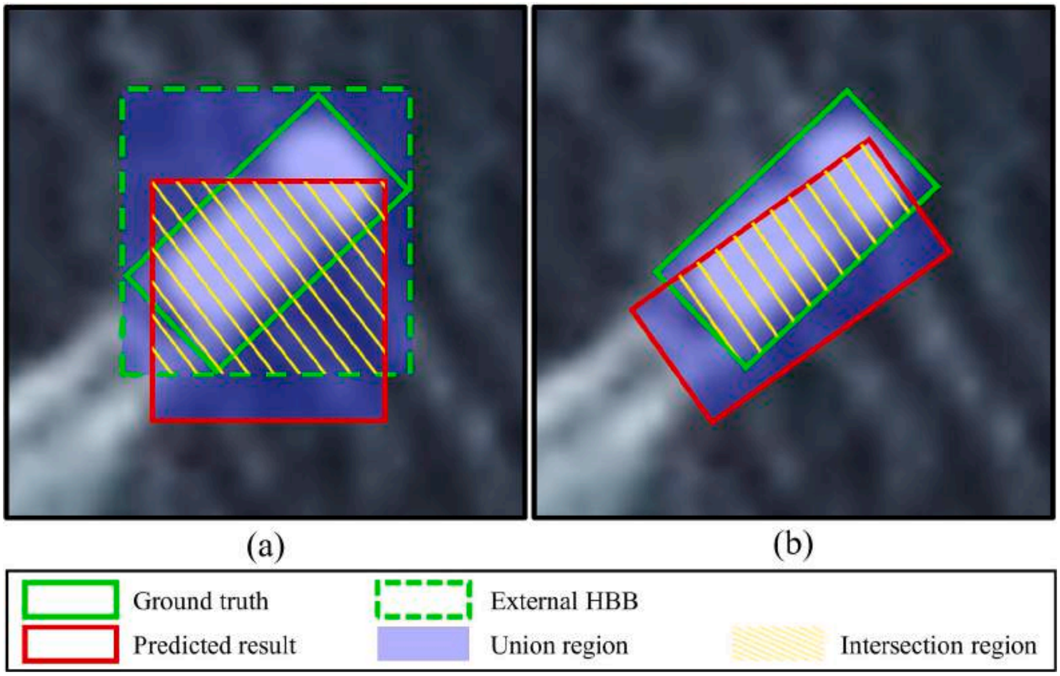

Fig. 13. Visualization of the intersection and union regions. (a) and (b) show the predicted results for HBB and OBB, respectively.

compute the intersection and union regions between the ground truth (i.e., OBB) and predicted bounding box (i.e., OBB), as shown in Fig. 13(b). The AUC of the success plot (i.e., SR) is used to rank trackers.

Overall, the larger the PR, NPR, SR, and FPS values, the better the tracking performance.

5.6. Selecting 33 SOTA trackers for evaluation

In OOTB, we compare and analyze 33 representative SOTA trackers with a total of 58 models covering different features, backbones, and tracker tags. The compared algorithms are CSK (Henriques et al., 2012), SAMF (Li and Zhu, 2015), DAT (Possegger et al., 2015), KCF (Henriques et al., 2015), SRDCF, Staple, DSST, BACF (Galoogahi et al., 2017b), SiamRPN (Li et al., 2018a), DaSiamRPN (Zhu et al., 2018), ARCF (Huang et al., 2019), SiamRPN++ (Li et al., 2019a), UpdateNet (Zhang et al., 2019), SiamDW (Zhang, 2019b), SiamMask (Wang et al., 2019c), SiamBAN (Chen et al., 2020), SiamFC++ (Xu et al., 2020), AutoTrack (Li et al., 2020), CFME (Xuan et al., 2020), SiamGAT (Guo et al., 2021), LightTrack (Yan et al., 2021), Stark (Yan et al., 2021a), SiamCAR (Cui et al., 2022a), OSTrack (Ye et al., 2022a), SimTrack (Chen et al., 2022a), DF (https://github.com/YZCU/DF) (Chen. et al., 2022), RAMC (Chen et al., 2022c), SBT (Xie et al., 2022), GRM (Gao et al., 2023), SeqTrack (Chen et al., 2023a), ARTrack (Wei et al., 2023), ODTrack (Zheng et al., 2024), and SMAT (Yelluru Gopal and Amer, 2024). These SOTAs cover all mainstream tracking paradigms. Usually, a tracker is trained using distinctive models and datasets to meet diverse needs. To explore suitable methods and models for SOT in SV, we benchmark 33 SOTAs with a total of 58 models.

6. Experiments and analysis

In this section, we perform comprehensive experiments and analysis. Section 6.1 quantitatively compares all trackers including overall, category-based, and attribute-based evaluations. Section 6.2 presents the qualitative results. While Section 6.3 compares and analyzes the running speeds.

6.1. Quantitative evaluations

6.1.1. Overall evaluation results

Here we provide a comprehensive assessment on OOTB. To be fair, we conduct experiments using the officially provided models and corresponding configuration parameters.

Table 6 shows the characteristics, overall results, category-based results, and running speed for all 33 trackers with a total of 58 models. The benchmark includes four metrics (i.e., PR, NPR, SR, and FPS). Fig. 14 displays the precision plot, normalized precision plot, and success plot for the top 30 trackers. The values in the legend denote PR, NPR, and SR, respectively. SiamCAR with default parameters attains outstanding performance in both precision plot and normalized precision plot with a PR of 0.824 and NPR of 0.779. Additionally, SiamFC++ with AlexNet adopts the anchor-free idea and proposes a set of object state estimation guidelines, which secures top ranking in the SV tracking.

In the success plot, SiamDW with CIRNext22 performs remarkably well with an SR of 0.645, surpassing the fourth-place DF by 2.2 %. RAMC can combine the appearance and motion features to deal with object rotation and tracking drifts. It achieves a satisfactory SR of 0.598. Since RAMC can estimate compact OBB, it is prone to tracking position bias. Therefore, the results obtained in the precision map and normalized precision map are relatively poor. Furthermore, DSST produces satisfactory results in PR, NPR, and SR scores. Notably, SiamDW including SiamDW_CIRNext22 SiamDW_CIRIncep22, and SiamDW_CIResNet22, achieves competitive PR, NPR, and SR scores. This indicates that the residual modules and deep-wide network structure of SiamDW aid in mining the semantic information of SV objects.

6.1.2. Category-based evaluation results

In this section, we perform a category-based evaluation on OOTB. As shown in Fig. 8, SV objects can typically be divided into four categories. Table 6 summarizes the characteristics and results of the category-based evaluation. Fig. 15 shows the precision plot, normalized precision plot, and success plot for the top 30 trackers. It is observed that the plane is the easiest object to be tracked. The top 30 trackers exhibit tighter tracking curves and higher PR, NPR, and SR scores. This is attributed to the large size and the significant texture and structural information,

Table 6
Characteristics and results for 33 SOTAs with a total of 58 models. All SOTAs and their models are evaluated to explore the optimal framework and model. The results include accuracy metrics (i.e., PR, NPR, and SR) and speed metric (i.e., FPS).

Tracker	Venue	Feature/Backbone	Tracker tag	Specific tracker name	FPS		Overall Result			Category-based result											
										Car			Ship			Plane			Train		
					CPU	GPU	PR	NPR	SR	PR	NPR	SR	PR	NPR	SR	PR	NPR	SR	PR	NPR	SR
CSK	ECCV 2012	I	Default	CSK	465.8	—	0.540	0.578	0.478	0.483	0.471	0.359	0.564	0.554	0.459	0.737	0.827	0.740	0.233	0.507	0.412
SAMF	ECCV 2015	HOG + CN + I	Default	SAMF	60.1	—	0.590	0.589	0.459	0.614	0.555	0.407	0.514	0.492	0.390	0.792	0.846	0.698	0.203	0.392	0.308
DAT	CVPR 2015	CH	Default	DAT	323.7	—	0.474	0.569	0.453	0.519	0.472	0.376	0.610	0.660	0.505	0.384	0.711	0.594	0.089	0.380	0.292
KCF	TPAMI 2015	HOG	Default	KCF	388.5	—	0.493	0.529	0.428	0.486	0.454	0.343	0.492	0.494	0.389	0.619	0.761	0.676	0.214	0.390	0.308
SRDCF	ICCV 2015	HOG	Default	SRDCF	5.0	—	0.766	0.744	0.584	0.779	0.705	0.547	0.890	0.818	0.663	0.856	0.892	0.693	0.112	0.324	0.241
Staple	CVPR 2016	HOG + CN	Default	Staple	145.2	—	0.772	0.738	0.607	0.899	0.793	0.649	0.799	0.723	0.607	0.809	0.853	0.705	0.024	0.247	0.177
DSST	TPAMI 2017	HOG + I	Default	DSST	313.5	—	0.801	0.770	0.621	0.895	0.787	0.609	0.787	0.717	0.581	0.867	0.903	0.780	0.261	0.518	0.396
BACF	ICCV 2017	HOG	Default	BACF	84.1	—	0.795	0.750	0.547	0.847	0.750	0.499	0.897	0.800	0.612	0.849	0.861	0.673	0.117	0.327	0.249
SiamRPN	CVPR 2018	AlexNet	Default	SiamRPN	—	102.5	0.605	0.624	0.464	0.616	0.544	0.442	0.643	0.597	0.460	0.741	0.862	0.560	0.105	0.464	0.337
DaSiamRPN	ECCV 2018	AlexNet	Default	DaSiamRPN_Def	—	94.6	0.704	0.738	0.577	0.851	0.754	0.605	0.613	0.658	0.537	0.803	0.881	0.647	0.072	0.543	0.393
		AlexNet	OTB	DaSiamRPN_OTB	—	168.9	0.733	0.749	0.588	0.872	0.777	0.618	0.725	0.710	0.569	0.744	0.841	0.638	0.110	0.517	0.383
		AlexNet	VOT	DaSiamRPN_VOT	—	132.2	0.578	0.646	0.515	0.809	0.722	0.580	0.358	0.461	0.384	0.617	0.834	0.641	0.097	0.385	0.295
ARCF	ICCV 2019	HOG + CN + I	Default	ARCF	32.5	—	0.751	0.735	0.603	0.803	0.713	0.558	0.828	0.744	0.609	0.807	0.897	0.795	0.147	0.399	0.309
SiamRPN++	CVPR 2019	ResNet-50	Default	SiamRPN++_Def	—	70.9	0.758	0.735	0.589	0.843	0.746	0.600	0.799	0.737	0.597	0.829	0.863	0.694	0.079	0.354	0.252
		ResNet-50	L-T	SiamRPN++_LT	—	42.5	0.683	0.674	0.537	0.737	0.652	0.524	0.705	0.702	0.552	0.815	0.844	0.688	0.046	0.260	0.177
UpdateNet	ICCV 2019	AlexNet	DaSiamRPN	UpdateNet	—	46.0	0.652	0.708	0.550	0.853	0.752	0.600	0.453	0.532	0.452	0.747	0.875	0.623	0.101	0.624	0.440
SiamDW	CVPR 2019	CIResNet22	SiamFC	SiamDW_CIResNet22	—	88.9	0.777	0.749	0.630	0.780	0.702	0.577	0.857	0.789	0.670	0.954	0.937	0.820	0.077	0.369	0.273
		CIRIncep22	SiamFC	SiamDW_CIRIncep22	—	161.0	0.800	0.764	0.644	0.823	0.738	0.613	0.895	0.812	0.696	0.927	0.933	0.799	0.094	0.321	0.242
		CIRNext22	SiamFC	SiamDW_CIRNext22	—	125.6	0.794	0.765	0.645	0.836	0.747	0.617	0.857	0.787	0.672	0.944	0.937	0.822	0.038	0.349	0.255
SiamMask	CVPR 2019	ResNet-50	HBB	SiamMask_HBB	—	108.7	0.747	0.729	0.552	0.829	0.739	0.581	0.824	0.773	0.610	0.793	0.830	0.575	0.033	0.297	0.188
		ResNet-50	OBB	SiamMask_OBB	—	90.4	0.692	0.663	0.426	0.824	0.683	0.417	0.787	0.702	0.443	0.606	0.779	0.555	0.027	0.162	0.092
SiamBAN	CVPR 2020	ResNet-50	OTB	SiamBAN_OTB	—	56.3	0.754	0.733	0.523	0.814	0.700	0.532	0.765	0.752	0.581	0.906	0.905	0.562	0.076	0.394	0.206
		ResNet-50	VOT	SiamBAN_VOT	—	42.9	0.760	0.734	0.517	0.811	0.692	0.522	0.780	0.746	0.589	0.914	0.906	0.551	0.087	0.456	0.194
SiamFC++	AAAI 2020	AlexNet	OTB	SiamFC++_Alex	—	257.2	0.797	0.778	0.616	0.813	0.736	0.595	0.919	0.833	0.668	0.921	0.924	0.711	0.053	0.441	0.321
		GoogLeNet	OTB	SiamFC++_Google	—	122.6	0.668	0.672	0.530	0.716	0.650	0.517	0.600	0.625	0.505	0.915	0.921	0.720	0.039	0.286	0.186
AutoTrack	CVPR 2020	HOG + CN + I	Default	AutoTrack	114.7	—	0.760	0.728	0.558	0.831	0.717	0.517	0.831	0.743	0.556	0.801	0.872	0.752	0.124	0.373	0.259
CFME	TGRS 2020	HOG	Default	CFME	7.2	—	0.753	0.714	0.610	0.813	0.725	0.593	0.887	0.798	0.686	0.777	0.808	0.743	0.026	0.174	0.127
SiamGAT	CVPR 2021	GoogLeNet	GOT-10k	SiamGAT_GOT10k	—	56.8	0.692	0.686	0.474	0.774	0.693	0.495	0.662	0.691	0.531	0.851	0.883	0.521	0.017	0.155	0.093
		GoogLeNet	LaSOT	SiamGAT_LaSOT	—	44.4	0.463	0.463	0.338	0.478	0.398	0.315	0.350	0.356	0.301	0.753	0.842	0.519	0.009	0.133	0.104
		GoogLeNet	OTB-UAV	SiamGAT_OTB-UAV	—	113.1	0.768	0.744	0.550	0.810	0.730	0.551	0.837	0.780	0.623	0.896	0.910	0.604	0.051	0.276	0.194
		GoogLeNet	TrackingNet	SiamGAT_TrackingNet	—	69.8	0.789	0.757	0.577	0.818	0.727	0.575	0.871	0.790	0.637	0.923	0.926	0.632	0.074	0.369	0.266
LightTrack	CVPR 2021	Custom	Mobile	LightTrack	—	109.0	0.611	0.616	0.487	0.608	0.545	0.422	0.606	0.596	0.499	0.858	0.891	0.715	0.028	0.305	0.172
Stark	ICCV 2021	ResNet-50	S50	Stark_S50	—	97.7	0.571	0.573	0.451	0.408	0.365	0.292	0.838	0.782	0.616	0.748	0.787	0.613	0.062	0.349	0.266
		ResNet-50	ST50	Stark_ST50	—	87.1	0.610	0.601	0.477	0.487	0.432	0.337	0.841	0.765	0.613	0.782	0.804	0.647	0.048	0.356	0.274
		ResNet-101	ST101	Stark_ST101	—	58.2	0.622	0.618	0.486	0.524	0.467	0.363	0.768	0.729	0.579	0.822	0.844	0.670	0.119	0.397	0.299
SiamCAR	IJCV 2022	ResNet-50	Default	SiamCAR_Def	—	36.6	0.824	0.779	0.607	0.849	0.759	0.578	0.946	0.848	0.679	0.943	0.932	0.742	0.051	0.274	0.184
		ResNet-50	LaSOT	SiamCAR_LaSOT	—	47.9	0.757	0.716	0.529	0.781	0.654	0.508	0.805	0.756	0.608	0.935	0.931	0.600	0.057	0.334	0.208
		ResNet-50	GOT-10k	SiamCAR_GOT10k	—	57.8	0.731	0.710	0.538	0.809	0.722	0.570	0.677	0.651	0.510	0.933	0.930	0.658	0.035	0.287	0.173
OSTrack	ECCV 2022	ViT-Base	256-Default	OSTrack_256Def	—	73.5	0.665	0.650	0.515	0.587	0.518	0.402	0.635	0.598	0.494	0.910	0.921	0.706	0.492	0.726	0.602
		ViT-Base	256-GOT-10 k	OSTrack_256GOT10k	—	90.4	0.585	0.588	0.439	0.391	0.348	0.258	0.802	0.723	0.532	0.862	0.908	0.696	0.121	0.461	0.329
		ViT-Base	384-Default	OSTrack_384Def	—	36.6	0.693	0.677	0.530	0.542	0.482	0.376	0.839	0.781	0.618	0.915	0.919	0.701	0.374	0.636	0.527
		ViT-Base	384-GOT-10 k	OSTrack_384GOT10k	—	39.3	0.543	0.561	0.417	0.316	0.284	0.206	0.742	0.718	0.519	0.869	0.894	0.692	0.147	0.506	0.371
SimTrack	ECCV 2022	ViT-Base	ViT-B/16	SimTrack	—	59.5	0.535	0.544	0.430	0.365	0.330	0.265	0.633	0.593	0.484	0.823	0.857	0.645	0.286	0.580	0.476
DF	JSTARS 2022	HOG + CN + GCS	Default	DF	79.6	—	0.758	0.747	0.623	0.898	0.794	0.642	0.834	0.750	0.628	0.711	0.860	0.763	0.024	0.247	0.178
RAMC	RS 2022	HOG + OF	Default	RAMC	25.1	—	0.781	0.703	0.598	0.747	0.624	0.507	0.773	0.663	0.564	0.881	0.865	0.788	0.702	0.768	0.636
SBT	CVPR 2022	SBT-Base	Default	SBT_Def	—	30.4	0.700	0.682	0.525	0.552	0.489	0.373	0.953	0.843	0.662	0.850	0.895	0.677	0.228	0.533	0.413
		SBT-Base	GOT-10k	SBT_GOT10k	—	20.8	0.552	0.581	0.419	0.444	0.404	0.292	0.758	0.713	0.515	0.707	0.876	0.637	0.031	0.248	0.162
GRM	CVPR 2023	ViT-Base	256-Default	GRM_Def	—	21.1	0.645	0.629	0.500	0.503	0.448	0.355	0.673	0.640	0.524	0.923	0.921	0.704	0.511	0.682	0.571
		ViT-Base	256-GOT-10 k	GRM_GOT10k	—	22.8	0.568	0.580	0.427	0.398	0.354	0.263	0.723	0.709	0.512	0.878	0.914	0.686	0.095	0.368	0.266
		ViT-Large	320-Large	GRM_L	—	12.2	0.662	0.650	0.502	0.476	0.427	0.329	0.821	0.756	0.589	0.898	0.912	0.686	0.430	0.684	0.556
SeqTrack	CVPR2023	ViT-Base	256-Default	SeqTrack_Def	—	14.8	0.726	0.720	0.550	0.633	0.558	0.422	0.848	0.783	0.610	0.902	0.919	0.691	0.344	0.762	0.595
		ViT-Base	256-GOT-10 k	SeqTrack_GOT10k	—	17.1	0.600	0.627	0.464	0.498	0.446	0.326	0.680	0.665	0.503	0.868	0.908	0.679	0.154	0.624	0.433
		ViT-Large	256-Large	SeqTrack_L	—	10.7	0.723	0.706	0.528	0.643	0.562	0.422	0.784	0.729	0.558	0.892	0.921	0.646	0.475	0.742	0.613
		ViT-Large	256-GOT-10 k	SeqTrack_LGOT10k	—	10.5	0.573	0.605	0.432	0.426	0.375	0.269	0.759	0.719	0.514	0.805	0.900	0.638	0.100	0.562	0.401
ARTrack	CVPR2023	ViT-Base	256-Default	ARTrack	—	13.3	0.762	0.731	0.549	0.674	0.600	0.452	0.903	0.820	0.639	0.857	0.918	0.641	0.493	0.593	0.487
ODTrack	AAAI2024	ViT-Base	384-Default	ODTrack_Def	—	14.6	0.686	0.695	0.526	0.686	0.601	0.445	0.715	0.707	0.558	0.808	0.878	0.647	0.290	0.630	0.491
		ViT-Large	384-Large	ODTrack_L	—	9.0	0.708	0.715	0.534	0.665	0.587	0.445	0.719	0.716	0.561	0.918	0.929	0.646	0.343	0.751	0.576
SMAT	WACV 2024	MobileViTv2	Default	SMAT	—	77.1	0.688	0.665	0.526	0.599	0.531	0.417	0.874	0.793	0.628	0.860	0.883	0.711	0.101	0.335	0.244

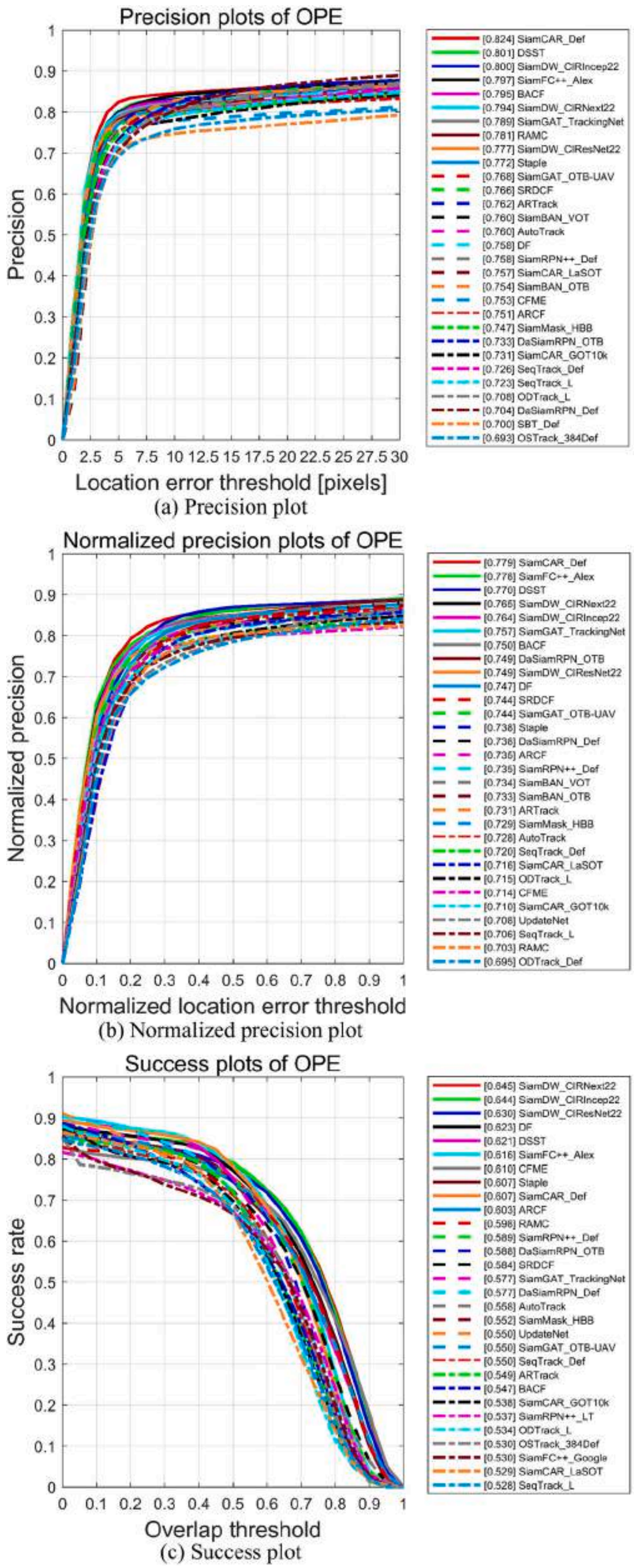

Fig. 14. Overall results for the top 30 trackers on OOTB. (a) Precision plot. (b) Normalized precision plot. (c) Success plot. The values in the legend denote PR, NPR, and SR, respectively.

which facilitates the extraction of discriminative features. In this case, RAMC, DF, and CFME, three tailored trackers for SV tracking, perform well, ranking 5th, 7th, and 9th in the success plot, respectively. This is because of their ability to combine appearance and motion features to mitigate tracking drifts. Ship objects have relatively large sizes and stable motion patterns. However, the interference of tail waves and complex backgrounds may cause tracking drifts. Car objects face various challenges due to their small size, complex motion patterns, and similar contexts. SOTA trackers, such as DF, Staple, UpdateNet, SiamFC++, and SiamCAR, can adapt to object changes and produce acceptable results. However, compared to plane and ship objects, the small size of cars leads to lower PR, NPR, and SR scores for almost all trackers. For the train object, it is one of the most challenging categories, and almost all SOTAs exhibit inferior performance. This is mainly due to the larger aspect ratios and more frequent non-rigid deformations compared to other categories. In this case, almost all the top trackers such as SeqTrack, OSTrack, ODTrack, GRM, and OSTrack use CNN features, as shown in Fig. 15(l). In contrast, hand-crafted feature-based trackers such as CSK, DSST, and SAMF have difficulty adapting to complex scenarios due to limited representations. While most trackers achieve the worst results in the train category, indicating that more attention should be paid to addressing this issue.

6.1.3. *Attribute-based evaluation results*

To evaluate the strengths and limitations of trackers, we perform attribute-based evaluation. Table 7 shows the SR scores for each attribute, and Fig. 16 presents the success plot for the top 30 trackers. DF and CFME achieve significant results in terms of PO, FO, IV, MB, BC, LT, IM, and AM, due to their ability to sense the tracking confidence of current frames and predict the object position in subsequent frames. The most prevalent challenge is IPR, where RAMC achieves the optimal SR score of 0.601, which is 1.7 % higher than the second-best SiamDW (SiamDW_CIRIncep22). SiamRPN++ (SiamRPN++_Def) obtains an SR score of 0.577 and ranks 3-rd. Besides RAMC, the best three correlation filter-based trackers are DSST, DF, and Staple, with SR scores of 0. 531, 0.524, and 0.522, respectively. Regarding the FO attribute, CFME ranks first with an SR score of 0.490, followed by SiamDW (SiamDW_CIRNext22), SiamDW (SiamDW_CIResNet22), and DF. These SOTAs encounter a severe drop in SR score, which indicates that FO is extremely challenging in SV. Fig. 16 also demonstrates that deep learning-based trackers are more robust in handling challenging attributes in the SV tracking domain. Whereas the correlation filter-based trackers, except for DF, DSST, and CFME, are relatively weak in addressing these challenges. Table 8 presents the NPR scores of all trackers on 12 attributes, and Fig. 17 presents the normalized precision plot for the top 30 trackers. It is found that the NPR scores also show a severe decrease in FO, confirming that FO is extremely challenging in SV tracking.

6.2. *Qualitative evaluations*

For qualitative evaluation, we visualize nine SOTA trackers including ODTrack (ODTrack_L), SeqTrack (SeqTrack_Def), ARTrack, CFME, SiamFC++ (SiamFC++_Alex), DSST, SiamDW (SiamDW_CIRNext22), RAMC, and DF, covering a wide range of tracking paradigms. To showcase their performance, we have selected six sequences with diverse attributes and object categories, namely *Car_16*, *Ship_10*, *Plane_2*, *Plane_21*, *Train_1*, and *Train_5*. Fig. 18 shows the qualitative results to help intuitively understand the tracking performance.

In the *Car_16* sequence, the object undergoes in-plane rotation, full occlusion, and illumination variation, which challenge the tracking algorithms. The results show that SOTA trackers encounter tracking failure when the car is completely occluded. Compared to other trackers, DF and DSST perform better and cross the obstacle. CFME is also capable of overcoming occlusion but may fail in the case of long-duration occlusion. The *Train_1* sequence, as one of the most challenging sequences,

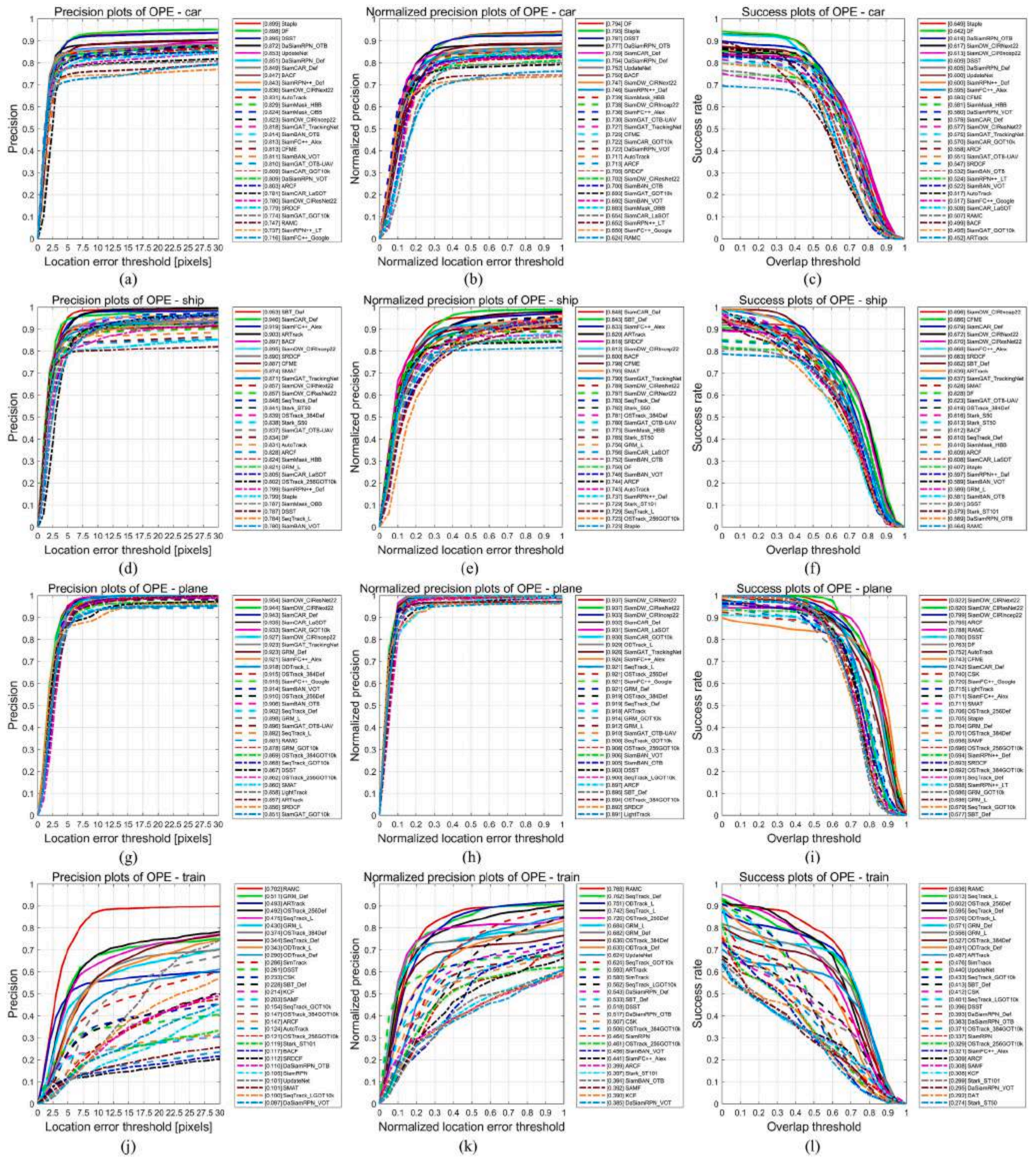

Fig. 15. The precision plot (column 1), normalized precision plot (column 2), and success plot (column 3) for the top 30 trackers. Rows 1 to 4 show the results for car, ship, plane, and train, respectively.

encounters deformation, in-plane rotation, illumination variation, motion blur, and out-of-normal. In particular, the train object tends to undergo non-rigid deformation, making it extremely difficult to track accurately. In this case, DSST, DF, and CFME initially drift away from the object (e.g., frame #0164). While RAMC, ARTrack, and SiamFC++ (SiamFC++_Alex) can adapt to object changes and achieve better tracking performance. More visual samples can be found in Fig. 18.

Based on qualitative results, we can draw several conclusions. Firstly, combining appearance with motion features is useful in handling multiple challenges. Secondly, deep feature-based trackers are more robust in the SV tracking domain than hand-crafted feature-based trackers. Thirdly, tracking the train object is extremely demanding and requires more attention. Finally, the SOT of SV is far from being well resolved. There is still room for improvement in this field.

Table 7
SR scores for each attribute.

Tracker	Venue	Specific tracker name	DEF	IPR	PO	FO	IV	MB	BC	OON	SA	LT	IM	AM
CSK	ECCV 2012	CSK	0.522	0.443	0.355	0.310	0.478	0.477	0.451	0.467	0.365	0.319	0.379	0.288
SAMF	ECCV 2015	SAMF	0.417	0.412	0.457	0.375	0.451	0.497	0.454	0.430	0.410	0.313	0.439	0.325
DAT	CVPR 2015	DAT	0.446	0.385	0.419	0.310	0.475	0.465	0.428	0.408	0.340	0.326	0.366	0.358
KCF	TPAMI 2015	KCF	0.375	0.363	0.398	0.327	0.423	0.462	0.419	0.392	0.348	0.288	0.357	0.294
SRDCF	ICCV 2015	SRDCF	0.447	0.494	0.547	0.379	0.590	0.601	0.594	0.454	0.515	0.535	0.560	0.522
Staple	CVPR 2016	Staple	0.429	0.522	0.604	0.400	0.607	0.625	0.621	0.446	0.543	0.536	0.652	0.657
DSST	TPAMI 2017	DSST	0.517	0.531	0.619	0.404	0.611	0.642	0.629	0.525	0.577	0.542	0.593	0.585
BACF	ICCV 2017	BACF	0.445	0.450	0.538	0.438	0.540	0.568	0.561	0.422	0.454	0.479	0.541	0.469
SiamRPN	CVPR 2018	SiamRPN	0.524	0.460	0.417	0.282	0.446	0.443	0.438	0.410	0.407	0.351	0.459	0.363
DaSiamRPN	ECCV 2018	DaSiamRPN_Def	0.541	0.574	0.585	0.412	0.555	0.540	0.565	0.523	0.514	0.496	0.554	0.607
		DaSiamRPN_OTB	0.535	0.567	0.603	0.406	0.580	0.592	0.578	0.512	0.533	0.512	0.585	0.644
		DaSiamRPN_VOT	0.491	0.515	0.562	0.317	0.492	0.487	0.487	0.427	0.483	0.430	0.514	0.600
ARCF	ICCV 2019	ARCF	0.513	0.494	0.520	0.395	0.611	0.625	0.607	0.479	0.510	0.519	0.601	0.528
SiamRPN++	CVPR 2019	SiamRPN++_Def	0.487	0.577	0.533	0.354	0.581	0.578	0.576	0.451	0.536	0.504	0.620	0.607
		SiamRPN++_LT	0.447	0.518	0.512	0.387	0.539	0.525	0.510	0.435	0.448	0.440	0.568	0.495
UpdateNet	ICCV 2019	UpdateNet	0.545	0.562	0.585	0.412	0.517	0.502	0.531	0.543	0.510	0.455	0.491	0.607
SiamDW	CVPR 2019	SiamDW_CIResNet22	0.500	0.553	0.568	0.457	0.655	0.655	0.623	0.472	0.533	0.550	0.656	0.578
		SiamDW_CIRIncep22	0.494	0.584	0.532	0.444	0.655	0.660	0.653	0.477	0.565	0.544	0.642	0.588
		SiamDW_CIRNext22	0.501	0.577	0.567	0.457	0.655	0.666	0.647	0.485	0.556	0.553	0.639	0.640
SiamMask	CVPR 2019	SiamMask_HBB	0.437	0.524	0.513	0.399	0.516	0.541	0.553	0.439	0.487	0.506	0.589	0.555
		SiamMask_OBB	0.305	0.423	0.378	0.280	0.417	0.414	0.425	0.323	0.346	0.344	0.486	0.382
SiamBAN	CVPR 2020	SiamBAN_OTB	0.424	0.534	0.510	0.335	0.504	0.533	0.516	0.398	0.469	0.459	0.530	0.536
		SiamBAN_VOT	0.409	0.515	0.485	0.323	0.505	0.542	0.522	0.394	0.451	0.465	0.517	0.524
SiamFC++	AAAI 2020	SiamFC++_Alex	0.506	0.572	0.520	0.432	0.611	0.630	0.619	0.519	0.533	0.533	0.599	0.562
		SiamFC++_Google	0.404	0.530	0.495	0.402	0.503	0.538	0.525	0.435	0.473	0.377	0.564	0.502
AutoTrack	CVPR 2020	AutoTrack	0.462	0.475	0.496	0.366	0.557	0.578	0.561	0.426	0.474	0.454	0.550	0.460
CFME	TGRS 2020	CFME	0.351	0.448	0.596	0.490	0.624	0.649	0.629	0.455	0.515	0.544	0.680	0.628
SiamGAT	CVPR 2021	SiamGAT_GOT10k	0.312	0.468	0.437	0.351	0.456	0.495	0.484	0.349	0.427	0.398	0.565	0.445
		SiamGAT_LaSOT	0.292	0.279	0.271	0.213	0.351	0.354	0.323	0.241	0.291	0.238	0.299	0.226
		SiamGAT_OTB-UAV	0.421	0.530	0.512	0.432	0.545	0.560	0.550	0.419	0.476	0.480	0.545	0.508
		SiamGAT_TrackingNet	0.464	0.557	0.548	0.449	0.566	0.588	0.580	0.481	0.521	0.501	0.591	0.552
LightTrack	CVPR 2021	LightTrack	0.400	0.444	0.380	0.344	0.483	0.480	0.468	0.386	0.394	0.324	0.525	0.366
Stark	ICCV 2021	Stark_S50	0.390	0.392	0.277	0.331	0.468	0.465	0.443	0.393	0.307	0.327	0.313	0.289
		Stark_ST50	0.381	0.409	0.326	0.306	0.490	0.496	0.474	0.422	0.324	0.366	0.386	0.345
		Stark_ST101	0.461	0.429	0.340	0.366	0.496	0.507	0.476	0.458	0.352	0.365	0.443	0.360
SiamCAR	IJCV 2022	SiamCAR_Def	0.434	0.561	0.519	0.430	0.607	0.626	0.612	0.425	0.522	0.526	0.589	0.577
		SiamCAR_LaSOT	0.425	0.501	0.403	0.344	0.534	0.555	0.534	0.418	0.439	0.439	0.536	0.468
		SiamCAR_GOT10k	0.410	0.524	0.525	0.424	0.499	0.528	0.532	0.403	0.455	0.457	0.545	0.564
OSTrack	ECCV 2022	OSTrack_256Def	0.643	0.498	0.438	0.389	0.503	0.520	0.497	0.636	0.456	0.377	0.508	0.438
		OSTrack_256GOT10k	0.476	0.394	0.269	0.218	0.469	0.485	0.431	0.386	0.350	0.311	0.408	0.270
		OSTrack_384Def	0.604	0.478	0.415	0.428	0.549	0.552	0.519	0.579	0.435	0.436	0.488	0.431
		OSTrack_384GOT10k	0.475	0.396	0.222	0.219	0.445	0.447	0.401	0.389	0.300	0.270	0.260	0.199
SimTrack	ECCV 2022	SimTrack	0.545	0.393	0.310	0.306	0.458	0.452	0.404	0.510	0.336	0.290	0.339	0.278
DF	JSTARS 2022	DF	0.420	0.524	0.588	0.455	0.637	0.636	0.636	0.441	0.561	0.544	0.633	0.655
RAMC	RS 2022	RAMC	0.694	0.601	0.526	0.319	0.603	0.563	0.575	0.621	0.524	0.466	0.582	0.483
SBT	CVPR 2022	SBT_Def	0.503	0.448	0.357	0.415	0.553	0.582	0.524	0.520	0.409	0.430	0.456	0.391
		SBT_GOT10k	0.375	0.361	0.314	0.383	0.435	0.460	0.403	0.366	0.293	0.302	0.357	0.303
GRM	CVPR 2023	GRM_Def	0.634	0.471	0.426	0.318	0.512	0.527	0.475	0.609	0.414	0.371	0.454	0.383
		GRM_GOT10k	0.445	0.378	0.279	0.235	0.449	0.455	0.411	0.350	0.307	0.278	0.392	0.255
		GRM_L	0.641	0.489	0.399	0.387	0.519	0.527	0.482	0.574	0.408	0.371	0.408	0.366
SeqTrack	CVPR2023	SeqTrack_Def	0.651	0.518	0.445	0.376	0.555	0.559	0.545	0.587	0.453	0.454	0.483	0.476
		SeqTrack_GOT10k	0.514	0.444	0.370	0.394	0.450	0.491	0.449	0.503	0.373	0.327	0.395	0.326
		SeqTrack_L	0.650	0.516	0.471	0.420	0.514	0.517	0.524	0.620	0.454	0.445	0.484	0.454
		SeqTrack_LGOT10k	0.493	0.410	0.295	0.308	0.446	0.457	0.424	0.428	0.346	0.333	0.330	0.300
ARTrack	CVPR2023	ARTrack	0.600	0.538	0.432	0.371	0.560	0.569	0.540	0.545	0.453	0.445	0.483	0.460
ODTrack	AAAI2024	ODTrack_Def	0.575	0.501	0.444	0.431	0.513	0.500	0.513	0.533	0.466	0.431	0.496	0.419
		ODTrack_L	0.628	0.511	0.459	0.377	0.530	0.529	0.529	0.614	0.469	0.446	0.499	0.446
SMAT	WACV 2024	SMAT	0.475	0.482	0.467	0.371	0.532	0.561	0.505	0.428	0.400	0.411	0.469	0.413

6.3. Running speed analysis

Table 6 presents the FPS metric on Central Processing Unit (CPU) and Graphics Processing Unit (GPU) devices. Traditional SOTA trackers, such as CSK, SAMF, DAT, Staple, and AutoTrack, mainly utilize the CPU device. For trackers using deep features, they usually rely on the GPU device and are tested on the NVIDIA GeForce RTX 4060 GPU. CSK achieves the fastest speed with 465.8 FPS on CPU. Benefiting from the efficient calculation of the circulant matrix and Fourier domain, KCF achieves second place with 388.5 FPS on the CPU. DAT follows with 323.7 FPS. For those deep trackers, SiamFC++ (SiamFC++_Alex) holds the highest speed of 257.2 FPS on GPU. DaSiamRPN (DaSiamRPN_OTB) follows with 168.9 FPS. In contrast, recent SOTAs achieve relatively low tracking speeds, such as SMAT, ODTrack, ARTrack, SeqTrack, GRM, and SBT. Therefore, realizing the accuracy-speed trade-off is subject to further research in the field of satellite video object tracking.

7. Discussion and recommendations for future work

Remote sensing Earth observation techniques have achieved vigorous development in change detection (Wang et al., 2022c), anomaly detection (Cheng et al., 2024; Lin et al., 2023), clustering (Guan et al., 2023), segmentation (Wang et al., 2022b), etc. However, previous research has mainly focused on image data. The emergence of video satellites has opened up a new era of remote sensing Earth observation from static images to real-time videos. SOT in SV is one of

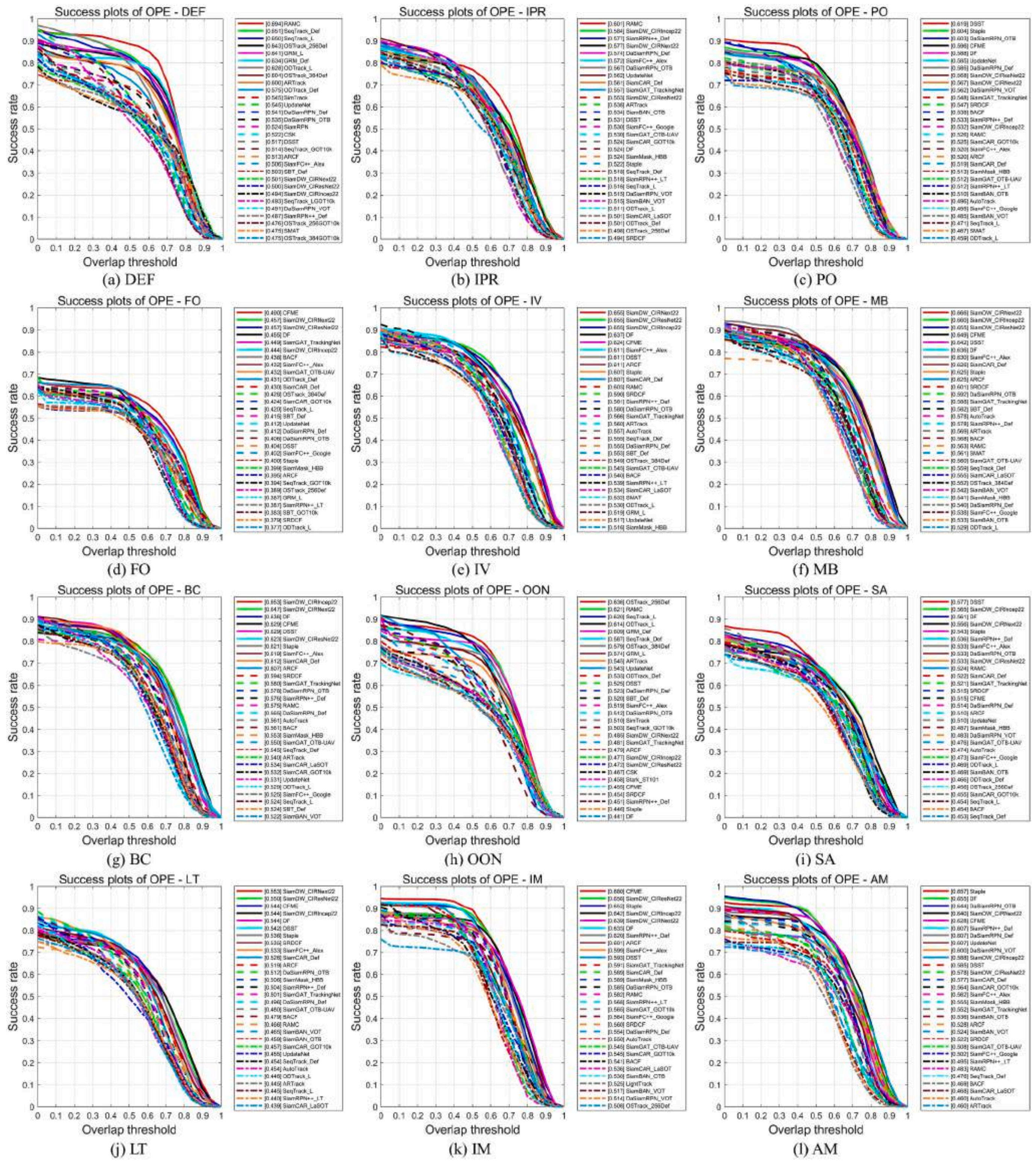

Fig. 16. The success plot of each attribute for the top 30 trackers. (a) DEF. (b) IPR. (c) PO. (d) FO. (e) IV. (f) MB. (g) BC. (h) OON. (i) SA. (j) LT. (k) IM. (l) AM.

the most fundamental tasks in the intelligent interpretation of remote sensing, holding potential applications in traffic sensing and modeling, wildfire suppression, sustainable fisheries (Shao et al., 2021), etc. Although some progress has been made, there are still issues that impede development. In this article, firstly, we review various tracking paradigms and frameworks that cover both general video and satellite video domains to discover potential prospects in the SV tracking domain. Secondly, we develop the first available oriented object tracking benchmark OOTB for SOT in SV. OOTB benchmarks various SOTAs and aims to identify the strengths of trackers and explore the intrinsic factors that contribute to effective tracking. Through comprehensive comparison and analysis, we have summarized several thoughts on how to facilitate SV object tracking.

Table 8
NPR scores for each attribute.

Tracker	Venue	Specific tracker name	DEF	IPR	PO	FO	IV	MB	BC	OON	SA	LT	IM	AM
CSK	ECCV 2012	CSK	0.628	0.562	0.444	0.387	0.568	0.575	0.547	0.583	0.447	0.395	0.493	0.370
SAMF	ECCV 2015	SAMF	0.534	0.555	0.606	0.491	0.571	0.630	0.582	0.568	0.533	0.403	0.608	0.444
DAT	CVPR 2015	DAT	0.555	0.501	0.525	0.395	0.596	0.584	0.537	0.508	0.426	0.424	0.468	0.445
KCF	TPAMI 2015	KCF	0.476	0.475	0.509	0.413	0.516	0.557	0.519	0.516	0.439	0.363	0.491	0.393
SRDCF	ICCV 2015	SRDCF	0.587	0.670	0.701	0.478	0.754	0.755	0.750	0.597	0.664	0.672	0.719	0.681
Staple	CVPR 2016	Staple	0.535	0.661	0.734	0.482	0.741	0.738	0.752	0.559	0.669	0.650	0.792	0.793
DSST	TPAMI 2017	DSST	0.654	0.700	0.770	0.515	0.752	0.781	0.780	0.680	0.724	0.699	0.771	0.752
BACF	ICCV 2017	BACF	0.585	0.638	0.769	0.542	0.747	0.778	0.774	0.584	0.659	0.700	0.786	0.736
SiamRPN	CVPR 2018	SiamRPN	0.672	0.624	0.509	0.347	0.618	0.613	0.586	0.528	0.517	0.447	0.579	0.437
DaSiamRPN	ECCV 2018	DaSiamRPN_Def	0.706	0.742	0.718	0.511	0.723	0.701	0.721	0.685	0.660	0.630	0.716	0.755
		DaSiamRPN_OTB	0.678	0.743	0.758	0.515	0.750	0.752	0.735	0.667	0.673	0.663	0.741	0.820
		DaSiamRPN_VOT	0.622	0.651	0.701	0.404	0.624	0.605	0.603	0.551	0.597	0.529	0.674	0.728
ARCF	ICCV 2019	ARCF	0.632	0.643	0.645	0.476	0.734	0.751	0.742	0.624	0.636	0.645	0.776	0.679
SiamRPN++	CVPR 2019	SiamRPN++_Def	0.614	0.730	0.664	0.431	0.735	0.716	0.722	0.578	0.667	0.641	0.776	0.768
		SiamRPN++_LT	0.557	0.661	0.628	0.474	0.686	0.658	0.640	0.552	0.565	0.565	0.721	0.622
UpdateNet	ICCV 2019	UpdateNet	0.714	0.728	0.709	0.509	0.676	0.654	0.680	0.716	0.653	0.571	0.632	0.758
SiamDW	CVPR 2019	SiamDW_CIResNet22	0.622	0.690	0.689	0.541	0.773	0.767	0.736	0.595	0.648	0.659	0.807	0.688
		SiamDW_CIRIncep22	0.594	0.725	0.642	0.531	0.775	0.774	0.772	0.590	0.674	0.643	0.769	0.691
		SiamDW_CIRNext22	0.610	0.728	0.689	0.536	0.773	0.778	0.761	0.599	0.665	0.660	0.760	0.759
SiamMask	CVPR 2019	SiamMask_HBB	0.579	0.703	0.650	0.515	0.700	0.708	0.733	0.586	0.636	0.665	0.760	0.717
		SiamMask_OBB	0.471	0.643	0.598	0.462	0.635	0.647	0.667	0.489	0.568	0.590	0.713	0.650
SiamBAN	CVPR 2020	SiamBAN_OTB	0.631	0.753	0.668	0.475	0.724	0.746	0.714	0.579	0.643	0.603	0.697	0.715
		SiamBAN_VOT	0.668	0.748	0.663	0.467	0.726	0.757	0.724	0.608	0.637	0.620	0.696	0.708
SiamFC++	AAAI 2020	SiamFC++_Alex	0.634	0.738	0.655	0.542	0.780	0.813	0.776	0.658	0.671	0.682	0.753	0.696
		SiamFC++_Google	0.519	0.678	0.644	0.531	0.651	0.688	0.661	0.579	0.599	0.476	0.695	0.636
AutoTrack	CVPR 2020	AutoTrack	0.610	0.652	0.666	0.471	0.720	0.736	0.741	0.598	0.632	0.636	0.766	0.633
CFME	TGRS 2020	CFME	0.418	0.558	0.726	0.575	0.727	0.749	0.736	0.556	0.616	0.653	0.819	0.766
SiamGAT	CVPR 2021	SiamGAT_GOT10k	0.440	0.681	0.591	0.459	0.692	0.728	0.686	0.477	0.613	0.545	0.777	0.651
		SiamGAT_LaSOT	0.381	0.391	0.343	0.255	0.505	0.478	0.430	0.297	0.383	0.300	0.373	0.287
		SiamGAT_OTB-UAV	0.570	0.740	0.690	0.549	0.753	0.768	0.733	0.577	0.655	0.634	0.774	0.690
		SiamGAT_TrackingNet	0.623	0.745	0.688	0.538	0.756	0.781	0.753	0.624	0.681	0.645	0.752	0.722
LightTrack	CVPR 2021	LightTrack	0.549	0.587	0.489	0.434	0.615	0.602	0.592	0.523	0.515	0.431	0.657	0.493
Stark	ICCV 2021	Stark_S50	0.488	0.498	0.350	0.417	0.605	0.602	0.560	0.488	0.389	0.426	0.379	0.352
		Stark_ST50	0.473	0.529	0.428	0.406	0.621	0.629	0.593	0.532	0.407	0.468	0.485	0.436
		Stark_ST101	0.578	0.550	0.448	0.470	0.638	0.646	0.603	0.578	0.446	0.476	0.551	0.462
SiamCAR	IJCV 2022	SiamCAR_Def	0.567	0.736	0.702	0.545	0.777	0.806	0.786	0.578	0.676	0.695	0.783	0.756
		SiamCAR_LaSOT	0.591	0.690	0.539	0.466	0.738	0.745	0.711	0.570	0.607	0.577	0.702	0.610
		SiamCAR_GOT10k	0.570	0.691	0.666	0.543	0.682	0.698	0.692	0.566	0.610	0.595	0.696	0.717
OSTrack	ECCV 2022	OSTrack_256Def	0.777	0.642	0.558	0.502	0.636	0.662	0.624	0.778	0.582	0.482	0.634	0.567
		OSTrack_256GOT10k	0.635	0.536	0.367	0.305	0.645	0.656	0.572	0.508	0.473	0.440	0.528	0.366
		OSTrack_384Def	0.734	0.617	0.528	0.541	0.709	0.714	0.662	0.703	0.558	0.563	0.606	0.554
		OSTrack_384GOT10k	0.625	0.545	0.298	0.303	0.607	0.608	0.540	0.510	0.410	0.396	0.341	0.282
SimTrack	ECCV 2022	SimTrack	0.668	0.505	0.388	0.373	0.588	0.569	0.508	0.615	0.422	0.366	0.420	0.341
DF	JSTARS 2022	DF	0.518	0.663	0.715	0.576	0.761	0.749	0.761	0.558	0.681	0.660	0.787	0.796
RAMC	RS 2022	RAMC	0.821	0.707	0.640	0.388	0.706	0.659	0.681	0.739	0.628	0.574	0.714	0.593
SBT	CVPR 2022	SBT_Def	0.635	0.586	0.466	0.538	0.723	0.770	0.681	0.659	0.546	0.575	0.589	0.516
		SBT_GOT10k	0.517	0.504	0.433	0.518	0.620	0.645	0.554	0.496	0.402	0.434	0.477	0.428
GRM	CVPR 2023	GRM_Def	0.764	0.599	0.536	0.386	0.650	0.671	0.592	0.734	0.517	0.467	0.555	0.486
		GRM_GOT10k	0.586	0.518	0.366	0.325	0.628	0.627	0.553	0.454	0.418	0.398	0.505	0.348
		GRM_L	0.788	0.635	0.511	0.497	0.684	0.693	0.625	0.709	0.533	0.490	0.518	0.472
SeqTrack	CVPR2023	SeqTrack_Def	0.817	0.683	0.590	0.497	0.731	0.749	0.710	0.743	0.598	0.611	0.634	0.630
		SeqTrack_GOT10k	0.689	0.607	0.503	0.526	0.619	0.659	0.603	0.677	0.513	0.452	0.511	0.450
		SeqTrack_L	0.804	0.679	0.623	0.547	0.703	0.714	0.695	0.764	0.606	0.594	0.633	0.607
		SeqTrack_LGOT10k	0.672	0.568	0.410	0.409	0.647	0.649	0.590	0.567	0.493	0.486	0.437	0.413
ARTrack	CVPR2023	ARTrack	0.756	0.722	0.567	0.473	0.761	0.775	0.717	0.679	0.604	0.597	0.635	0.620
ODTrack	AAAI2024	ODTrack_Def	0.734	0.669	0.598	0.568	0.684	0.667	0.675	0.681	0.606	0.572	0.668	0.563
		ODTrack_L	0.815	0.682	0.607	0.500	0.723	0.725	0.702	0.785	0.621	0.593	0.649	0.587
SMAT	WACV 2024	SMAT	0.601	0.620	0.608	0.479	0.679	0.709	0.635	0.536	0.507	0.538	0.574	0.516

7.1. Synergy of appearance information and motion cues

Combining appearance information and motion cues can effectively handle challenging tracking scenarios. In particular, the use of optical flow and historical trajectory is useful when objects exhibit motion properties similar to those of surrounding objects, but in opposite motion directions. One such scenario is anisotropic motion (i.e., AM), where an object moves with similar amplitude to surrounding objects, but in the opposite direction. In this case, relying on appearance information alone may not be sufficient to achieve accurate tracking. This is because there is little difference in the object's appearance. Therefore, it is difficult to distinguish it from the surrounding objects. By leveraging optical flow information, it is possible to discriminate the objects from the background based on their relative motion (Hu et al., 2020). Additionally, incorporating object trajectory information can further improve tracking performance in challenging scenarios. By analyzing the historical trajectory, it is able to predict future locations and adjust accordingly (Yang et al., 2023).

7.2. Dense object

The dense object is a significant challenge in the SV tracking domain. It is necessary to exploit more discriminative features to accurately identify and track objects (Song et al., 2022). One approach is to analyze the spatial distribution of the tracked object and surrounding objects. In this way, the tracker can identify and track the object of interest, even

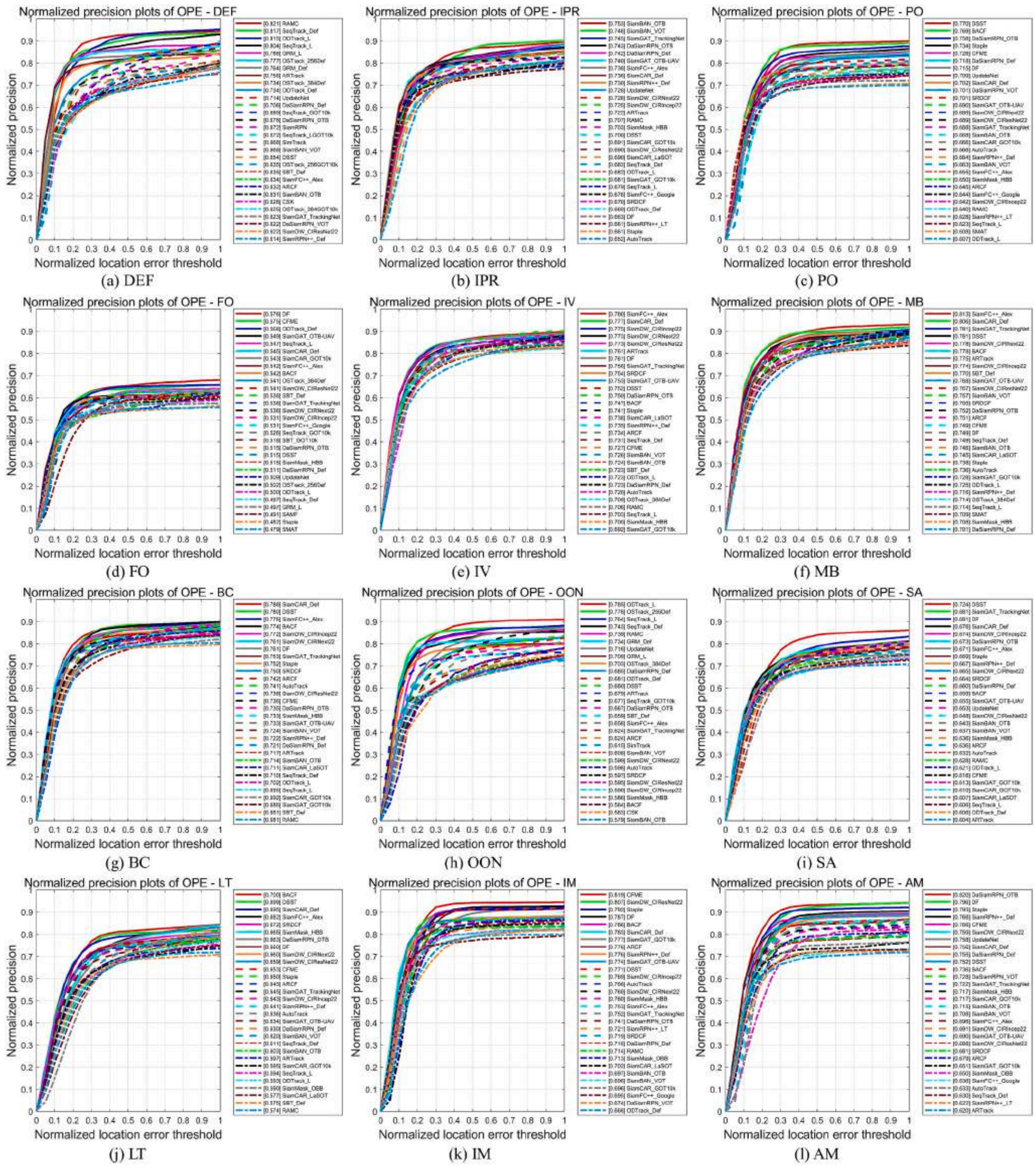

Fig. 17. The normalized precision plot of each attribute for the top 30 trackers. (a) DEF. (b) IPR. (c) PO. (d) FO. (e) IV. (f) MB. (g) BC. (h) OON. (i) SA. (j) LT. (k) IM. (l) AM.

when multiple objects are close. To accomplish this, the tracker should extract and analyze highly discriminative object features. In addition, a tracker is expected to possess the ability to detect the object's motion in its vicinity. This involves analyzing the motion patterns of nearby objects, as well as detecting any changes in their motion states. In summary, achieving accurate tracking under dense objects requires the tracker to exploit more discriminative features and detect the motion state of nearby objects.

7.3. Motion estimation

The non-stationary background is an issue for SOT in SV. With the high-speed moving platform and nadir view, trackers need to eliminate background motion and focus on the actual motion. Additionally, by

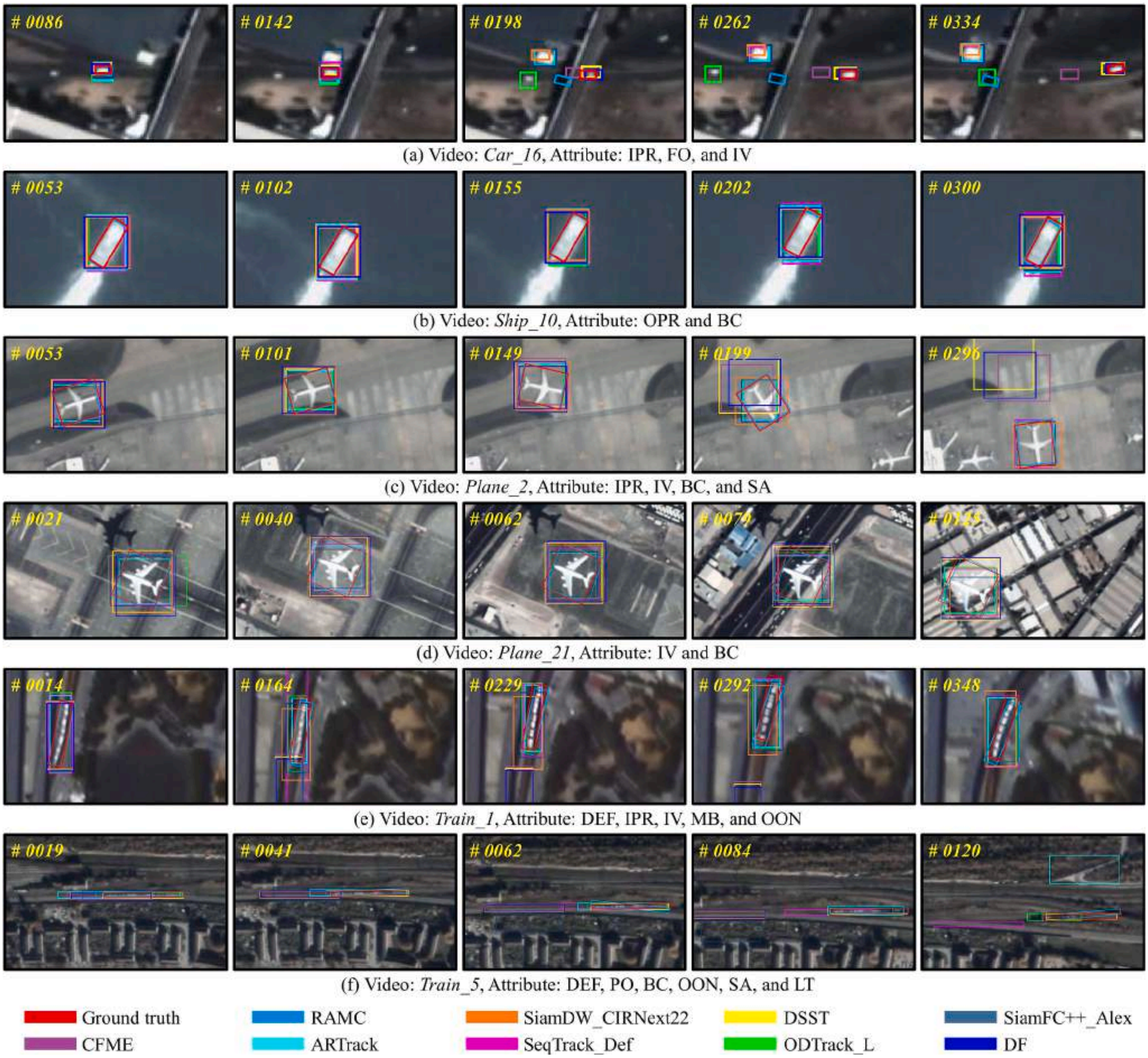

Fig. 18. Qualitative results for nine SOTAs. The current frame is shown in the upper left corner.

integrating multi-modal data such as Global Navigation Satellite System data (Zhou et al., 2021), high-resolution optical images (Guan et al., 2022), radar data (Garnot et al., 2022), and synthetic aperture radar data (Peng et al., 2023), we would capture precise geographic coordinates and object trajectories. As a result, the tracker can get a more complete picture of the object to predict its future motion and adjust accordingly.

7.4. Precise object representations

Most trackers can only generate HBB results that lack important semantic information such as orientation and shape. This can lead to performance degradation, particularly for SV objects with non-rigid deformation. Therefore, a superior tracker is expected to yield an accurate representation of the object, such as center position, scale, orientation, and shape (Chen et al., 2024). By leveraging these additional cues, it is possible to achieve more accurate and robust tracking performance.

7.5. Suitable backbones and features

In the SV tracking domain, the backbone and training dataset are typically borrowed from GV. Considering the significant differences between the SV and GV domains, it is urgent to develop the backbones and features suitable for SV object tracking. Moreover, pre-training using massive remote sensing datasets may result in substantial performance improvements.

7.6. Video enhancement

Video enhancement is a viable option for improving tracking performance. The tracker can benefit from enhancement processes such as video adjustment and reconstruction (Wang et al., 2023). In addition, the space–time super-resolution (Xiao et al., 2022) enables to obtain fine-grained features in both spatial and temporal dimensions, which provides rich motion cues and minimizes the risk of tracking drifts.

8. Conclusions

In this article, we first present a systematic review of tracking methods and datasets followed by introducing the proposed oriented object tracking benchmark OOTB. It is the first publicly available benchmark with high-quality oriented bounding box annotations and high-precision evaluation protocols in the satellite video single object tracking domain. OOTB includes 110 sequences captured by multiple satellite constellations, with a total of 29,890 frames, covering diverse object categories. To ensure comprehensive and fair evaluation, a protocol is proposed. We also benchmark 33 SOTA trackers including 58 models with different features, backbones, and tracker tags on OOTB. Extensive experiments and analysis are conducted in terms of the overall, category-based, and attribute-based results. Furthermore, the qualitative evaluation and running speed analysis demonstrate that satellite video object tracking remains challenging and far from being resolved. Finally, several thoughts on facilitating satellite video tracking tasks are summarized. It is believed that this work will spark interest in satellite video tracking, which in turn will lead to advances in remote sensing Earth observation. Future work will explore the fields of video object segmentation and scene recognition.

CRediT authorship contribution statement

Yuzeng Chen: Conceptualization, Methodology, Validation, Formal analysis, Writing - Original Draft, Visualization, Writing - Review & Editing. **Yuqi Tang:** Resources, Supervision, Data Curation, Writing - Review & Editing, Funding acquisition. **Yi Xiao:** Writing - Review & Editing, Visualization, Methodology, Project administration. **Qiangqiang Yuan:** Investigation, Supervision, Methodology, Writing - Review & Editing, Funding acquisition. **Yuwei Zhang:** Writing - Review & Editing. **Fengqing Liu:** Writing - Review & Editing. **Jiang He:** Writing - Review & Editing. **Liangpei Zhang:** Writing - Review & Editing, Validation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Sincerely thanks to the editors and reviewers for their valuable comments and suggestions. This work was supported by the National Natural Science Foundation of China under Grant 42230108 and 42271411. We are grateful to the institutions and companies for providing the satellite data.

References

Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.S., 2016a. Staple: Complementary Learners for Real-Time Tracking, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1401-1409.

Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S., 2016. Fully-convolutional siamese networks for object tracking. Proc. Eur. Conf. Comput. Vis. (ECCV) /IEEE Trans. Signal Process. 850–865.

Bhat, G., Danelljan, M., Van Gool, L., Timofte, R., Ieee, 2019. Learning Discriminative Model Prediction for Tracking, Ieee I Conf Comp Vis, pp. 6181-6190.

Bhat, G., Danelljan, M., Van Gool, L., Timofte, R., 2020. Know Your Surroundings: Exploiting Scene Information for Object Tracking. Proc. Eur. Conf. Comput. Vis. (ECCV).

Bi, F., Sun, J., Han, J., Wang, Y., Bian, M., 2021. Remote sensing target tracking in satellite videos based on a variable-angle-adaptive Siamese network. IET Image Processing 15, 1987–1997.

Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M., 2010. Visual object tracking using adaptive correlation filters. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 2544–2550.

Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the Devil in the Details: Delving Deep into Convolutional Nets. Proceedings of the British Machine Vision Conference.

Chen, Z.D., Zhong, B.N., Li, G.R., Zhang, S.P., Ji, R.R., Ieee, 2020. Siamese box adaptive network for visual tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6667-6676.

Chen, X., Yan, B., Zhu, J.W., Wang, D., Yang, X.Y., Lu, H.C., Ieee Comp, S.O.C., 2021. Transformer Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Electr Network, pp. 8122-8131.

Chen., Y., Tang., Y., Yin., Z., Han., T., Zou., B., Feng., H., 2022. Single Object Tracking in Satellite Videos: A Correlation Filter-Based Dual-Flow Tracker. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 6687-6698.

Chen, B.Y., Li, P.X., Sun, C., Wang, D., Yang, G., Lu, H.C., 2019. Multi attention module for visual tracking. Pattern Recognition 87, 80–93.

Chen, B., Li, P., Bai, L., Qiao, L., Shen, Q., Li, B., Gan, W., Wu, W., Ouyang, W., 2022a. Backbone is all your need: a simplified architecture for visual object tracking. Proc. Eur. Conf. Comput. Vis. (ECCV) 375–392.

Chen, X., Peng, H., Wang, D., Lu, H., Hu, H., 2023a. SeqTrack: Sequence to Sequence Learning for Visual Object Tracking. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14572–14581.

Chen, Y.Z., Tang, Y.Q., Han, T., Zhang, Y.W., Zou, B., Feng, H.H., 2022c. RAMC: A rotation adaptive tracker with motion constraint for satellite video single-object tracking. Remote Sens. 14, 3108.

Chen, Y., Tang, Y., Yuan, Q., Zhang, L., 2024. REPS: Rotation equivariant Siamese network enhanced by probability segmentation for satellite video tracking. International Journal of Applied Earth Observation and Geoinformation 128, 103741.

Chen, S.L., Wang, T.Y., Wang, H.S., Wang, Y.M., Hong, J.Z., Dong, T.C., Li, Z., 2022b. Vehicle tracking on satellite video based on historical model. IEEE J Sel. Top. Appl. Earth Obs. Remote Sens. 15, 7784–7796.

Chen, Y., Yuan, Q., Tang, Y., Xiao, Y., He, J., Zhang, L., 2023b. SPIRIT: spectral awareness interaction network with dynamic template for hyperspectral object tracking. IEEE Trans. Geosci. Remote Sens. 1–16.

Cheng, X., Zhang, M., Lin, S., Li, Y., Wang, H., 2024. Deep self-representation learning framework for hyperspectral anomaly detection. IEEE Trans. Instrum. Measur. 73, 1–16.

Cui, Y.Y., Hou, B.A., Wu, Q., Ren, B., Wang, S., Jiao, L.C., 2022b. Remote Sensing Object Tracking with Deep Reinforcement Learning Under Occlusion. IEEE Trans. Geosci. Remote Sens. 60.

Cui, Y., Guo, D.Y., Shao, Y.Y., Wang, Z.H., Shen, C.H., Zhang, L.Y., Chen, S.Y., 2022a. Joint classification and regression for visual tracking with fully convolutional siamese networks. Int. J. Comput. Vis. 130, 550–566.

Cui, Y., Jiang, C., Wu, G., Wang, L., 2024. MixFormer: end-to-end tracking with iterative mixed attention. IEEE Trans. Pattern Anal. Mach. Intell. 1–18.

Dai, K., Wang, D., Lu, H., Sun, C., Li, J., 2019. Visual Tracking via Adaptive Spatially-Regularized Correlation Filters, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR).

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 886–893.

Danelljan, M., Bhat, G., Gladh, S., Khan, F.S., Felsberg, M., 2019. Deep motion and appearance cues for visual tracking. Pattern Recognition Letters 124, 74–81.

Danelljan, M., Khan, F.S., Felsberg, M., van de Weijer, J., Adaptive Color Attributes for Real-Time Visual Tracking. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2014, pp. 1090-1097.

Danelljan, M., Hager, G., Khan, F.S., Felsberg, M., 2015a. Convolutional Features for Correlation Filter Based Visual Tracking, Proc. IEEE Int. Conf. Comput. Vis. (ICCV).

Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M., 2017a. ECO: Efficient Convolution Operators for Tracking, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6931-6939.

Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M., Soc, I.C., 2019b. ATOM: Accurate Tracking by Overlap Maximization, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, pp. 4655-4664.

Danelljan, M., Van Gool, L., Timofte, R., Ieee, 2020. Probabilistic Regression for Visual Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Electr Network, pp. 7181-7190.

Danelljan, M., Hager, G., Khan, F.S., Felsberg, M., 2015. Learning spatially regularized correlation filters for visual tracking. Proc. IEEE Int. Conf. Comput. Vis. (ICCV) 4310–4318.

Danelljan, M., Hager, G., Khan, F.S., Felsberg, M., 2017. Discriminative scale space tracking. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1561–1575.

Du, B., Sun, Y., Cai, S., Wu, C., Du, Q., 2018. Object tracking in satellite videos by fusing the kernel correlation filter and the three-frame-difference Algorithm. IEEE Geosci. Remote Sens. Lett. 15, 168–172.

Du, B., Cai, S.H., Wu, C., 2019. Object tracking in satellite videos based on a multiframe optical flow tracker. IEEE J Sel. Top. Appl. Earth Obs. Remote Sens. 12, 3043–3055.

Fan, Y., Qian, Y., Xie, F.-L., Soong, F.K., 2014. TTS Synthesis with Bidirectional LSTM Based Recurrent Neural Networks, pp. 1964-1968.

Fan, H., Ling, H.B., Soc, I.C., 2019b. Siamese Cascaded Region Proposal Networks for Real-Time Visual Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, pp. 7944-7953.

Fan, H., Lin, L.T., Yang, F., Chu, P., Deng, G., Yu, S.J., Bai, H.X., Xu, Y., Liao, C.Y., Ling, H.B., Soc, I.C., 2019. LaSOT: A high-quality benchmark for large-scale single object tracking. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 5369–5378.

Fan, H., Miththanthaya, H.A., Harshit, H., Rajan, S.R., Liu, X., Zou, Z., Lin, Y., Ling, H., 2021. Transparent object tracking benchmark. IEEE I Conf Comp Vis 10714–10723.

Feng, J., Zeng, D.N., Jia, X.P., Zhang, X.R., Li, J., Liang, Y.P., Jiao, L.C., 2021. Cross-frame keypoint-based and spatial motion information-guided networks for moving vehicle detection and tracking in satellite videos. ISPRS J. Photogramm. Remote Sens. 177, 116–130.

Fu, Z., Fu, Z., Liu, Q., Cai, W., Wang, Y., 2022. SparseTT: Visual Tracking with Sparse Transformers. ArXiv abs/2205.03776.

Fu, C.H., Xu, J.T., Lin, F.L., Guo, F.Y., Liu, T.C., Zhang, Z.J., 2020. Object saliency-aware dual regularized correlation filter for real-time aerial tracking. IEEE Trans. Geosci. Remote Sens. 58, 8940–8951.

Galoogahi, H.K., Sim, T., Lucey, S., 2015. Correlation Filters with Limited Boundaries. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 4630–4638.

Galoogahi, H.K., Fagg, A., Huang, C., Ramanan, D., Lucey, S., 2017a. Need for speed: a benchmark for higher frame rate object tracking. Ieee I Conf Comp Vis 1134–1143.

Galoogahi, H.K., Fagg, A., Lucey, S., 2017b. Learning background-aware correlation filters for visual tracking. IEEE I Conf Comp Vis 1144–1152.

Gao, J., Zhang, T., Xu, C., 2019. Graph convolutional tracking. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 4644–4654.

Gao, S., Zhou, C., Ma, C., Wang, X., Yuan, J., 2022. AiATrack: attention in attention for transformer visual tracking. Proc. Eur. Conf. Comput. Vis. (ECCV) 146–164.

Gao, S., Zhou, C., Zhang, J., 2023. Generalized Relation Modeling for Transformer Tracking. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18686–18695.

Garnot, V.S., Landrieu, L., Chehata, N., 2022. Multi-modal temporal attention models for crop mapping from satellite time series. ISPRS J. Photogramm. Remote Sens. 187, 294–305.

Guan, R., Li, Z., Li, X., Tang, C., Feng, R., 2023. Contrastive multi-view subspace clustering of hyperspectral images based on graph convolutional networks. arXiv e-prints, arXiv:2312.06068.

Guan, R., Li, Z., Li, T., Li, X., Yang, J., Chen, W., 2022. Classification of heterogeneous mining areas based on ResCapsNet and Gaofen-5 imagery. Remote Sens. 14, 3216.

Guo, J., Xu, T.F., Jiang, S.W., Shen, Z.Y., Ieee, 2018. Generating Reliable Online Adaptive Templates for Visual Tracking, 25th IEEE International Conference on Image Processing (ICIP), Athens, GREECE, pp. 226-230.

Guo, D.Y., Shao, Y.Y., Cui, Y., Wang, Z.H., Zhang, L.Y., Shen, C.H., Ieee Comp, S.O.C., 2021. Graph Attention Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Electr Network, pp. 9538-9547.

Guo, Y.J., Yang, D.Q., Chen, Z.Z., 2019. Object tracking on satellite videos: a correlation filter-based tracking method with trajectory correction by Kalman filter. IEEE J Sel. Top. Appl. Earth Obs. Remote Sens. 12, 3538–3551.

Han, B., Comaniciu, D., Zhu, Y., Davis, L.S., 2008. Sequential kernel density approximation and its application to real-time visual tracking. IEEE Trans. Pattern Anal. Mach. Intell. 30, 1186–1197.

He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J., Ieee, 2016. Deep Residual Learning for Image Recognition, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, pp. 770-778.

He, A.F., Luo, C., Tian, X.M., Zeng, W.J., Ieee, 2018. A Twofold Siamese Network for Real-Time Object Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, pp. 4834-4843.

He, Q.B., Sun, X., Yan, Z.Y., Li, B.B., Fu, K., 2022. Multi-object tracking in satellite videos with graph-based multitask modeling. IEEE Trans. Geosci. Remote Sens. 60.

Held, D., Thrun, S., Savarese, S., 2016. Learning to Track at 100 FPS with Deep Regression Networks, Proc. Eur. Conf. Comput. Vis. (ECCV), Amsterdam, NETHERLANDS, pp. 749-765.

Henriques, J.F., Caseiro, R., Martins, P., Batista, J., 2012. Exploiting the circulant structure of tracking-by-detection with kernels. Proc. Eur. Conf. Comput. Vis. (ECCV) 702–715.

Henriques, J.F., Caseiro, R., Martins, P., Batista, J., 2015. High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. 37, 583–596.

Hu, Z.P., Yang, D.Q., Zhang, K., Chen, Z.Z., 2020. Object tracking in satellite videos based on convolutional regression network with appearance and motion features. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 783–793.

Huang, Z.Y., Fu, C.H., Li, Y.M., Lin, F.L., Lu, P., 2019. Learning aberrance repressed correlation filters for real-time UAV tracking. Proc. IEEE Int. Conf. Comput. Vis. (ICCV) 2891–2900.

Huang, L.H., Zhao, X., Huang, K.Q., 2021. GOT-10k: a large high-diversity benchmark for generic object tracking in the wild. IEEE Trans. Pattern Anal. Mach. Intell. 43, 1562–1577.

Javed, S., Danelljan, M., Shahbaz Khan, F., Haris Khan, M., Felsberg, M., Matas, J., 2021. Visual Object Tracking with Discriminative Filters and Siamese Networks: A Survey and Outlook. arXiv e-prints, arXiv:2112.02838.

Javed, S., Danelljan, M., Khan, F.S., Khan, M.H., Felsberg, M., Matas, J., 2022. Visual Object Tracking with Discriminative Filters and Siamese Networks: A Survey and Outlook. IEEE Trans. Pattern Anal. Mach. Intell. PP.

Jepson, A.D., Fleet, D.J., El-Maraghi, T.F., 2003. Robust online appearance models for visual tracking. IEEE Trans. Pattern Anal. Mach. Intell. 25, 1296–1311.

Jiao, L., Zhang, X., Liu, X., Liu, F., Yang, S., Ma, W., Li, L., Chen, P., Feng, Z., Guo, Y., Tang, X., Hou, B., Zhang, X., Bai, J., Quan, D., Zhang, J., 2023. Transformer meets remote sensing video detection and tracking: a comprehensive survey. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 1–46.

Jung, I., Son, J., Baek, M., Han, B., 2018. Real-Time MDNet, Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, GERMANY, pp. 89-104.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. J. Basic Eng. 82, 35–45.

Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Zajc, L.C., Vojir, T., Bhat, G., Lukezic, A., Eldesokey, A., Fernandez, G., Garcia-Martin, A., Iglesias-Arias, A., Alatan, A.A., Gonzalez-Garcia, A., Petrosino, A., Memarmoghadam, A., Vedaldi, A., Muhic, A., He, A.F., Smeulders, A., Perera, A.G., Li, B., Chen, B.Y., Kim, C., Xu, C. S., Xiong, C.Z., Tian, C., Luo, C., Sun, C., Hao, C., Kim, D., Mishra, D., Chen, D.M., Wang, D., Wee, D., Gavves, E., Gundogdu, E., Velasco-Salido, E., Khan, F.S., Yang, F., Zhao, F., Li, F., Battistone, F., De Ath, G., Subrahmanyam, G., Bastos, G., Ling, H.B., Galoogahi, H.K., Lee, H., Li, H.J., Zhao, H.J., Fan, H., Zhang, H.G., Possegger, H., Li, H.Q., Lu, H.C., Zhi, H., Li, H.Y., Lee, H., Chang, H.J., Drummond, I., Valmadre, J., Martin, J.S., Chahl, J., Choi, J.Y., Li, J., Wang, J.Q., Qi, J.Q., Sung, J., Johnander, J., Henriques, J., Choi, J., van de Weijer, J., Herranz, J.R., Martinez, J.M., Kittler, J., Zhuang, J.F., Gao, J.Y., Grm, K., Zhang, L.C., Wang, L.J., Yang, L.X., Rout, L., Si, L., Bertinetto, L., Chu, L.T., Che, M.Q., Maresca, M.E., Danelljan, M., Yang, M.H., Abdelpakey, M., Shehata, M., Kang, M., Lee, N., Wang, N., Miksik, O., Moallem, P., Vicente-Monivar, P., Senna, P., Li, P.X., Torr, P., Raju, P.M., Qian, R.H., Wang, Q., Zhou, Q., Guo, Q., Martin-Nieto, R., Gorthi, R.K., Tao, R., Bowden, R., Everson, R., Wang, R.L., Yun, S., Choi, S., Vivas, S., Bai, S., Huang, S.P., Wu, S.H., Hadfield, S., Wang, S.W., Golodetz, S., Ming, T., Xu, T.Y., Zhang, T.Z., Fischer, T., Santopietro, V., Struc, V., Wei, W., Zuo, W.M., Feng, W., Wu, W., Zou, W., Hu, W.M., Zhou, W.G., Zeng, W.J., Zhang, X.F., Wu, X.H., Wu, X.J., Tian, X.M., Li, Y., Lu, Y., Law, Y.W., Wu, Y., Demiris, Y., Yang, Y.C., Jiao, Y.F., Li, Y.H., Zhang, Y.H., Sun, Y.X., Zhang, Z., Zhu, Z., Feng, Z.H., Wang, Z.H., He, Z.Q., 2018. The Sixth Visual Object Tracking VOT2018 Challenge Results, Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, GERMANY, pp. 3-53.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 84–90.

Kugarajeevan, J., Kokul, T., Ramanan, A., Fernando, S., 2023. Transformers in single object tracking: an experimental survey. IEEE Access 11, 80297–80326.

Law, H., Deng, J., 2020. CornerNet: detecting objects as paired keypoints. Int. J. Comput. Vis. 128, 642–656.

Li, Y.F., Bian, C.J., 2022. Object Tracking in Satellite Videos: A Spatial-Temporal Regularized Correlation Filter Tracking Method With Interacting Multiple Model. IEEE Geosci. Remote Sens. Lett. 19.

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J., 2019a. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4277-4286.

Li, Y., Zhu, J., Hoi, S.C.H., Song, W., Wang, Z., Liu, H., Aaai, 2019c. Robust Estimation of Similarity Transformation for Visual Object Tracking, 33rd AAAI Conference on Artificial Intelligence / 31st Innovative Applications of Artificial Intelligence Conference / 9th AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, HI, pp. 8666-8673.

Li, Y.F., Bian, C.J., Chen, H.Z., 2022b. Object Tracking in Satellite Videos: Correlation Particle Filter Tracking Method With Motion Estimation by Kalman Filter. IEEE Trans. Geosci. Remote Sens. 60.

Li, Y.X., Jiao, L.C., Huang, Z.J., Zhang, X., Zhang, R.H., Song, X., Tian, C.X., Zhang, Z.X., Liu, F., Shuyuan, Y., Hou, B., Ma, W.P., Liu, X., Li, L.L., 2022c. Deep Learning-Based Object Tracking in Satellite Videos: A Comprehensive Survey With a New Dataset. IEEE Geosci. Remote Sens. Mag.

Li, Y., Fu, C., Ding, F., Huang, Z., Lu, G., 2020. AutoTrack: Towards High-Performance Visual Tracking for UAV with Automatic Spatio-Temporal Regularization. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).

Li, A.N., Lin, M., Wu, Y., Yang, M.H., Yan, S.C., 2016. NUS-PRO: a new visual tracking challenge. IEEE Trans. Pattern Anal. Mach. Intell. 38, 335–349.

Li, X., Ma, C., Wu, B., He, Z., Yang, M.-H., 2019b. Target-Aware Deep Tracking. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).

Li, F., Tian, C., Zuo, W., Zhang, L., Yang, M.-H., 2018b. Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).

Li, C.L., Xue, W.L., Jia, Y.Q., Qu, Z.C., Luo, B., Tang, J., Sun, D.D., 2022a. LasHeR: A large-scale high-diversity benchmark for RGBT tracking. IEEE Trans. Image Process. 31, 392–404.

Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X., 2018a. High performance visual tracking with siamese region proposal network. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 8971–8980.

Li, S., Zhou, Z., Zhao, M., Yang, J., Guo, W., Lv, Y., Kou, L., Wang, H., Gu, Y., 2023. A multi-task benchmark dataset for satellite video: object detection, tracking, and segmentation. IEEE Trans. Geosci. Remote Sens. 1-1.

Li, Y., Zhu, J., 2015. A scale adaptive kernel correlation filter tracker with feature integration. Proc. Eur. Conf. Comput. Vis. (ECCV) 254–265.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C. L., 2014. Microsoft COCO: Common Objects in Context, Proc. Eur. Conf. Comput. Vis. (ECCV), Zurich, SWITZERLAND, pp. 740-755.

Lin, L., Fan, H., Zhang, Z., Xu, Y., Ling, H., 2022. SwinTrack: a simple and strong baseline for transformer tracking. NeurIPS 16 (743–716), 754.

Lin, S., Zhang, M., Cheng, X., Shi, L., Gamba, P., Wang, H., 2023. Dynamic low-rank and sparse priors constrained deep autoencoders for hyperspectral anomaly detection. IEEE Trans. Instrum. Measur. 1-1.

Liu, Y.S., Liao, Y.R., Lin, C.B., Li, Z.M., Yang, X.Y., Zhang, A.D., Ieee, 2021. Object Tracking in Satellite Videos Based on Improved Correlation Filters. 2021 13th International Conference on Communication Software and Networks (ICCSN), 323-331.

Liu, Y.S., Liao, Y.R., Lin, C.B., Jia, Y.T., Li, Z.M., Yang, X.Y., 2022. Object Tracking in Satellite Videos Based on Correlation Filter with Multi-Feature Fusion and Motion Trajectory Compensation. Remote Sens. 14.

Lukezic, A., Matas, J., Kristan, M., Ieee, 2020. D3S-A Discriminative Single Shot Segmentation Tracker, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Electr Network, pp. 7131-7140.

Lukezic, A., Vojir, T., Zajc, L.C., Matas, J., Kristan, M., 2018. Discriminative correlation filter tracker with channel and spatial reliability. Int. J. Comput. Vis. 126, 671–688.

Ma, C., Huang, J.B., Yang, X.K., Yang, M.H., Ieee, 2015a. Hierarchical Convolutional Features for Visual Tracking, IEEE International Conference on Computer Vision, Santiago, CHILE, pp. 3074-3082.

Ma, C., Yang, X.K., Zhang, C.Y.Y., Yang, M.H., 2015b. Long-term Correlation Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, pp. 5388-5396.

Ma, N.N., Zhang, X.Y., Zheng, H.T., Sun, J., 2018b. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design, 15th European Conference on Computer Vision (ECCV), Munich, GERMANY, pp. 122-138.

Ma, D., Bu, W., Wu, X., 2018a. Multi-Scale Recurrent Tracking via Pyramid Recurrent Network and Optical Flow, p. 242.

Marchi, E., Ferroni, G., Eyben, F., Gabrielli, L., Squartini, S., Schuller, B., Ieee, 2014. Multi-Resolution Linear Prediction Based Features for Audio Onset Detection with Bidirectional LSTM Neural Networks, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, ITALY.

Marvasti-Zadeh, S.M., Cheng, L., Ghanei-Yakhdan, H., Kasaei, S., 2022. Deep learning for visual tracking: a comprehensive survey. IEEE Trans. Intell. Transp. Syst. 23, 3943–3968.

Mayer, C., Danelljan, M., Bhat, G., Paul, M., Pani Paudel, D., Yu, F., Van Gool, L., 2022. Transforming Model Prediction for Tracking. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).

Mayer, C., Danelljan, M., Pani Paudel, D., Van Gool, L., 2021. Learning Target Candidate Association to Keep Track of What Not to Track. in Proc. IEEE Int. Conf. Comput. Vis. (ICCV).

Mueller, M., Smith, N., Ghanem, B., 2016. A Benchmark and Simulator for UAV Tracking, Proc. Eur. Conf. Comput. Vis. (ECCV), Amsterdam, NETHERLANDS, pp. 445-461.

Mueller, M., Smith, N., Ghanem, B., 2017. Context-Aware Correlation Filter Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, pp. 1387-1395.

Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B., 2018. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild, 15th European Conference on Computer Vision (ECCV), Munich, GERMANY, pp. 310-32S7.

Nam, H., Baek, M., Han, B., 2016. Modeling and Propagating CNNs in a Tree Structure for Visual Tracking. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).

Nam, H., Han, B., 2016. Learning multi-domain convolutional neural networks for visual tracking. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 4293–4302.

Peng, Y.F., He, J., Yuan, Q.Q., Wang, S.X., Chu, X.D., Zhang, L.P., 2023. Automated glacier extraction using a Transformer based deep learning approach from multi-sensor remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 202, 303–313.

Pham, H., Guan, M.Y., Zoph, B., Le, Q.V., Dean, J., 2018. Efficient Neural Architecture Search via Parameter Sharing, 35th International Conference on Machine Learning (ICML), Stockholm, SWEDEN.

Possegger, H., Mauthner, T., Bischof, H., Ieee, 2015. In Defense of Color-based Model-free Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2113-2120.

Ren, S.Q., He, K.M., Girshick, R., Sun, J., 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137–1149.

Ruan, L., Guo, Y.J., Yang, D.Q., Chen, Z.Z., 2022. Deep Siamese Network with Motion Fitting for Object Tracking in Satellite Videos. IEEE Geosci. Remote Sens. Lett. 19.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z.H., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

Shao, J., Du, B., Wu, C., Pingkun, Y., 2019a. PASiam: Predicting Attention Inspired Siamese Network, for Space-Borne Satellite Video Tracking. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 1504–1509.

Shao, J., Du, B., Wu, C., Zhang, L., 2019b. Can we track targets from space? A hybrid kernel correlation filter tracker for satellite video. IEEE Trans. Geosci. Remote Sens. 57, 8719–8731.

Shao, J., Du, B., Wu, C., Zhang, L.F., 2019c. Tracking objects from satellite videos: a velocity feature based correlation filter. IEEE Trans. Geosci. Remote Sens. 57, 7860–7871.

Shao, J., Du, B., Wu, C., Gong, M., Liu, T., 2021. HRSiam: high-resolution siamese network, towards space-borne satellite video tracking. IEEE Trans. Image Process. 30, 3056–3068.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. in ICLR 2015.

Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M., 2014. Visual tracking: an experimental survey. IEEE Trans. Pattern Anal. Mach. Intell. 36, 1442–1468.

Song, W., Jiao, L.C., Liu, F., Liu, X., Li, L.L., Yang, S.Y., Hou, B.A., Zhang, W.H., 2022. A Joint Siamese Attention-Aware Network for Vehicle Object Tracking in Satellite Videos. IEEE Trans. Geosci. Remote Sens. 60.

Song, Y.B., Ma, C., Wu, X.H., Gong, L.J., Bao, L.C., Zuo, W.M., Shen, C.H., Lau, R.W.H., Yang, M.H., 2018. VITAL: visual tracking via adversarial learning. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) 8990–8999.

Sundermeyer, M., Schlueter, R., Ney, H., 2012. LSTM Neural Networks for Language Modeling, 13th Annual conference of the International Speech Communication Association 2012, vol. 2: 13th Annual conference of the International Speech Communication Association 2012 (INTERSPEECH 2012), 9-13 September 2012, Portland, Oregon, USA, Portland, OR(US), pp. 194-197.

Szegedy, C., Liu, W., Jia, Y.Q., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., Ieee, 2015. Going Deeper with Convolutions, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, pp. 1-9.

Tao, R., Gavves, E., Smeulders, A.W.M., Ieee, 2016. Siamese Instance Search for Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, pp. 1420-1429.

Tian, Z., Shen, C.H., Chen, H., He, T., Ieee, 2019. FCOS: Fully Convolutional One-Stage Object Detection, Ieee I Conf Comp Vis, Seoul, SOUTH KOREA, pp. 9626-9635.

Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., Torr, P.H.S., Ieee, 2017. (CF-NET) End-to-end representation learning for Correlation Filter based tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 5000-5008.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need, 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA.

Wang, F., Cao, P., Li, F., Wang, X., He, B., Sun, F., 2022a. WATB: Wild Animal Tracking Benchmark. Int. J. Comput, Vis.

Wang, J.X., Chen, S.B., Ding, C.H.Q., Tang, J., Luo, B., 2022b. RanPaste: paste consistency and pseudo label for semisupervised remote sensing image semantic segmentation. IEEE Trans. Geosci. Remote Sens. 60, 1–16.

Wang, N., Zhou, W.G., Tian, Q., Hong, R.C., Wang, M., Li, H.Q., Ieee, 2018. Multi-Cue Correlation Filters for Robust Visual Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, pp. 4844-4853.

Wang, N., Song, Y.B., Ma, C., Zhou, W.G., Liu, W., Li, H.Q., Soc, I.C., 2019b. Unsupervised Deep Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, pp. 1308-1317.

Wang, N., Zhou, W.G., Wang, J., Li, H.Q., Ieee Comp, S.O.C., 2021. Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Electr Network, pp. 1571-1580.

Wang, J.X., Li, T., Chen, S.B., Tang, J., Luo, B., Wilson, R.C., 2022c. Reliable contrastive learning for semi-supervised change detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Wang, G., Luo, C., Xiong, Z., Zeng, W., 2019a. SPM-tracker: series-parallel matching for real-time visual object tracking. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 3638–3647.

Wang, H., Sun, S.X., Ren, P., 2023. Meta underwater camera: A smart protocol for underwater image enhancement. ISPRS J. Photogramm. Remote Sens. 195, 462–481.

Wang, Y.M., Wang, T.Y., Zhang, G., Cheng, Q., Wu, J.Q., 2020. Small target tracking in satellite videos using background compensation. IEEE Trans. Geosci. Remote Sens. 58, 7010–7021.

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.S., 2019c. Fast Online Object Tracking and Segmentation: a Unifying Approach. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 1328–1338.

Wei, X., Bai, Y., Zheng, Y., Shi, D., Gong, Y., 2023. Autoregressive Visual Tracking, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9697-9706.

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T.S., Yan, S.C., 2010. Sparse representation for computer vision and pattern recognition. Proc. IEEE 98, 1031–1044.

Wu, J.L., Su, X., Yuan, Q.Q., Shen, H.F., Zhang, L.P., 2022. Multivehicle Object Tracking in Satellite Video Enhanced by Slow Features and Motion Features. IEEE Trans. Geosci. Remote Sens. 60.

Wu, Y., Lim, J., Yang, M.H., 2013. Online object tracking: a benchmark. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 2411–2418.

Wu, Y., Lim, J., Yang, M.H., 2015. Object tracking benchmark. IEEE Trans. Pattern Anal. Mach. Intell. 37, 1834–1848.

Xiao, Y., Yuan, Q., He, J., Zhang, Q., Sun, J., Su, X., Wu, J., Zhang, L., 2022. Space-time super-resolution for satellite video: a joint framework based on multi-scale spatial-temporal transformer. Int. J. Appl. Earth Obs. Geoinf. 108, 102731.

Xie, S.N., Girshick, R., Dollar, P., Tu, Z.W., He, K.M., Ieee, 2017. Aggregated Residual Transformations for Deep Neural Networks, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, pp. 5987-5995.

Xie, F., Wang, C., Wang, G., Yang, W., Zeng, W., 2021. Learning Tracking Representations via Dual-Branch Fully Transformer Networks, 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 2688-2697.

Xie, F., Wang, C., Wang, G., Cao, Y., Yang, W., Zeng, W., 2022. Correlation-Aware Deep Tracking, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8741-8750.

Xie, F., Chu, L., Li, J., Lu, Y., Ma, C., 2023. VideoTrack: Learning to Track Objects via Video Transformer, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22826-22835.

Xing, D., Evangeliou, N., Tsoukalas, A., Tzes, A., 2022. Siamese Transformer Pyramid Networks for Real-Time UAV Tracking, 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1898-1907.

Xu, N., Yang, L.J., Fan, Y.C., Yang, J.C., Yue, D.C., Liang, Y.C., Price, B., Cohen, S., Huang, T., 2018. YouTube-VOS: Sequence-to-Sequence Video Object Segmentation, 15th European Conference on Computer Vision (ECCV), Munich, GERMANY, pp. 603-619.

Xu, Y.D., Wang, Z.Y., Li, Z.X., Yuan, Y., Yu, G., Assoc Advancement Artificial, I., 2020. SiamFC plus plus : Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines, Proc. AAAI Conf. Artif. Intell. (AAAI), pp. 12549-12556.

Xu, T.Y., Feng, Z.H., Wu, X.J., Kittler, J., 2019. Joint group feature selection and discriminative filter learning for robust visual object tracking. IEEE I Conf Comp Vis 7949–7959.

Xuan, S.Y., Li, S.Y., Han, M.F., Wan, X., Xia, G.S., 2020. Object tracking in satellite videos by improved correlation filters with motion estimations. IEEE Trans. Geosci. Remote Sens. 58, 1074–1086.

Xuan, S.Y., Li, S.Y., Zhao, Z.F., Zhou, Z., Zhang, W.F., Tan, H., Xia, G.S., Gu, Y.F., 2021. Rotation adaptive correlation filter for moving object tracking in satellite videos. Neurocomputing 438, 94–106.

Yan, B., Peng, H., Fu, J., Wang, D., Lu, H., 2021a. Learning Spatio-Temporal Transformer for Visual Tracking. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).
Yan, B., Peng, H., Wu, K., Wang, D., Fu, J., Lu, H., 2021. LightTrack: finding lightweight neural networks for object tracking via one-shot architecture search. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 15175–15184.
Yang, T.Y., Chan, A.B., Ieee, 2017. Recurrent Filter Learning for Visual Tracking, Ieee I Conf Comp Vis, Venice, ITALY, pp. 2010-2019.
Yang, J., Pan, Z., Wang, Z., Lei, B., Hu, Y., 2023. SiamMDM: an adaptive fusion network with dynamic template for real-time satellite video single object tracking. IEEE Trans. Geosci. Remote Sens. 61, 1–19.
Ye, B., Chang, H., Ma, B., Shan, S., 2022a. Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework. in Proc. Eur. Conf. Comput. Vis. (ECCV).
Ye, M., Shen, J.B., Lin, G.J., Xiang, T., Shao, L., Hoi, S.C.H., 2022. Deep learning for person re-identification: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. 44, 2872–2893.
Yelluru Gopal, G., Amer, M.A., 2024. Separable Self and Mixed Attention Transformers for Efficient Object Tracking. in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 6708-6717.
Yin, Q., Hu, Q.Y., Liu, H., Zhang, F., Wang, Y.Q., Lin, Z.P., An, W., Guo, Y.L., 2022. Detecting and Tracking Small and Dense Moving Objects in Satellite Videos: A Benchmark. IEEE Trans. Geosci. Remote Sens. 60.
You, S., Zhu, H., Li, M., Li, Y., 2019. A Review of Visual Trackers and Analysis of its Application to Mobile Robot. arXiv e-prints, arXiv:1910.09761.
Yun, S., Choi, J., Yoo, Y., Yun, K., Choi, J.Y., 2018. Action-driven visual object tracking with deep reinforcement learning. IEEE Trans. Neural Netw. Learn. Syst. 29, 2239–2252.
Zhang, L.C., Gonzalez-Garcia, A., Van De Weijer, J., Danelljan, M., Khan, F.S., Ieee, 2019a. Learning the Model Update for Siamese Trackers, Ieee I Conf Comp Vis, Seoul, SOUTH KOREA, pp. 4009-4018.
Zhang, Z., Peng, H., Fu, J., Li, B., Hu, W., 2020. Ocean: Object-aware Anchor-free Tracking. Proc. Eur. Conf. Comput. Vis. (ECCV).
Zhang, W.H., Jiao, L.C., Liu, F., Li, L.L., Liu, X., Liu, J., 2022. MBLT: Learning Motion and Background for Vehicle Tracking in Satellite Videos. IEEE Trans. Geosci. Remote Sens. 60.
Zhang, X.D., Zhu, K., Chen, G.Z., Liao, P.Y., Tan, X.L., Wang, T., Li, X.W., 2023. High-resolution satellite video single object tracking based on thicksiam framework. GIsci. Remote Sens. 60.
Zhang, Z.P., Peng, H.W., Soc, I.C., 2019b. Deeper and Wider Siamese Networks for Real-Time Visual Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, pp. 4586-4595.
Zhao, M.Q., Li, S.Y., Xuan, S.Y., Kou, L.X., Gong, S., Zhou, Z., 2022. SatSOT: A Benchmark Dataset for Satellite Video Single Object Tracking. IEEE Trans. Geosci. Remote Sens. 60.
Zhao, F., Wang, J.Q., Wu, Y., Tang, M., 2019. Adversarial deep tracking. IEEE Trans. Circuits Syst. Video Technol. 29, 1998–2011.
Zheng, Y., Zhong, B., Liang, Q., Mo, Z., Zhang, S., Li, X., 2024. ODTrack: Online Dense Temporal Token Learning for Visual Tracking. in Proc. AAAI Conf. Artif. Intell. (AAAI).
Zhou, T., Hasheminasab, S.M., Habib, A., 2021. Tightly-coupled camera/LiDAR integration for point cloud generation from GNSS/INS-assisted UAV mapping systems. ISPRS J. Photogramm. Remote Sens. 180, 336–356.
Zhu, Z., Wu, W., Zou, W., Yan, J.J., Ieee, 2018b. End-to-end flow correlation tracking with spatial-temporal attention, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, pp. 548-557.
Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W., 2018. Distractor-aware siamese networks for visual object tracking. Proc. Eur. Conf. Comput. Vis. (ECCV) 103–119.
Zhu, K., Zhang, X.D., Chen, G.Z., Tan, X.L., Liao, P.Y., Wu, H.Y., Cui, X.J., Zuo, Y.A., Lv, Z.Y., 2021. Single object tracking in satellite videos: deep siamese network incorporating an interframe difference centroid inertia motion model. Remote Sens. 13.