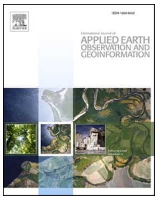
International Journal of Applied Earth Observation and Geoinformation 128 (2024) 103741

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



REPS: Rotation equivariant Siamese network enhanced by probability segmentation for satellite video tracking

Yuzeng Chen[a], Yuqi Tang[b,*], Qiangqiang Yuan[a], Liangpei Zhang[c]

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

ARTICLE INFO

Keywords:
Video satellite
Object tracking
Siamese network
Semantic representation

ABSTRACT

Satellite video is an emerging surface observation data that has drawn increasing interest due to its potential in spatiotemporal dynamic analysis. Single object tracking of satellite videos allows the continuous acquisition of the positions and ranges of objects and establishes the correspondences in the video sequence. However, small-sized objects are vulnerable to rotation and non-rigid deformation. Moreover, the horizontal bounding box of most trackers has difficulty in providing accurate semantic representations such as object position, orientation, and spatial distribution. In this article, we propose a unified framework, named rotation equivariant Siamese network enhanced by probability segmentation (REPS), to enhance the tracking accuracy and semantic representations simultaneously. First, to deal with the inconsistency of representations, we design a rotation equivariant (RE) Siamese network architecture to detect the rotation variations of objects right from the start frame, achieving the RE tracking. Second, a pixel-level (PL) refinement is proposed to refine the spatial distribution of objects. In addition, we proposed an adaptive Gaussian fusion that synergizes tracking and segmentation results to obtain compact outputs for satellite object representations. Extensive experiments on satellite videos demonstrate the superiority of the proposed approach. The code will be available at https://github.com/YZCU/REPS

1. Introduction

Satellite video (SV), being an emerging remote sensing data, provides a wealth of spatiotemporal information on specific scenarios (Xiao et al., 2022a). It can be applied to diverse scenarios such as traffic surveillance, stereo mapping, and disaster response. Launched by Skybox Imaging in 2013, the SkySat-1 satellite can shoot the panchromatic SV with a 1.1 m ground sample distance (GSD) and 30 frames per second (FPS). While the Jilin-1 satellite constellation can capture 30 FPS RGB videos with 0.92 m GSD. Recently, the Luojia-3-01 satellite has been launched, which has the capability of multi-mode optical imaging, intelligent processing in orbit, and real-time transmission. The advancement of satellites has enriched the remote sensing observation capability (He et al., 2023).

Single object tracking (SOT) of SV, one of the most fundamental tasks of remote sensing intelligence interpretation, has potential prospects in traffic awareness and simulation, fighting wildfires, sustainable fishing (Shao et al., 2021), etc. Given the starting object state, the SOT task seeks to determine its subsequent states and establish the object correspondences in the video (Chen et al., 2023b). However, it also involves some challenges. First, considering the long-distance imaging platform, the spatial and spectral resolution of SV is limited and results in objects with finite features (Chen. et al., 2022; Xiao et al., 2022b). Second, the low contrast between foreground and background restricts the discriminability of objects (Yin et al., 2022). Moreover, objects are susceptible to rotation and non-rigid deformation in the nadir view, which can change the spatial layout of objects and cause tracking drift. Due to these issues, trackers tend to experience tracking failures and produce unsatisfactory performance.

Existing methods of SOT in SV mainly include two paradigms: the correlation filter (CF) and deep learning (DL). CF-based paradigms start by pre-training a filter on all samples and then use it to track the object (Javed et al., 2022). Several trackers are inheriting the CF-based paradigms. For instance, VCF (Shao et al., 2019b) and HKCF (Shao et al., 2019a) incorporate the kernelized correlation filter (KCF) (Henriques et al., 2015) with Lucas–Kanade optical flow (Patel and Upadhyay, 2013) for satellite object tracking and achieve superior results. KCF_TFD (Du et al., 2018), CFME (Xuan et al., 2020), and HMTS (Chen et al.,

[*] Corresponding author.

https://doi.org/10.1016/j.jag.2024.103741
Received 1 October 2023; Received in revised form 20 February 2024; Accepted 26 February 2024

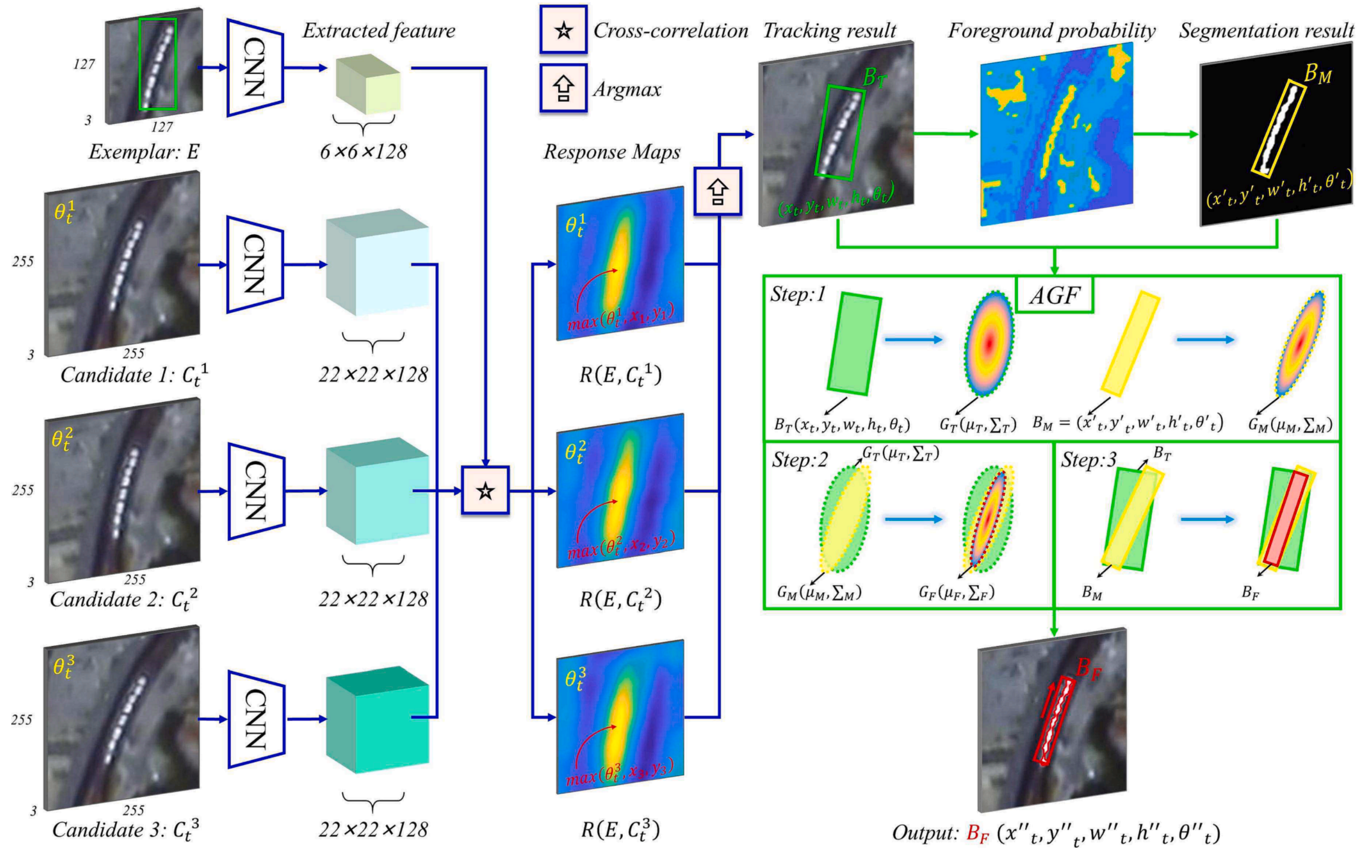


Fig. 1. Overview of REPS. It consists of RE tracking connected with blue lines and PL refinement connected with green lines. The RE tracking enumerates several candidates C_t^i with θ_t^i. The features extracted from exemplar E and candidates C are cross-correlated to yield a group of response maps $R(E, C)$. The preliminary tracking results with OBB are obtained by finding the maximum of $R(E, C)$. It is then refined by the PL refinement that can generate the mask from the object probability map. Meanwhile, it imposes an AGF and outputs the final result B_F. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2022b) also use the KCF to implement SOT in SV. While WTIC (Wang et al., 2020), CFKF (Guo et al., 2019), DF (Chen. et al., 2022), and CPKF (Li et al., 2022) are modeled on the CSK (Henriques et al., 2012), DSST (Danelljan et al., 2014), Staple (Bertinetto et al., 2016a), and STRCF (Li et al., 2018b), respectively, obtaining the competitive speed. These trackers typically exploit static intra-frame features and dynamic inter-frame features for tracking SV objects. And they jointly use hand-crafted features such as histogram of oriented gradients (HOG) (Dalal and Triggs, 2005), color names (CN) (Danelljan et al., 2014b), and color histogram (CH) (Bertinetto et al., 2016a) and the motion models such as Kalman filter (Kalman, 1960), motion smoothness (Wang et al., 2020), and motion trajectory averaging (Xuan et al., 2020). It is noted that most trackers inherit the CF and exploit hand-crafted features for SV object tracking, which may yield unsatisfactory accuracy due to the limitation of hand-crafted features (Shao et al., 2021). Recently, the DL-based paradigms have drawn a great deal of attention due to their robust object representation. For instance, AD-OHNet (Cui et al., 2022) blends the spatiotemporal contexts, appearance model, and motion vectors to extract discriminative features to obtain precise object position. In addition, the Siamese neural network (SNN) is widely used for satellite object tracking that is capable of obtaining discriminative features and achieves competing performance. For instance, HRSiam (Shao et al., 2021), SiamMDM (Yang et al., 2023), and ThickSiam (Zhang et al., 2023) inherit the SNN, achieving the accuracy-speed trade-off in SV object tracking. During tracking, object rotation is common in SV. Several works have been developed to address this issue. On the one hand, some trackers exploit rotation invariance features to deal with the rotation issue. For example, WTIC (Wang et al., 2020) uses the Gabor filter to train a feature representation model that is adaptable for object rotation. CPKF and DF use the CH features to generate a response map that is inherently robust to object rotation. However, the output of the above trackers is a horizontal bounding box (HBB) with unsatisfactory semantic representations. On the other hand, the trackers strive to extract a series of rotation patches with an angle pool to achieve a better match with the template. RACF (Xuan et al., 2021) takes a similar strategy to address the rotation issue and proposes a method to estimate the scale change even with its output of the HBB. In RAMC (Chen et al., 2022c), object rotation is analyzed with a focus on illustrating the consistency of the spatial layout of the object and background between the current and initial frames. It also achieves adaptive angle estimation with its output of the oriented bounding box (OBB). However, the hand-crafted features of CF-based trackers may lead to performance constraints, especially for slight angle variations. The SNN is capable of obtaining discriminative features and is expected to handle the rotation problem of SV objects.

An object usually exhibits arbitrary positions and orientations from frame to frame. However, the Siamese network uses the region with a fixed orientation (zero degree) to locate the object. Considering that the Siamese network is not equivariant to object rotation, it may perform poorly on the orientation of less represented objects in the training set. Towards this end, the rotation equivariant for SOT (Gupta et al., 2021) and multiple object detection (Zhu et al., 2022) have been explored. However, these methods track or detect objects by rotating templates or filters, which makes it difficult to detect small angle changes. Moreover, the output object angles are fixed and discontinuous. In addition, ignoring the characteristics of the object itself also produces inaccurate representations such as position, orientation, and spatial distribution. Few methods utilize spatial distribution features of objects (e.g., segmentation masks) to facilitate semantic representations and improve performance.

To realize continuous tracking of the object with arbitrary orientations, we propose a rotation equivariant Siamese network enhanced by probability segmentation (REPS) for SV object tracking. First, to achieve rotation equivariant tracking, we propose a rotation equivariant (RE) Siamese network architecture for dealing with the inconsistency of semantic representations, which can detect rotation variations right from the start frame to acquire the precise object position and motion orientation. Considering that objects are homogeneous, we further propose a pixel-level (PL) refinement method that can capture specific object distribution by constructing the foreground probability map online. Meanwhile, an adaptive Gaussian fusion (AGF) is proposed to synergize the tracking and segmentation results to obtain compact output. The proposed REPS can obtain precision location and further semantic representation with OBBs and segmentation masks that are consistent with realistic states.

The remainder of this work is presented below. Section 2 depicts the proposed approach. Section 3 presents experiments and analysis. Extensive discussion is included in Section 4. Lastly, Section 5 formulates general conclusions, emphasizes the main contribution, and provides thoughts on future work.

2. Methodology

2.1. Overall architecture

Fig. 1 presents the architecture of REPS including RE tracking and PL refinement. The former leverages the proposed RE Siamese network to deal with the semantic inconsistency issue, obtaining the accurate position and orientation information of objects. Its input is a batch of patches cropped from the initial and current frames of the SV sequence. Let E and C denote the patch of the exemplar and the candidate, respectively. C can be regarded as a set of candidate patches with the same dimensionality but larger size than the candidate E. Its output is a series of response maps illustrating the similarity of the E and each patch of C. In the tracking stage, these candidates are first obtained by the previous object state such as the previous object position and orientation. The preliminary tracking result is determined by response maps. Furthermore, the result is refined by the latter (i.e., PL refinement) which can construct the foreground object probability map and achieve the AGF. The object probability distribution is tapped to overcome the tracking challenges (e.g., deformation and motion blur) and extract accurate information (e.g., segmentation mask). Specifically, the proposed AGF strategy can adaptively fuse tracking and segmentation results to enhance the tracking performance.

2.2. RE tracking

The proposed REPS constructs a rotation equivariant architecture for SV tracking which is modeled on the fully convolutional Siamese network (Bertinetto et al., 2016b). The main benefit of employing the fully convolutional structure is that a larger candidate region can be fed into the network without adjusting it to the size of the exemplar. The feature extraction function $f: Q \rightarrow P$ would be equivariant if

$$f(\varphi^Q(q)) = \varphi^P(f(q)), q \in Q, \tag{1}$$

where $f(\bullet)$ denotes the feature extraction operation, and φ represents the rotation transformation. As mentioned above, the Siamese network is not equivariant, which may cause unsatisfactory results, especially for rotating objects. To realize rotation equivariant tracking, a RE architecture for tracking satellite objects without additional data augmentation or redundant parameters is proposed, which can rotate candidate patches to capture small angle changes. Considering that object variations are relatively slight in adjacent frames, the angle pool $\{\alpha_i = [-2, 0, 2], i \in 1, 2, 3\}$ is explored to enumerate the angle variations and achieve a better match for tracking. The RE tracking consists of two streams. One is the template stream where the exemplar patch E of the initial frame is input. Another one is the detection stream, which enters a batch of candidate patches C_t^i with an angle pool at frame t. As shown in Fig. 1, the exemplar E, a region including the object $(x_1, y_1, w_1, h_1, \theta_1)$ and background, is first extracted from the initial frame, where (x_1, y_1) indicates the center of the object, (w_1, h_1) represents the object's width and height, and θ_1 is the initial angle. We then extract candidates C_t^i by angle θ_t^i that records the rotation angles of the object at frame t. θ_t^i is computed by

$$\theta_t^i = \theta_{t-1} + \alpha_i, \tag{2}$$

where θ_{t-1} denotes the rotation angle of the interest object at frame $t-1$. Both streams share the same convolutional neural network (CNN) $f(\bullet)$, so the same transformation is implicitly imposed on the exemplar and candidates. Then, the transformed inputs $f(E)$ and $f(C_t^i)$ are cross-correlated to generate response maps by

$$R(E, C_t^i) = f(E)*f(C_t^i) + b, \tag{3}$$

where $f(E)$ and $f(C_t^i)$ denote extracted features of E and C_t^i, * is the cross-correlation, b is a bias signal, and $R(E, C_t^i)$ is a batch of response maps defined on the angle pool. Finally, the object state $(x_t, y_t, w_t, h_t, \theta_t)$ of frame t is derived by

$$(x_t, y_t, w_t, h_t, \theta_t) = \underset{x,y,i}{argmax}(R(E, C_t^i)), (i = 1, 2, 3). \tag{4}$$

To achieve the RE tracking, the fully convolutional Siamese network is pre-trained over more than two million labels from the ImageNet dataset (Russakovsky et al., 2015). The parameter ϑ of the network is determined by taking the Stochastic Gradient Descent (SGD) to minimize the loss $\mathscr{L}$ by

$$\underset{\vartheta}{argmax}(\mathbb{E}_{(e,c,y)}(\mathscr{L}(g(e, c; \vartheta), y))), \tag{5}$$

where $\mathbb{E}$ is the expectation, g denotes a score map, $y[u] \in \{-1, 1\}$ represents the labels for each position $u \in \mathscr{D}$ of the response map, and (e, c) is a pair of training samples concerning exemplar and candidate patches.

2.3. PL refinement

The RE tracking guarantees accurate object positions and orientations, but it ignores the significance of spatial distribution that can boost the tracking and semantic representation. Therefore, a PL refinement is proposed. It first generates the segmentation mask by building the per-pixel foreground probability map online. Subsequently, an AGF that hybridizes the tracking and segmentation results is proposed to obtain the accurate object position, orientation, and spatial range. The PL refinement includes the generation of segmentation masks and the refinement with AGF.

2.3.1. Generation of segmentation masks

Considering the homogeneity of the object, we explore the spectral statistic feature that is inherently insensitive to deformation to obtain the per-pixel probability in the foreground region $\mathscr{O} \subset \mathbb{Z}^2$ and the background region $\mathscr{B} \subset \mathbb{Z}^2$. For constructing the per-pixel probability maps, the linear regression is introduced to get per-frame objective loss $\mathscr{L}(x_t, p_t, \delta)$ by

$$L(g_t, p_t, \delta) = \frac{1}{|\mathscr{O}|}\sum_{u \in \mathscr{O}}(\delta^T \phi_g[u] - 1)^2 + \frac{1}{|\mathscr{B}|}\sum_{u \in \mathscr{B}}(\delta^T \phi_g[u])^2, \tag{6}$$

where p_t is the desired object position of an image g_t at frame t, the K

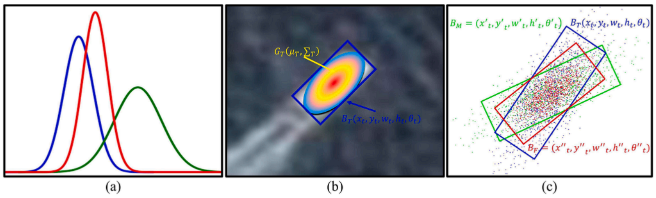

$G_T(\mu_T, \sum_T)$
$B_T(x_t, y_t, w_t, h_t, \theta_t)$
$B_M = (x'_t, y'_t, w'_t, h'_t, \theta'_t)$
$B_T(x_t, y_t, w_t, h_t, \theta_t)$
$B_F = (x''_t, y''_t, w''_t, h''_t, \theta''_t)$
(a)
(b)
(c)

Fig. 2. (a) Visualization of the 1-D Gaussian product process, where the blue and green lines indicate the observed and predicted states, respectively. The red line indicates the result of the Gaussian product. The result has a mean value between the observed and predicted states and a smaller variance than both. (b) shows the object $B_T(x_t, y_t, w_t, h_t, \theta_t)$ and its 2-D Gaussian distribution $G_T(\mu_T, \sum_T)$. (c) illustrates the optimization process of AGF, where the red, green, and blue dots follow the Gaussian distribution of corresponding bounding boxes, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Details of satellite video datasets.

Datasets	FS	OS	FN	OC	DS
Minneapolis	4090 × 2160	16.3 × 18.9	268	Car	Jilin-1
Sydney	4096 × 3072	11.0 × 22.0	320	Ship	Jilin-1
Atlanta	12000 × 5000	52.5 × 44.6	296	Plane	Jilin-1
Vancouver	3840 × 2160	22.9 × 78.3	405	Train	ISS
Dubai	4096 × 3072	12.6 × 10.7	147	Car	Jilin-1

FS = Frame Size. OS = Object Size. FN = Frame Number. OC = Object Category. DS = Data Source. ISS = International Space Station.

channel feature $\phi_g : \mathscr{P} \rightarrow \mathbb{R}^K$ is extracted from g_t and defined on grid $\mathscr{P} \subset \mathbb{Z}^2$, and δ denotes the model parameter with K feature channels. Based on the one-point assumption and ridge regression, it has

$$\delta^j = \frac{\rho^j(\mathscr{O})}{\rho^j(\mathscr{O}) + \rho^j(\mathscr{B}) + \lambda}, j = 1, \cdots, K, \tag{7}$$

where $\rho^j(\bullet) = \mathscr{Z}^j(\bullet)/|\bullet|$ is the pixel proportion of a region in feature channel j, and $\mathscr{Z}^j(\bullet)$ denotes the number of pixels. The object probability map is then constructed, and $\rho(\mathscr{O})$ and $\rho(\mathscr{B})$ are also updated to adapt to object changes.

To further generate the binary mask, we propose a PL segmentation method based on self-similarity. In the initialization stage, the foreground probability map is treated as the training data $s_i = \{s_1, s_2, \cdots, s_N\}$ with a total of N pixels. We then compute the per-pixel class (i.e., foreground or background) $c_i = \{c_1, c_2, \cdots, c_N\}$ by thresholding, in which the paired training set would be $\{(s_i, c_i), i = 1, 2, \cdots, N\}$. Accordingly, the paired test set would be (s'_i, c'_i), where s'_i is a pixel value of the foreground probability map, and c'_i is the desired pixel class. To compute c'_i, we build the distance metric $d_i = \left\{(s'_i - s)^2, i = 1, 2, \cdots, N\right\}$ to measure similarity and identify the class of pixel i. When ranking d_i in descending order, the predicted class c'_i of per-pixel probability s'_i is set as the majority of k classes. Finally, we could obtain all c'_i according to s'_i and generate the segmentation mask to compact the real-world object state.

2.3.2. Refinement with AGF

How to exploit the synergized advantages of tracking and segmentation remains a challenge. The RE tracking architecture estimates the OBB that illustrates the consistency of spatial layout between current and initial frames. However, it ignores the characteristics of the object itself to a certain extent. As shown in Fig. 1, the OBB $B_T = (x_t, y_t, w_t, h_t, \theta_t)$ fails to shape the object precisely (e.g., orientation and size). In contrast, the minimum bounding box $B_M = (x'_t, y'_t, w'_t, h'_t, \theta'_t)$ of the segmentation mask is more compact and would complement the RE tracking result. Thus, a novel AGF that synergizes the tracking and segmentation results is proposed for SV object tracking. As shown in Fig. 2(a), to estimate the true state from observation terms with errors, two 1-D Gaussian functions are multiplied to obtain a new 1-D Gaussian distribution function inspired by the Kalman filter. Considering that SV objects usually present an elliptical-like distribution (Chen et al., 2022c), it can be approximated as a 2-D Gaussian distribution, as shown in Fig. 2(b). $(\mu_T, \sum_T)$ denotes the mean and the variance of Gaussian function G_T. Motivated by the principles of Gaussian modeling and overlap calculation (Yang et al., 2022), the AGF first approximates B_T and B_M as the Gaussian distributions, and the product of distributions is

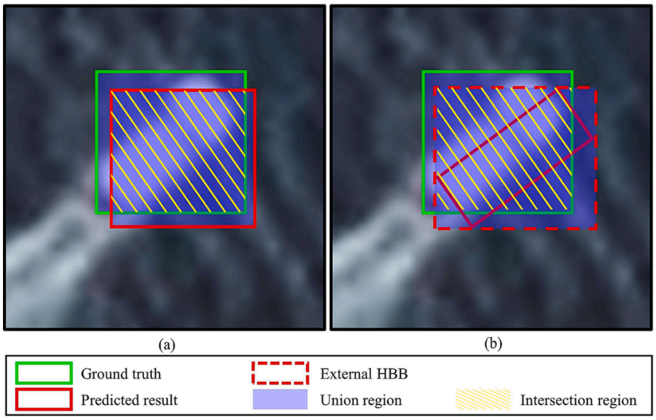


Fig. 4. Visualization of the overlap. (a) and (b) show the predicted results in HBB and OBB cases, respectively.

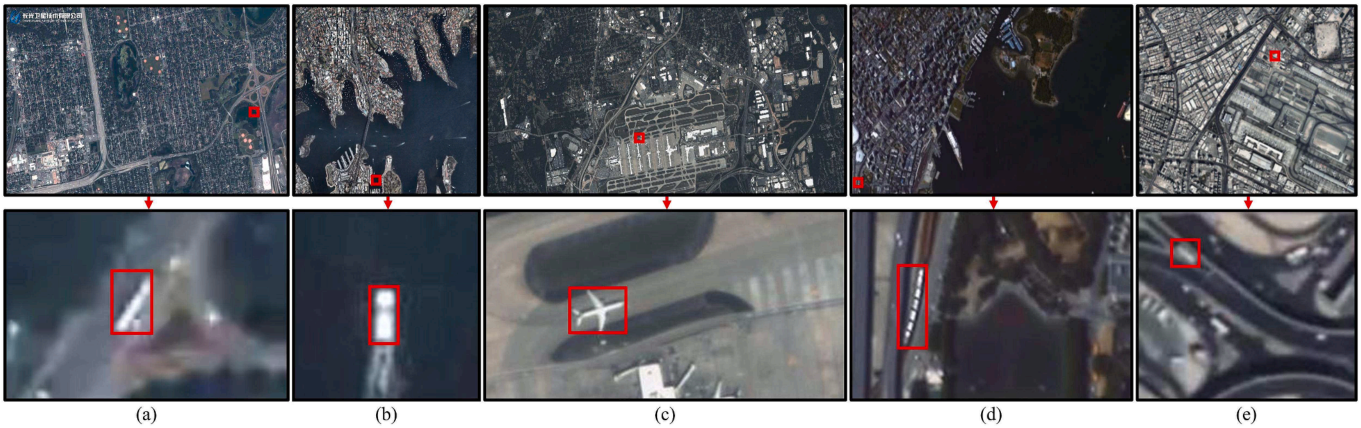
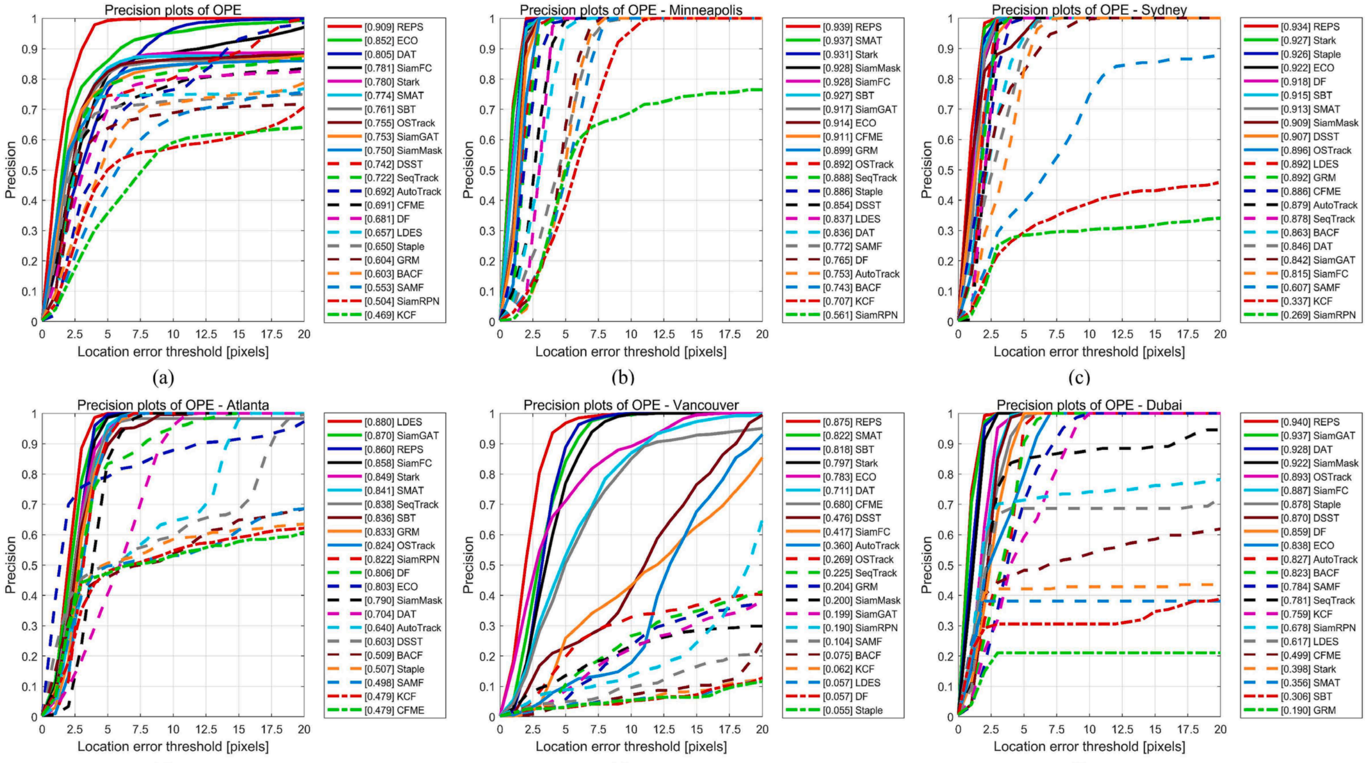


Fig. 3. Satellite video datasets. (a) Minneapolis, Car. (b) Sydney, Ship. (c) Atlanta, Plane. (d) Vancouver, Train. (e) Dubai, Car.

Table 2
Experimental results of overall datasets and characteristics of trackers.

Trackers	Venue	Feature/Backbone	MS	Output	Pre	Suc	FPS	
							CPU	GPU
SAMF	ECCV 2015	HOG + CN + I	✔	HBB	0.553	0.406	76.0	-
DAT	CVPR 2015	CH	✔	HBB	**0.805**	**0.679**	**283.7**	-
KCF	TPAMI 2015	HOG	-	HBB	0.469	0.350	**463.3**	-
Staple	CVPR 2016	HOG + CN	✔	HBB	0.650	0.517	127.6	-
SiamFC	ECCV 2016	AlexNet	✔	HBB	0.781	0.647	-	**125.9**
DSST	TPAMI 2017	HOG + I	✔	HBB	0.742	0.631	**206.2**	-
BACF	ICCV 2017	HOG	✔	HBB	0.603	0.500	67.8	-
ECO	CVPR 2017	VGG-M	✔	HBB	**0.852**	0.649	2.5	-
SiamRPN	CVPR 2018	AlexNet	✔	HBB	0.504	0.457	-	**372.6**
LDES	AAAI 2019	HOG + CH	✔	OBB	0.657	0.468	18.0	-
SiamMask	CVPR 2019	ResNet-50	✔	OBB, Mask	0.750	0.488	-	**123.0**
AutoTrack	CVPR 2020	HOG + CN + I	✔	HBB	0.692	0.561	83.6	-
CFME	TGRS 2020	HOG	-	HBB	0.691	0.587	11.0	-
SiamGAT	CVPR 2021	GoogLeNet	✔	HBB	0.753	0.626	-	56.2
Stark	ICCV 2021	ResNet-101	✔	HBB	0.780	**0.703**	-	70.7
OSTrack	ECCV 2022	ViT-Base	✔	HBB	0.755	0.657	-	92.7
DF	JSTARS 2022	HOG + CN + GCS	-	HBB	0.681	0.525	89.5	-
SBT	CVPR 2022	SBT-Base	✔	HBB	0.761	0.678	-	56.1
GRM	CVPR 2023	ViT-Base	✔	HBB	0.604	0.550	-	66.1
SeqTrack	CVPR2023	ViT-Large	✔	HBB	0.722	0.622	-	10.8
SMAT	WACV 2024	MobileViTv2	✔	HBB	0.774	0.667	-	121.7
REPS	Ours	AlexNet	✔	OBB, Mask	**0.909**	**0.791**	-	11.2

The top three scores are bolded. MS denotes the mechanism for scale. For the Feature/Backbone, HOG = histogram of oriented gradients, CN = color name, I = intensity, CH = color histogram. AlexNet (Krizhevsky et al., 2017), VGG-M (Chatfield et al., 2014), ResNet-50 (He et al., 2016), GoogLeNet (Szegedy et al., 2015), ResNet-101 (He et al., 2016), ViT-Base (Dosovitskiy et al., 2020), SBT-Base (Xie et al., 2022), and MobileViTv2 (Mehta and Rastegari, 2022) denote the backbone networks.



(a) (b) (c)

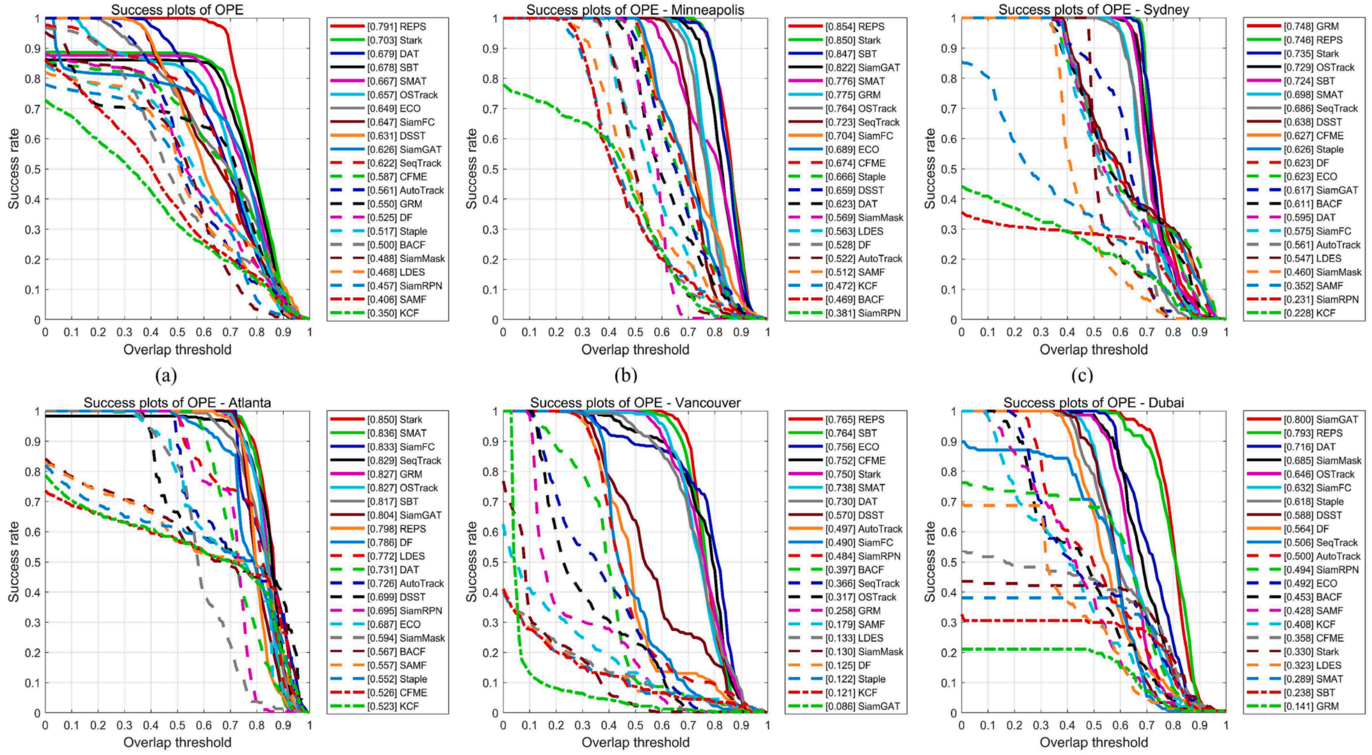
(d) (e) (f)

Fig. 5. The precision plot for overall and per-dataset. (a) overall. (b) Minneapolis. (c) Sydney. (d) Atlanta. (e) Vancouver. (f) Dubai. The values in the legend indicate Pre.

then calculated to obtain the fusion result $B_F = (x''_t, y''_t, w''_t, h''_t, \theta''_t)$, as shown in Fig. 2(c). Specifically, B_T and B_M are converted to Gaussian distributions $G_T(\mu_T, \sum_T)$ and $G_M(\mu_M, \sum_M)$ respectively via

$$\mu = (x, y)^T, \sum = R\Lambda\mathscr{R}^T, \tag{8}$$

where $\mathscr{R} = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix}$ is a rotation matrix, i.e., eigenvector



(a) (b) (c)

(d) (e) (f)

Fig. 6. The success plot for overall and per-dataset. (a) overall. (b) Minneapolis. (c) Sydney. (d) Atlanta. (e) Vancouver. (f) Dubai. The values in the legend indicate Suc.

Table 3
Experimental results of per-dataset.

Trackers	Venue	Feature/Backbone	Minneapolis		Sydney		Atlanta		Vancouver		Dubai	
			Pre	Suc	Pre	Suc	Pre	Suc	Pre	Suc	Pre	Suc
SAMF	ECCV 2015	HOG + CN + I	0.772	0.512	0.607	0.352	0.498	0.557	0.104	0.179	0.784	0.428
DAT	CVPR 2015	CH	0.836	0.623	0.846	0.595	0.704	0.731	0.711	0.730	**0.928**	**0.716**
KCF	TPAMI 2015	HOG	0.707	0.472	0.337	0.228	0.479	0.523	0.062	0.121	0.759	0.408
Staple	CVPR 2016	HOG + CN	0.886	0.666	**0.926**	0.626	0.507	0.552	0.055	0.122	0.878	0.618
SiamFC	ECCV 2016	AlexNet	0.928	0.704	0.815	0.575	0.858	**0.833**	0.417	0.490	0.887	0.632
DSST	TPAMI 2017	HOG + I	0.854	0.659	0.907	0.638	0.603	0.699	0.476	0.570	0.870	0.588
BACF	ICCV 2017	HOG	0.743	0.469	0.863	0.611	0.509	0.567	0.075	0.397	0.823	0.453
ECO	CVPR 2017	VGG-M	0.914	0.689	0.922	0.623	0.803	0.687	0.783	**0.756**	0.838	0.492
SiamRPN	CVPR 2018	AlexNet	0.561	0.381	0.269	0.231	0.822	0.695	0.190	0.484	0.678	0.494
LDES	AAAI 2019	HOG + CH	0.837	0.563	0.892	0.547	**0.880**	0.772	0.057	0.133	0.617	0.323
SiamMask	CVPR 2019	ResNet-50	0.928	0.569	0.909	0.460	0.790	0.594	0.200	0.130	0.922	0.685
AutoTrack	CVPR 2020	HOG + CN + I	0.753	0.522	0.879	0.561	0.640	0.726	0.360	0.497	0.827	0.500
CFME	TGRS 2020	HOG	0.911	0.674	0.886	0.627	0.479	0.526	0.680	0.752	0.499	0.358
SiamGAT	CVPR 2021	GoogLeNet	0.917	0.822	0.842	0.617	**0.870**	0.804	0.199	0.086	**0.937**	**0.800**
Stark	ICCV 2021	ResNet-101	**0.931**	**0.850**	**0.927**	**0.735**	0.849	**0.850**	0.797	0.750	0.398	0.330
OSTrack	ECCV 2022	ViT-Base	0.892	0.764	0.896	0.729	0.824	0.827	0.269	0.317	0.893	0.646
DF	JSTARS 2022	HOG + CN + GCS	0.765	0.528	0.918	0.623	0.806	0.786	0.057	0.125	0.859	0.564
SBT	CVPR 2022	SBT-Base	0.927	**0.847**	0.915	0.724	0.836	0.817	**0.818**	**0.764**	0.306	0.238
GRM	CVPR 2023	ViT-Base	0.899	0.775	0.892	**0.748**	0.833	0.827	0.204	0.258	0.190	0.141
SeqTrack	CVPR2023	ViT-Large	0.888	0.723	0.878	0.686	0.838	0.829	0.225	0.366	0.781	0.506
SMAT	WACV 2024	MobileViTv2	**0.937**	0.776	0.913	0.698	0.841	**0.836**	**0.822**	0.738	0.356	0.289
REPS	Ours	AlexNet	**0.939**	**0.854**	**0.934**	**0.746**	**0.860**	0.798	**0.875**	**0.765**	**0.940**	**0.793**

matrix of covariance $\sum$, and $\Lambda = \begin{bmatrix} w^2/4 & 0 \\ 0 & h^2/4 \end{bmatrix}$ denotes the eigenvalue matrix derived from scaling matrix $\mathscr{S} = \begin{bmatrix} w/2 & 0 \\ 0 & h/2 \end{bmatrix}$ via $\Lambda = \mathscr{S}\mathscr{S}^T$. With G_T and G_M, we can compute fusion distribution $\tau G_F\left(\mu_F, \sum_F\right)$ by

$$\tau G_F\left(\mu_F, \sum\nolimits_F\right) = G_M\left(\mu_M, \sum\nolimits_M\right) G_T\left(\mu_T, \sum\nolimits_T\right), \tag{9}$$

where $\sum_F = (1-K)\sum_M$. $\mu_F = \mu_M + K(\mu_T - \mu_M)$. K denotes the gain matrix with $K = \sum_M\left(\sum_M + \sum_T\right)^{-1}$. τ is a scaling term of G_F with

$$\tau = \left|2\pi\left(\left(\sum\nolimits_M + \sum\nolimits_T\right)\right)\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mu_M - \mu_T)^T\left(\sum_M + \sum_T\right)^{-1}(\mu_M - \mu_T)}, \tag{10}$$

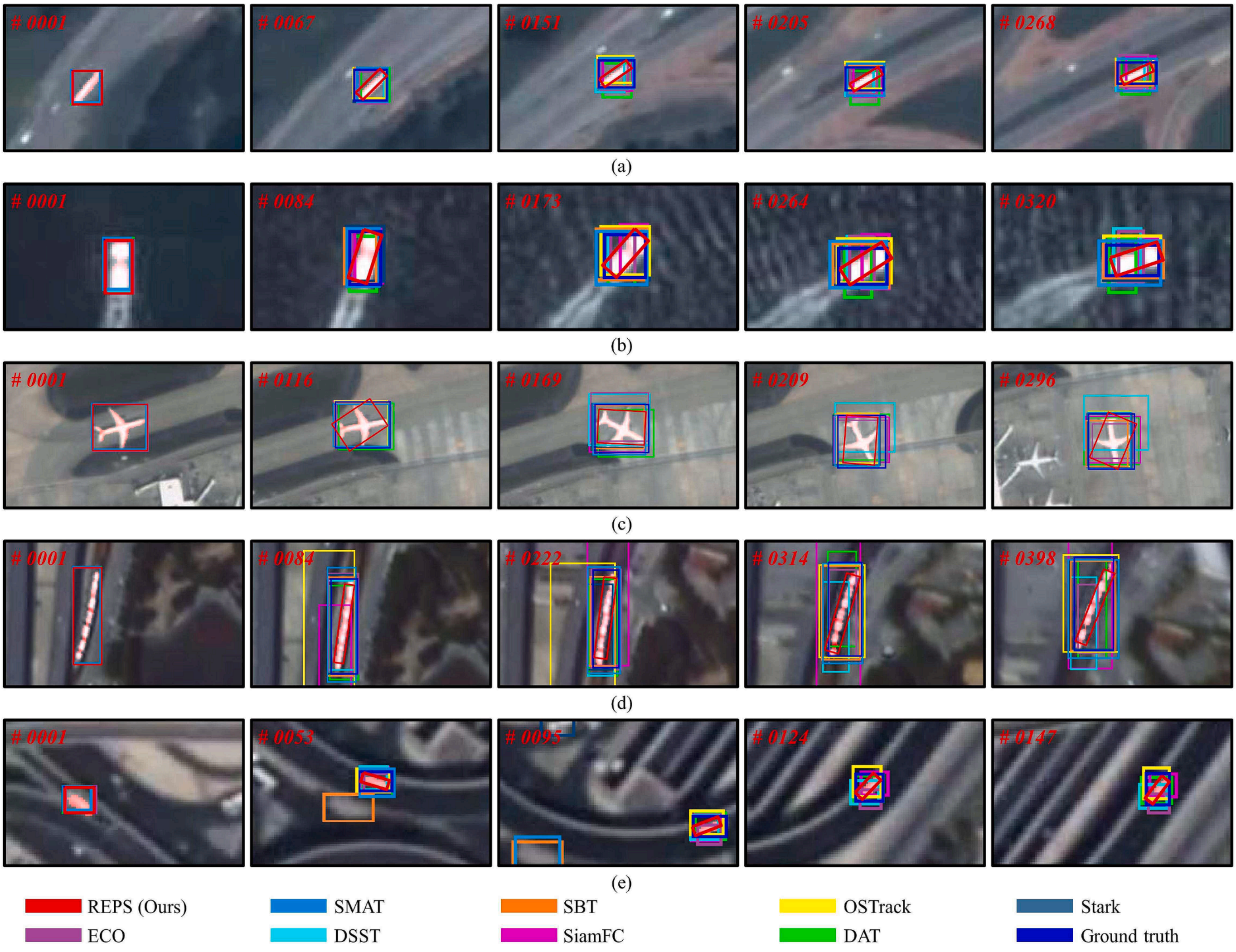


Fig. 7. Qualitative results of the top nine trackers. The current frame is displayed in the upper-left corner of each image. REPS can generate OBBs and segmentation masks marked by red pixels. (a) Minneapolis. (b) Sydney. (c) Atlanta. (d) Vancouver. (e) Dubai. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Components and results of different models.

Models	RE tracking	PL refinement	Pre	Suc
Model-1	–	–	0.768	0.640
Model-2	✔	–	0.871	0.638
Model-3	–	✔	0.871	0.768
REPS	✔	✔	0.909	0.791

where $|\bullet|$ denotes $det(\bullet)$. Considering the small-sized object with a close centroid between B_T and B_M, the position (x''_t, y''_t) of the fusion bounding box B_F is $\mu_F = \mu_M + K(\mu_T - \mu_M)$. The angle θ''_t is determined by finding the eigenvector matrix $\mathscr{R}$ of $\sum_F$, and size (w''_t, h''_t) can be obtained by eigenvalue matrix $\Lambda = \mathscr{S}\mathscr{S}^T$. Finally, we can obtain the desired B_F that synergizes the tracking and segmentation to output compact results for SV object representations.

3. Experiments and analysis

3.1. Experimental settings and datasets

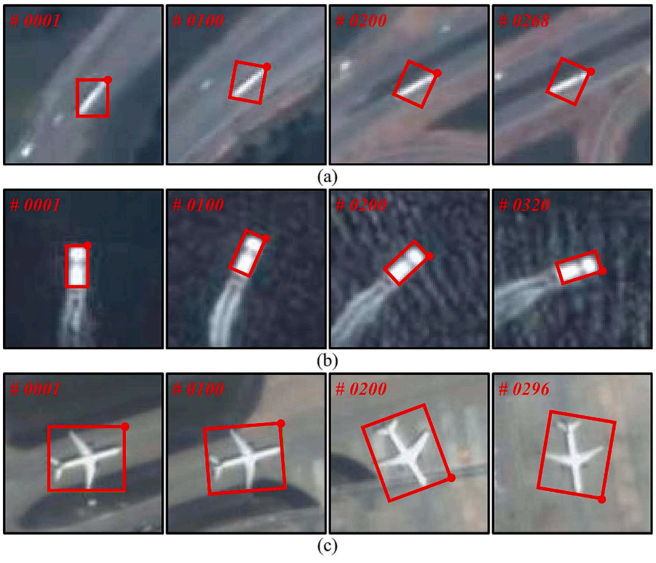
In the RE tracking, the initial angle θ_0 is set to zero, and the response maps are upsampled by three times to improve accuracy. The learning rate of the SGD optimizer anneals geometrically from 10^{-2} to 10^{-5}. The training process is over 50 epochs, and the size of mini-batches is 8. To adapt to slightly larger angle variations and balance accuracy and efficiency, the angle pool is set to [−2,0,2]. In the PL refinement, the number of bins for the color histograms is 2^3. k is set to 4 owing to a significant probability difference between objects and backgrounds. The proposed method is tested on a computer with a 3.2 GHz Intel(R) Xeon (R) Gold6134 CPU and NVIDIA GeForce RTX 2080 Ti GPU. We perform comprehensive experiments over space-borne video datasets. Table 1 provides detailed information on some of the datasets while Fig. 3 shows the first frames and tracked objects.

3.2. Evaluation metrics

Precision and success plots are utilized to benchmark the trackers (Wu et al., 2015). In the precision plot, the precision rate illustrates the proportion of frames where the center location error v is smaller than thresholds varied from 1 to 20 pixels. v is obtained by

$$v = \sqrt{(x - X)^2 + (y - Y)^2}, \tag{11}$$

where (x, y) and (X, Y) are the center of the predicted result r_p and the



(a)

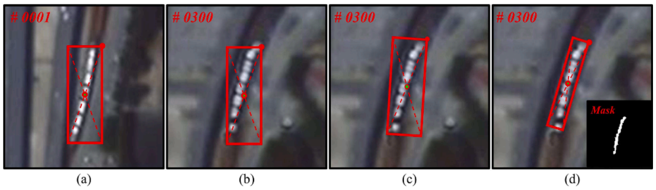
(b)

(c)

Fig. 8. Tracking examples of Model-2. The red dot denotes the consistency constraint of the OBB. (a) Minneapolis, Car. (b) Sydney, Ship. (c) Atlanta, Plane. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Components and results of different angle pools.

Models	Angle pool	Pre	Suc
Model-4	[−2,−1,0,1,2]	0.910	0.793
Model-5	[−5,−3,0,3,5]	0.899	0.760
Model-6	[−1,0,1]	0.911	0.791
Model-7	[−2,0,2]	0.909	0.791
Model-8	[−3,0,3]	0.911	0.788



(a)

(b)

(c)

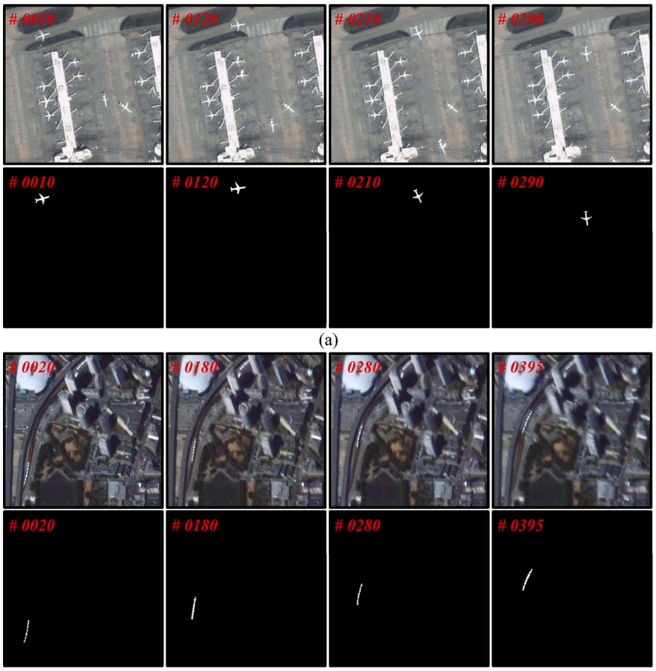
(d)

Fig. 9. Visualization examples of different stages. (a) shows the tracking initialization step with HBB. (b) is the HBB result of Model-1. (c) shows the OBB result of Model-2. It can be noticed that (b) and (c) are not centered on the train and estimate inaccurate real-world states. (d) presents the result of the REPS that provides the accurate OBB and segmentation result.

ground truth r_g, respectively. In the success plot, the success rate aims to calculate the percentage of successful frames where the overlap s surpasses the thresholds varied from 0 to 1. Given r_p and r_g, s is obtained by

$$s = \left| r_p \cap r_g \right| \Big/ \left| r_p \cup r_g \right|, \tag{12}$$

where $\cup$ is the union, $\cap$ is the intersection, and $|\bullet|$ denotes the pixel count in a given region. As mentioned above, the HBB has difficulty in describing the object orientation and spatial distribution when compared with the OBB. However, many trackers (Chen. et al., 2022; Xuan et al., 2020) can only receive the HBB for initialization and produce the HBB output. When r_p is the OBB format, it is converted to the external HBB followed by a calculation of s. Fig. 4 illustrates the calculation of s when the predicted results are HBB and OBB, respectively. All trackers are ranked by the area under the curve of the precision plot (Pre) and the success plot (Suc) (Shao et al., 2021). The median FPS is used to measure the running efficiency.



(a)

(b)

Fig. 10. Segmentation masks of the REPS. (a) Atlanta, Plane. (b) Vancouver, Train.

Table 6
Fine-grained attributes of the OOTB dataset.

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 an SV.
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.

3.3. Comparison with state-of-the-art trackers

In this section, we compare REPS with 21 state-of-the-art (SOTA) trackers in quantitative and qualitative terms. Compared trackers include SAMF (Li and Zhu, 2015), DAT (Possegger et al., 2015), KCF (Henriques et al., 2015), Staple (Bertinetto et al., 2016a), SiamFC (Bertinetto et al., 2016b), DSST (Danelljan et al., 2017b), BACF (Galoogahi et al., 2017), ECO (Danelljan et al., 2017a), SiamRPN (Li et al., 2018a), LDES (Li et al., 2019), SiamMask (Wang et al., 2019), AutoTrack (Li et al., 2020), CFME (Xuan et al., 2020), SiamGAT (Guo et al., 2021), Stark (Yan et al., 2021a), OSTrack (Ye et al., 2022), DF (Chen. et al., 2022) (https://github.com/YZCU/DF), SBT (Xie et al.,

Table 7
Detailed characteristics, overall results, and category-based results of trackers. All trackers are tested on the OOTB dataset.

Trackers	Venue	Feature/Backbone	Overall result		Category-based result							
					Car		Ship		Train		Plane	
			Pre	Suc	Pre	Suc	Pre	Suc	Pre	Suc	Pre	Suc
CSK	ECCV 2012	I	0.560	**0.482**	0.515	0.361	0.564	0.464	**0.294**	**0.413**	0.741	**0.752**
SAMF	ECCV 2015	HOG + CN + I	0.573	0.463	**0.602**	0.409	0.497	0.394	0.211	0.306	0.758	**0.707**
DAT	CVPR 2015	CH	0.535	0.457	0.507	0.380	0.677	0.509	0.182	0.292	0.556	0.601
KCF	TPAMI 2015	HOG	0.512	0.432	0.505	0.345	0.506	0.392	0.211	0.306	0.653	0.686
SiamRPN	CVPR 2018	AlexNet	0.586	0.468	0.579	**0.447**	0.604	0.464	0.156	0.335	0.751	0.563
SiamGAT	CVPR 2021	GoogLeNet	**0.677**	0.476	**0.723**	**0.498**	0.714	**0.533**	0.044	0.087	0.803	0.523
LightTrack	CVPR 2021	Custom	**0.595**	**0.489**	0.577	0.423	0.622	0.499	0.065	0.166	**0.808**	**0.724**
Stark	ICCV 2021	ResNet-101	**0.595**	**0.489**	0.490	0.364	0.746	**0.583**	0.178	0.298	0.768	0.678
OSTrack	ECCV 2022	ViT-Base	0.570	0.438	0.378	0.252	**0.770**	**0.534**	0.225	0.324	**0.813**	0.703
SimTrack	ECCV 2022	ViT-Base	0.520	0.430	0.350	0.262	0.615	0.483	**0.359**	**0.478**	0.777	0.652
SBT	CVPR 2022	SBT-Base	0.564	0.421	0.437	0.293	0.751	0.515	0.086	0.160	0.758	0.642
GRM	CVPR 2023	ViT-Base	0.565	0.427	0.380	0.258	**0.760**	0.513	0.177	0.263	**0.819**	0.694
SeqTrack	CVPR 2023	ViT-Large	0.577	0.433	0.403	0.270	**0.763**	0.514	**0.236**	**0.399**	0.803	0.643
REPS	Ours	AlexNet	**0.690**	**0.486**	**0.715**	**0.451**	0.739	0.512	0.170	0.274	0.797	0.605

ViT-Large (Dosovitskiy et al., 2020) denotes the backbone network.

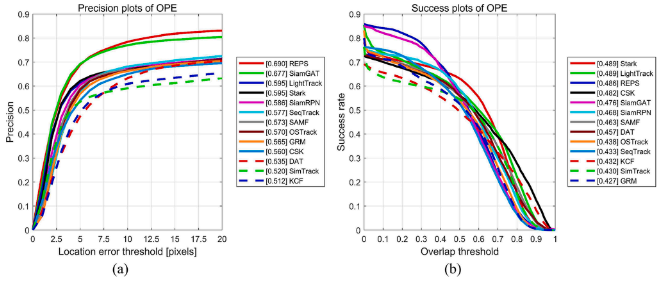


Fig. 11. Precision plot (a) and success plot (b) for the overall dataset.

2022), GRM (Gao et al., 2023), SeqTrack (Chen et al., 2023a), and SMAT (Yelluru Gopal and Amer, 2024), covering a variety of features, backbones, tracking paradigms, and applications.

3.3.1. Overall evaluations

Table 2 details the results of overall datasets and characteristics of trackers. Figs. 5(a) and 6(a) display the precision plot and success plot of all trackers on overall datasets, respectively. With Pre and Suc of 0.909 and 0.791, REPS obtains remarkable performance while KCF achieves the worst accuracy. Stark produces competitive results with Pre and Suc of 0.780 and 0.703, respectively. This is because Stark can use the transformer framework to capture the long-range dependency in spatial and temporal dimensions. Compared to Stark, the Pre and Suc of REPS have increased by 12.9% and 8.8%, respectively. SeqTrack designs a novel sequence-to-sequence tracking framework that inherits the encoder-decoder transformer architecture and casts tracking tasks as a sequence generation issue. SeqTrack obtains 0.722 and 0.622 for Pre and Suc, which are 18.7% and 16.9% lower than that of REPS. Compared to CFME which cooperates motion model to mitigate model drift for SV object tracking, REPS achieves gains of 21.8% and 20.4% in Pre and Suc, respectively. When compared to SNN-based trackers such as SiamFC, SiamRPN, OSTrack, and SMAT, REPS offers superior accuracy and an acceptable speed of 11.2 FPS. Experimental results show the competitiveness of REPS, demonstrating the importance of synergizing RE tracking and PL refinement.

3.3.2. Per-dataset evaluations

Figs. 5 and 6 exhibit the precision plot and success plot for overall and per-dataset, respectively. Table 3 presents the results of per-dataset. For the Pre metric, REPS ranks first in four (i.e., Minneapolis, Sydney, Vancouver, and Dubai) out of five datasets and achieves the third result in the Atlanta dataset. For the Suc metric, REPS ranks in the top two in four (i.e., Minneapolis, Sydney, Vancouver, and Dubai) out of five datasets. One of the most challenging datasets is the Vancouver, in which most trackers drift away from the object. REPS keeps track of the object and gains Pre of 0. 875 and Suc of 0.765, ranking first among compared trackers. In Dubai, the small-sized object undergoes fast rotation and background clutters. Benefiting from the rotation equivariant structure, REPS obtains excellent performance with Pre and Suc of 0.940 and 0.793, respectively.

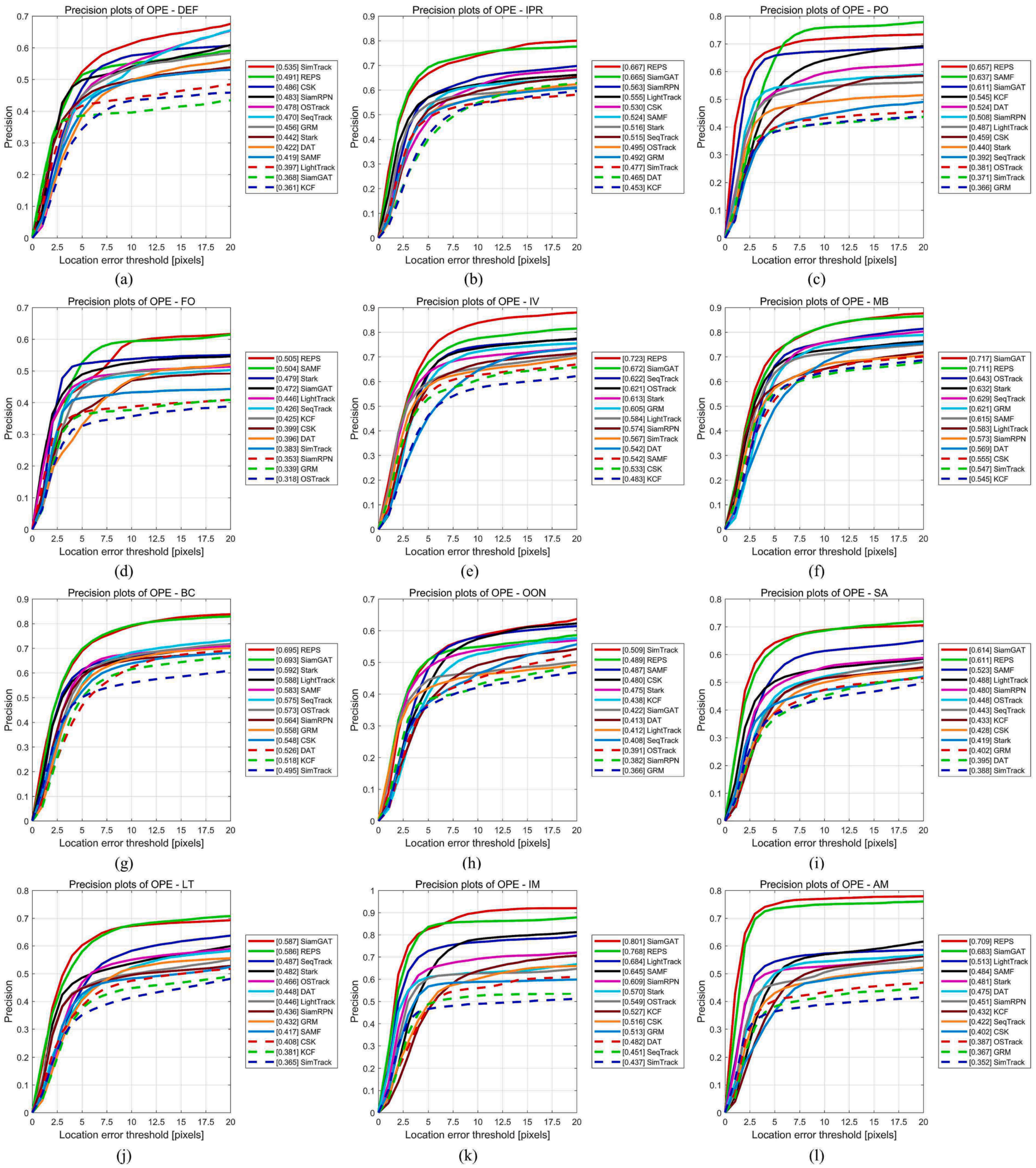
3.3.3. Qualitative evaluations

The qualitative results are shown in Fig. 7. In Vancouver, a train is experiencing rotation, motion blur, and illumination variation. In particular, train tends to have non-rigid deformation while car, ship, and plane usually experience rigid deformation. Therefore, it is difficult to track the train object. In this case, REPS can successfully track the object, whereas a lot of trackers either experience tracking drift or predict inaccurate position and scale. In other cases, REPS also obtains impressive performance. Moreover, it is capable of generating OBBs and segmentation masks for arbitrary objects, which better match real-world object states.

4. Discussions

4.1. Discussion on the rotation equivariant tracking

The RE tracking can deal with the inconsistency of semantic representations and detect rotation variations right from the start frame. To study its effectiveness, we conduct four sets of experiments including Model-1, Model-2, Model-3, and REPS. Detailed components and results are shown in Table 4. By comparing Model-3 and REPS, it can be seen that Pre and Suc reduce by 3.8% and 2.3%, respectively, after dropping the RE tracking component from REPS. Due to the absence of RE tracking, Model-3 cannot adaptively deal with the inconsistency of representations, which makes it difficult to achieve rotation equivariant tracking. To further validate the RE tracking, we compare Model-1 and Model-2. It can be seen that both of them yield comparable Suc, but the Pre is enhanced by 10.3% after introducing the RE tracking. Due to the nadir view, the rotation of objects is common in satellite object tracking (Chen et al., 2022c; Xuan et al., 2021). However, the general CNN has few abilities to achieve rotation equivarience. The RE tracking can achieve rotational equivariance and detect the rotation variation right from the start frame. Therefore, Model-2 can obtain precise positions and orientations. Fig. 8 shows the tracking examples of Model-2. It can be observed that Model-2 sensitively detects the small angle variations



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Fig. 12. The precision plot of per-attribute. (a) DEF. (b) IPR. (c) PO. (d) FO. (e) IV. (f) MB. (g) BC. (h) OON. (i) SA. (j) LT. (k) IM. (l) AM.

of SV objects, which demonstrates the effectiveness of the RE tracking.

4.2. Discussion on the angle pool

Here, we discuss the effectiveness of the number and interval of the angle pool. It is a key parameter for balancing accuracy and speed and facilitating RE tracking. Considering the steady motion of the object, it is difficult to make sudden angle changes. To this end, we conduct five sets of experiments with different angle pools, as shown in Table 5. It is observed that a large angle range tends to weaken the tracking performance, as seen in Model-5. This is because the object usually could not undergo drastic angle changes between adjacent frames. When the angle pool is within a certain range, the tracking performance is similar and comparable such as Model-4, Model-6, and Model-7. In addition, an

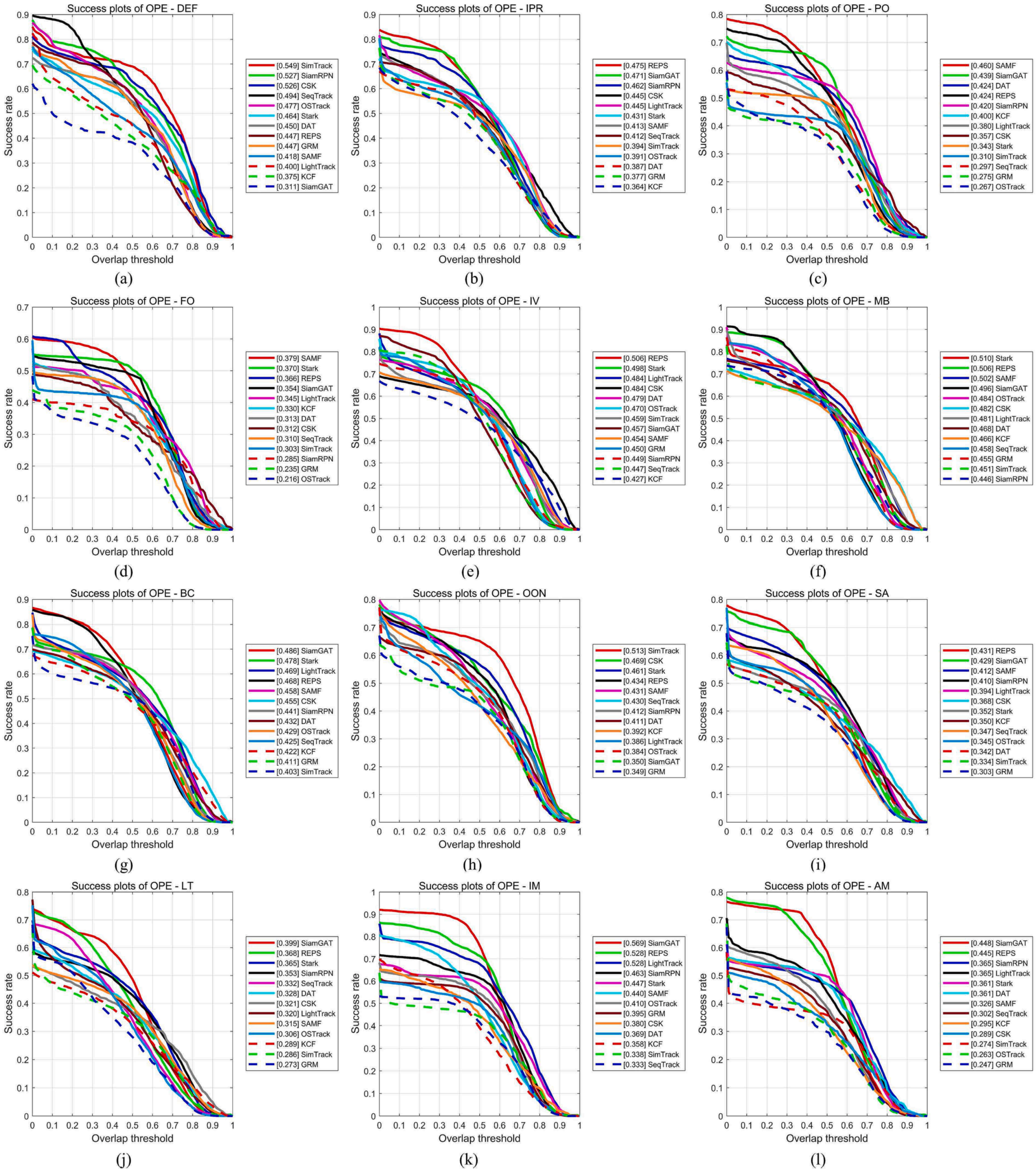


Fig. 13. The success plot of per-attribute. (a) DEF. (b) IPR. (c) PO. (d) FO. (e) IV. (f) MB. (g) BC. (h) OON. (i) SA. (j) LT. (k) IM. (l) AM.

excessive angle pool would cause high complexity and affect efficiency. Therefore, the angle pool [−2,0,2] is selected to achieve the accuracy-speed trade-off and prepare to handle slightly larger angle changes.

4.3. Discussion on the pixel-level refinement

Here, we discuss the effectiveness of the PL refinement component. Detailed components and results can be seen in Table 4. By comparing Model-3 and Model-1, it is observed that Pre decreases from 0.871 to 0.768 while Suc decreases from 0.768 to 0.640 after removing the PL refinement. This is because Model-1 cannot apply object masks to predict accurate states such as the center position and orientation, generating inferior results, as shown in Fig. 9(b). Furthermore, compared to Model-2, REPS improves by 3.8% in Pre and 15.3% in Suc. Model-2 can achieve rotation equivariant tracking and detect rotation variations right from the start frame. However, it only obtains the coarse-scale object states, which are inaccurate due to the initialization error caused by HBB, as shown in Fig. 9(c) and (a), respectively. In addition, the center may deviate from the object or not be on the objects that are prone to non-rigid deformation. Compared to Model-2, REPS further cooperates with the PL refinement to generate segmentation masks by constructing foreground probability maps online. Meanwhile, it also achieves the AGF to exploit the tracking and segmentation results and obtain more accurate position, orientation, and shape, as shown in Fig. 9 (d).

4.4. Discussion on k of the segmentation method

We explore the parameter k of the PL refinement component. Three sets of experiments including $k = 3$, $k = 4$, and $k = 5$ are designed. We find they produce the same results, which proves the proposed method is insensitive to k. The reason is that the SV object is homogeneous. There are usually significant differences between foregrounds and backgrounds. Considering the diversity and complexity of SV objects, it is challenging to track all types of objects, especially slender and tangled objects. Nevertheless, the proposed method is capable of capturing the compact spatial distribution, as shown in Fig. 10.

4.5. Discussion on the robustness and applicability

To validate the robustness and applicability of REPS, extensive experiments are further executed on a large-scale satellite video object tracking dataset, namely the oriented object tracking benchmark (OOTB) dataset (https://github.com/YZCU/OOTB), which is sampled from multiple satellite platforms such as Jilin-1, SkySat-1, and International Space Station (ISS). The multi-platform data would satisfy the need for dataset diversity and allow for better representation and generalization. The OOTB dataset consists of 110 sequences with a total of 29,890 frames and covers common object categories including 45 cars, 30 ships, 25 planes, and 10 trains. All sequences are manually annotated with high-quality bounding boxes and labeled with 12 fine-grained attributes, as shown in Table 6. On the OOTB dataset, we compare REPS with 13 representative trackers in terms of overall, per-category, and per-attribute. Compared trackers include CSK (Henriques et al., 2012), SAMF (Li and Zhu, 2015), DAT (Possegger et al., 2015), KCF (Henriques et al., 2015), SiamRPN (Li et al., 2018a), SiamGAT (Guo et al., 2021), LightTrack (Yan et al., 2021b), Stark (Yan et al., 2021a), OSTrack (Ye et al., 2022), SimTrack (Chen et al., 2022a), SBT (Xie et al., 2022), GRM (Gao et al., 2023), and SeqTrack (Chen et al., 2023a) and cover various features, backbones, paradigms, and applications. Table 7 shows the characteristics, overall results, and category-based results of trackers. Fig. 11 displays the precision plot and success plot for the overall dataset. REPS obtains outstanding performance with Pre and Suc of 0.690 and 0.486, respectively. To evaluate the strengths and limitations of trackers, we further perform attribute-based evaluations. Figs. 12 and 13 show the precision plot and success plot of per-attribute, respectively. Overall, extensive results would demonstrate the robustness and applicability of REPS under different scenarios, environmental conditions, and satellite sensors.

5. Conclusions

SOT in SV holds great promise in the field of satellite surveillance. The article proposes the REPS framework that explores SOT from the perspective of tracking and segmentation. To address the inconsistency of semantic representation, we design an RE architecture to achieve rotation equivariant tracking of SV objects. To improve the tracking accuracy and semantic representations simultaneously, a PL refinement is proposed to refine the spatial distribution of objects by constructing a per-pixel foreground probability map. Moreover, the proposed AGF synergizes the tracking and segmentation results to obtain compact outputs for satellite object representations. Extensive experiments validate the superiority of the proposed method. Future work will focus on multiple object tracking of satellite objects.

CRediT authorship contribution statement

Yuzeng Chen: Conceptualization, Formal analysis, Methodology, Validation, Writing – original draft. **Yuqi Tang:** Funding acquisition, Investigation, Writing – review & editing. **Qiangqiang Yuan:** Data curation, Project administration, Resources, Writing – review & editing. **Liangpei Zhang:** Funding acquisition, Visualization, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgments

We are grateful to anonymous reviewers and Editors. This work was supported by the National Natural Science Foundation of China under Grant 41971313, 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., 2016b. Fully-Convolutional Siamese Networks for Object Tracking. In: Proc. Eur. Conf. Comput. Vis. (ECCV)/IEEE Trans. Signal Process, pp. 850–865.

Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the devil in the details: delving deep into convolutional nets. In: Proceedings of the British Machine Vision Conference.

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. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 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., 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, 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, 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.

Cui, Y.Y., Hou, B.A., Wu, Q., Ren, B., Wang, S., Jiao, L.C., 2022. Remote sensing object tracking with deep reinforcement learning under occlusion. IEEE Trans. Geosci. Remote Sens. 60.

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

Danelljan, M., Häger, G., Shahbaz Khan, F., Felsberg, M., 2014. Accurate scale estimation for robust visual tracking. In:Proceedings of the British Machine Vision Conference, pp. 65.61–65.11.

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., Khan, F.S., Felsberg, M., van de Weijer, J., 2014b. Adaptive color attributes for real-time visual tracking. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR), pp. 1090–1097.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, p. arXiv:2010.11929.

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.

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

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.

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. In: 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.

Gupta, D.K., Arya, D., Gavves, E., 2021. Rotation equivariant siamese networks for tracking. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, pp. 12357–12366.

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

He, J., Yuan, Q., Li, J., Xiao, Y., Zhang, L., 2023. A self-supervised remote sensing image fusion framework with dual-stage self-learning and spectral super-resolution injection. ISPRS J. Photogramm. Remote Sens. 204, 131–144.

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.

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. In: IEEE Trans. Pattern Anal. Mach. Intell. PP.

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

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

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

Li, Y.F., Bian, C.J., Chen, H.Z., 2022. Object tracking in satellite videos: correlation particle filter tracking method with motion estimation by kalman filter. IEEE Trans. Geosci. Remote Sens. 60.

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

Li, F., Tian, C., Zuo, W., Zhang, L., Yang, M.H., 2018b. Learning spatial-temporal regularized correlation filters for visual tracking. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4904–4913.

Li, Y., Zhu, J., Hoi, S.C.H., Song, W., Wang, Z., Liu, H., Aaai, 2019. Robust estimation of similarity transformation for visual object tracking. In: 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., Fu, C., Ding, F., Huang, Z., Lu, G., 2020. AutoTrack: towards high-performance visual tracking for UAV with automatic spatio-temporal regularization. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).

Mehta, S., Rastegari, M., 2022. Separable Self-attention for Mobile Vision Transformers, p. arXiv:2206.02680.

Patel, D., Upadhyay, S., 2013. Optical flow measurement using Lucas Kanade method. Int. J. Comput. Appl. 61, 6–10.

Possegger, H., Mauthner, T., Bischof, H., Ieee, 2015. In defense of color-based model-free tracking. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2113–2120.

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., Zhang, L., 2019a. 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., 2019b. 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.

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. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, pp. 1–9.

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.S., 2019. Fast online object tracking and segmentation: a unifying approach. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1328–1338.

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.

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

Xiao, Y., Su, X., Yuan, Q., Liu, D., Shen, H., Zhang, L., 2022a. Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection. IEEE Trans. Geosci. Remote Sens. 60, 1–19.

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

Xie, F., Wang, C., Wang, G., Cao, Y., Yang, W., Zeng, W., 2022. Correlation-aware deep tracking. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8741–8750.

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., Wu, K., Wang, D., Fu, J., Lu, H., 2021b. LightTrack: finding lightweight neural networks for object tracking via one-shot architecture search. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 15175-15184.

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).

Yang, X., Zhou, Y., Zhang, G., Yang, J., Wang, W., Yan, J., Zhang, X., Tian, Q., 2022. The KFIoU Loss for Rotated Object Detection, p. arXiv:2201.12558.

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., 2022. Joint feature learning and relation modeling for tracking: a one-stream framework. In: Proc. Eur. Conf. Comput. Vis. (ECCV).

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), pp. 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.

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.

Zhu, K., Zhang, X.D., Chen, G.Z., Li, X.W., Cai, P.H., Liao, P.Y., Wang, T., 2022. Multi-oriented rotation-equivariant network for object detection on remote sensing images. IEEE Geosci. Remote Sens. Lett. 19.