*remote sensing*

MDPI

*Article*

# Single Object Tracking in Satellite Videos: Deep Siamese Network Incorporating an Interframe Difference Centroid Inertia Motion Model

**Kun Zhu [1], Xiaodong Zhang [1,*], Guanzhou Chen [1], Xiaoliang Tan [1], Puyun Liao [1], Hongyu Wu [2], Xiujuan Cui [3], Yinan Zuo [3] and Zhiyong Lv [4]**

[1] State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; zkun@whu.edu.cn (K.Z.); cgz@whu.edu.cn (G.C.); xl_tan@whu.edu.cn (X.T.); liaopuyun@whu.edu.cn (P.L.)
[2] Satellite Image Geometric Correction Processing of Chang Guang Satellite Technology Co., Ltd., Changchun 130051, China; wuhongyu@whu.edu.cn
[3] School of Geoscience, Yangtze University, Wuhan 430100, China; 17839164234@163.com (X.C.); 18562316570@163.com (Y.Z.)
[4] School of Computer and Engineering, Xi'an University of Technology, No. 5 Jin Hua South Road, Xi'an 710048, China; zhiyongLyu@xaut.edu.com
* Correspondence: zxdlmars@whu.edu.cn; Tel.: +86-27-6877-8033

**Abstract:** Satellite video single object tracking has attracted wide attention. The development of remote sensing platforms for earth observation technologies makes it increasingly convenient to acquire high-resolution satellite videos, which greatly accelerates ground target tracking. However, overlarge images with small object size, high similarity among multiple moving targets, and poor distinguishability between the objects and the background make this task most challenging. To solve these problems, a deep Siamese network (DSN) incorporating an interframe difference centroid inertia motion (ID-CIM) model is proposed in this paper. In object tracking tasks, the DSN inherently includes a template branch and a search branch; it extracts the features from these two branches and employs a Siamese region proposal network to obtain the position of the target in the search branch. The ID-CIM mechanism was proposed to alleviate model drift. These two modules build the ID-DSN framework and mutually reinforce the final tracking results. In addition, we also adopted existing object detection datasets for remotely sensed images to generate training datasets suitable for satellite video single object tracking. Ablation experiments were performed on six high-resolution satellite videos acquired from the International Space Station and "Jilin-1" satellites. We compared the proposed ID-DSN results with other 11 state-of-the-art trackers, including different networks and backbones. The comparison results show that our ID-DSN obtained a precision criterion of 0.927 and a success criterion of 0.694 with a frames per second (FPS) value of 32.117 implemented on a single NVIDIA GTX1070Ti GPU.

**Keywords:** satellite video; single object tracking; Siamese network; model drift; object detection; remote sensing



**Citation:** Zhu, K.; Zhang, X.; Chen, G.; Tan, X.; Liao, P.; Wu, H.; Cui, X.; Zuo, Y.; Lv, Z. Single Object Tracking in Satellite Videos: Deep Siamese Network Incorporating an Interframe Difference Centroid Inertia Motion Model. *Remote Sens.* **2021**, *13*, 1298. https://doi.org/10.3390/rs13071298

Academic Editor: Edoardo Pasolli

Received: 11 March 2021
Accepted: 26 March 2021
Published: 29 March 2021



## 1. Introduction

Single object tracking is a fundamental research area in the field of computer vision and has attracted wide attention, as it is applied in many fields, such as automatic driving, human-computer interactions, video surveillance, and augmented reality [1–4]. After capturing the position and extent of an object in the first frame of a video, single object tracking obtains the position and extent of the target in subsequent frames [5]. In actual applications, besides the tracking accuracy, speed is also a criterion to measure performance [6]. Occlusion, illumination, deformation and motion make single object tracking challenging [7,8]; therefore, many researchers have developed trackers to solve these problems.

Modern single object trackers focus on natural scene videos obtained by optical cameras, and can be broadly divided into two streams, namely correlation filter-based (CF-based) methods and deep learning-based (DL-based) methods [6,9]. Correlation filters originally came from the field of signal processing in the Fourier domain, while filter correlation indicates the similarity between two signals. Tracking methods based on correlation filters find a filter template, and then perform convolution with next frame. The area with the largest response value is considered to contain the predicted target. Since Minimum Output Sum of Squared Error filter (MOSSE) [10] introduced correlation filters in object tracking, many outstanding natural scene video tracking algorithms (e.g., Discriminative Scale Space Tracking (DSST) [11], Kernelized Correlation Filter (KCF) [12], Color Names (CN) [13]) were proposed. Due to the powerful feature extraction ability of convolutional neural networks (CNNs), these trackers now employ deep features to boost their tracking performance. Continuous Convolution Operators for Visual Tracking (C-COT) [14] and Efficient Convolution Operator (ECO) [15] adopt deep CNN features to improve the precision and speed of the trackers gradually.

Single object tracking methods based on deep learning have drawn much attention. Among them, the Siamese network-based trackers draw much attention. In object tracking tasks, these Siamese trackers normally consist of two branches: the template branch and the search branch. They learn a general similarity map by cross-correlation between the feature representations obtained from these two branches. After the offline training, the Siamese similarity function kept fixed during tracking [16]. Fully-Convolutional Siamese Networks (SiamFC) [17] utilizes fully-convolutional Siamese network in tracking field and enable to use the public datasets effectively. Siamese Region Proposal Network (SiamRPN) [6] utilizes a region proposal network (RPN) [18] to accommodate the diversity of object scales during tracking and regards the tracking as a one-shot local detection task. Distractor-aware Siamese Networks (DaSiamRPN) [19] utilizes static images from detection datasets through augmentation techniques to generate image pairs for dynamic tracking. SiamRPN++ [16] replaces the shallow backbone networks with modern deep neural networks, like ResNet50 [20], and achieves significant performance gains on multiple object tracking benchmarks.

High-resolution satellite videos are available with the development of remote sensing platforms for earth observation technologies [21]. These videos gazing objects in a specific area on the ground, enabling dynamic monitoring of moving objects [5,9]. The SkySat-1 satellite with a resolution of around 1 m launched by Skybox Imaging in 2013 is the first civilian satellite to capture objects in high-resolution panchromatic video [22]. The SkySat-1 captures up to 90-s video clips at 30 frames per second (FPS) (https://www.satimagingcorp.com/satellite-sensors/skysat-1/ (accessed on 18 April 2020)). In 2015, the Changchun Institute of Optics, Fine Mechanics, and Physics launched the China's first commercial remote sensing video satellite, "Jilin-1" satellite, captured at 25 FPS. It can obtain video data at a resolution of 1.12 m [23]. The resolution of the newest "Jilin-1" products can reach to 0.72 m. In 2016, the International Space Station (ISS) released a high-definition video captured at 30 FPS with a ground sample distance as fine as 1 m. The frame format of this video is 3840 $\times$ 2160 pixels, and it is used for object tracking contest by Deimos Imaging and Urthecast (http://www.grss-ieee.org/community/technical-committees/data-fusion (accessed on 9 October 2019)). The emergence of these video datasets will drive the development of satellite video object tracking technologies.

Compared with natural scene videos, single object tracking in satellite videos shows the following attributes [5,24], which make the natural scene-based object tracking algorithms inapplicable to satellite videos:

1. Overlarge frame size. The size of each frame in satellite video is larger than that of the natural scene video. Therefore, it is more time-consuming when dealing with satellite video. Some satellite video object tracking methods cannot achieve real-time processing [5,25].
2. Smaller object size. Some objects have only several pixels in satellite video, making it difficult for trackers in natural scenes to adapt the size changes.
3. Strong interferences. The interferences come from the high similarity among multiple moving targets and the poor distinguishability between the objects and the background.

To solve the above problems, our research exploits the powerful ability for feature extraction of convolutional neural network. By using existing object detection datasets from remote sensing images, a method for generating training datasets suitable for satellite video single object tracking is proposed. At the same time, a deep Siamese network (DSN) for obtaining the incipient tracking results and an interframe difference centroid inertia motion (ID-CIM) model for alleviating model drift are proposed. The main contributions of this paper are summarized as follows:

1. We propose a method for generating training datasets suitable for satellite video single object tracking tasks. Through the introduction of a variety of small objects, such as airplane, ship, vehicle, and other categories, the tracker can adapt to the size and scale changes of the targets. The effectiveness was validated in experiments with two public datasets DOTA [26] and DIOR [27].
2. We propose deep Siamese network (DSN) for real-time satellite video single object tracking. It is modified to adapt to different categories with large size differences in remote sensing images so as to train the datasets constructed in this paper. During testing, the template frame is only calculated once and can be considered as a one-shot local detection issue, which greatly accelerates the tracking speed.
3. We also propose the ID-CIM mechanism to alleviate model drift. By calculating the coordinates difference of centroid between the fixed interval frames in the first $N$ frames, the mechanism can find the movement pattern of the target. It involves multiple frames, thus avoiding contingency and improving robustness.

The ablation experiments were conducted on six satellite videos, and the tracking targets include trains, planes, a pair of vehicles and ships. We also compared the ID-DSN with other 11 state-of-the-art trackers, including different networks and backbones, to illustrate the effectiveness of the proposed framework and mechanisms. The rests of this paper is organized as follows. Section 2 introduces the related work involved in this work. Section 3 elaborates the proposed framework and mechanisms in detail. Section 4 implements ablation experiments and makes reliable analysis to the results. Finally, the conclusions are drawn in Section 6.

## 2. Related Work

In this section, we introduce some works related to the proposed framework according to the main contributions exhibited in Section 1. We also introduce the overall work flow of the Siamese network-based tracker and discuss existing satellite video object tracking methods and the mechanisms to alleviate model drift problem.

### *2.1. The Siamese Network-Based Trackers*

Single object tracking can be regarded as a similarity-learning problem, solvable with Siamese architectures [17,28,29]. The Siamese network-based trackers consist of two branches, namely the template branch and the search branch. The template branch encodes exemplar image $z$, while the search branch encodes candidate image $X$. In general, candidate image possesses a larger size than that of exemplar image. To find the object position in the candidate image, we calculate all possible locations and choose the candidate region with the maximum response value to the features that encoded in the exemplar image.

Shared processes in Siamese trackers are the template branch and the exemplar branch that undergo the same feature extraction operation, which can be represented by a convolutional embedding function $\varphi$. The encoded information from the two branches performs a cross-correlation operation on the final feature maps. Eventually, the candidate region with the maximum response value is considered as the updated location of the tracking object. The overall process can be expressed as:

$$f(z, X) = \varphi(z) * \varphi(X), \tag{1}$$

where $\varphi(z)$ and $\varphi(X)$, respectively, represent the encoded information from the template branch and the search branch, $*$ is the correlation operation, and $f(z, X)$ represents the similarity calculated by the cross-correlation operation. During offline training, the logistic loss function comprised by the positive and negative samples can be expressed as:

$$L(y, v) = log(1 + exp(-yv)), \tag{2}$$

where the values of $y$ belong to $\{1, -1\}$, which is used to classified whether the elements on the feature map are positive or negative. If the element is in a given neighborhood of center, it is assigned as a positive sample. Otherwise, it is a negative sample. $v$ is the real-valued score for each exemplar-candidate pair. Hence, for the parameters of the CNN $\theta$ embedded in the Siamese tracker, which can be obtained by Stochastic Gradient Descent (SGD) to the task:

$$\arg\min_{\theta} \sum L(y, f(z, X; \theta)). \tag{3}$$

During tracking, the search image centered at the object position in the previous frame is regarded as prior information. The candidate region with the maximum response value in the current frame is considered the updated location of the tracking object.

### *2.2. Satellite Video Single Object Tracking Methods*

Researchers have proposed several methods for single object tracking in satellite videos. Du et al. [5] proposed a multi-frame optical flow difference tracker (MOFT), which fused with the HSV color system and integral image to track targets in satellite videos. Shao et al. [30] proposed a velocity correlation filter (VCF) algorithm based on KCF [12], which employed velocity features to keep the discriminative ability high enough to detect moving objects in satellite videos and an inertia mechanism to alleviate model drift. In satellite videos, as backgrounds are complex with fewer features contained in the target, the tracker will lose the object, a condition called model drift, as shown in Figure 1. Shao et al. [31] proposed a hybrid kernel correlation filter (HKCF) tracker, which combines optical flow to detect variation pixels of the object and uses histogram of oriented gradient (HOG) for capturing the contour and texture information. Guo et al. [9] proposed a high-speed CF-based tracker, which utilized the global motion features of a moving object to constrain the tracking process, and adopted Kalman filter to correct the tracking trajectory of the moving target. Although these trackers can execute satellite video single object tracking, the hand-designed features, such as HOG and HSV color system, need to be carefully designed [32], and some of these algorithms are not equipped with a mechanism to alleviate model drift. In addition, some methods only tested on three videos, which does not guarantee the robustness of the tracker. Therefore, it is necessary to construct a tracker that can utilize deep features and alleviate model drift to realize satellite video single object tracking.

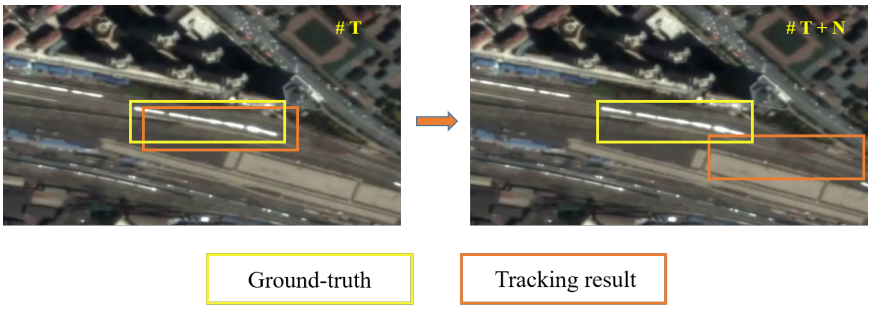



**Figure 1.** The phenomenon of model drift. When the overlap between the bounding-box of the tracking result and the bounding-box of ground-truth is less than the given threshold, it is considered that the tracker loses the target, also known as model drift.

### *2.3. The Mechanisms to Alleviate Model Drift*

In order to alleviate model drift problem, some approaches have been proposed by utilizing additional mechanisms. Zhang et al. [33] utilized an online learning method to re-detect the target. Their method is based on logistic regression classifier, which is updated before re-detecting the undetermined object to ensure the authenticity of tracking performance. Guan et al. [34] proposed a drift and occlusion detection method by considering the tracking loss in terms of spatial locality constraint and temporal continuity. These two criteria are fused to evaluate the current tracking status and trigger relevant mechanisms as needed. Luo et al. [35] designed an adaptive selection method for feature channels that obtains a feature map for each channel by correlation calculation. They adopted Peak to Sidelobe Ratio (PSR) indicator to reflect the capacity of feature channel to distinguish object and backgrounds, thus mitigating model drift. Huang et al. [36] proposed a bidirectional-tracking scheme, which contains three modules. The forward tracking module searches all candidate regions, and the backward tracking module calculates the confidence of each candidate region based on their historical information, while the integration module integrates the results of the other two modules to ascertain the final tracking target and the model update strategy for the current frame. Shao et al. [30] proposed an adaptive inertia emergency mechanism, which simultaneously calculates an inertia position in each frame. When the difference between the tracker result and the inertia emergency result is greater than a given threshold, the tracking result is replaced with the inertia result.

Generally, the methods in Reference [35,36] are complicated and consume multitudinous computing resources, thus slowing down the speed of trackers. The mechanism in Reference [30] contains many parameters, and the method over relies on the tracking results of the first several frames. In this paper, we propose a centroid inertia motion model based on interframe difference, because the target trajectory is consecutive and approximately linear in satellite video [9]. It simplifies the complicated mechanism, like in Reference [30], and contains less parameters. It does not contain complex principles, like Kalman filter [37], which can be regarded as an effective trajectory corrector, nor does it require multiple pre-defined matrices. In addition, the problem of burn-in period needs to be dealt with when Kalman filter is used [9]. This simple and effective mechanism can significantly improve the detection accuracy without reducing the detection speed by calculating the coordinate difference of centroids between the fixed interval frames in the first $N$ frames.

## 3. Methodology

In this section, we elaborate the architecture of the proposed ID-DSN in detail. The overall framework is shown in Figure 2. It mainly consists of two modules, namely the DSN module (shown as green and orange blocks) and the ID-CIM module (shown as blue blocks). The former module obtains the incipient tracking result based on deep Siamese neural networks, while the latter module alleviates model drift. These two mod-

ules complement each other to obtain the final tracking result. We also introduce the method of generating training datasets suitable for satellite video single object tracking.

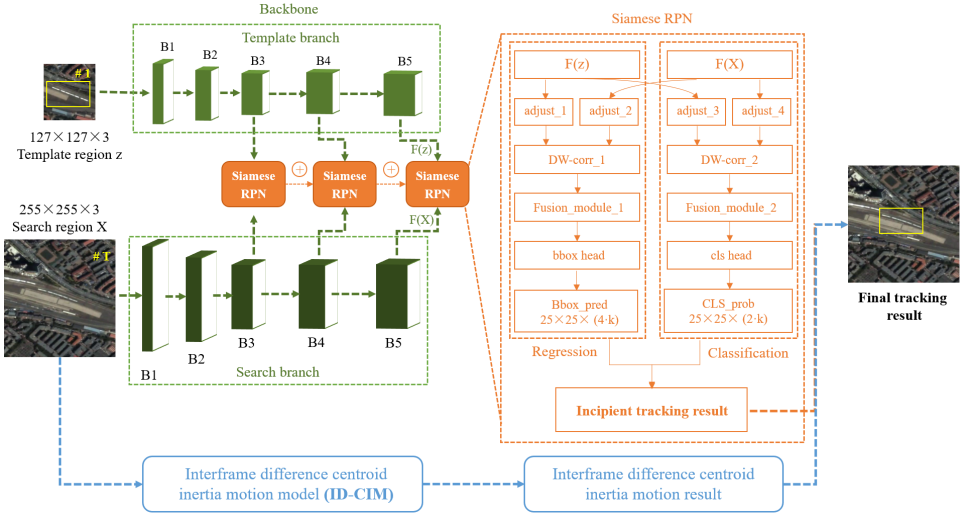



**Figure 2.** The overall structure of interframe difference deep Siamese network (ID-DSN). It consists of the DSN module (shown as green and orange blocks) and the interframe difference centroid inertia motion (ID-CIM) module (shown as blue blocks). These two modules mutually reinforce the final tracking result. The DSN inherently includes a template branch and a search branch, and they share weights.

### *3.1. The Structure of the DSN*

The DSN consists of backbone component and Siamese RPN component. The backbone component is used to extract the features of template branch and search branch. These two branches share weights, which are shown as green blocks in Figure 2. In this paper, we adopted the ResNet50-like structure. The original ResNet50 [20] can be divided into five stages, and the output of each stage's last residual block is denoted as $\{B_1, B_2, B_3, B_4, B_5\}$, with strides of {2, 4, 8, 16, 32} pixels corresponding to the input image. The size of the target in satellite video is relatively small; therefore, we reduced the strides of the last two stages from 16 and 32 to 8, that is, the strides of five stages are now {2, 4, 8, 8, 8} pixels, respectively. We utilized dilated convolutions [38] to increase the receptive field of the last two blocks, while the dilation rates are 2 and 4, respectively. In addition, these two blocks adjust channels to 256 through $1 \times 1$ convolution layer. The template branch and search branch undergo the same feature extraction operation, except that their input images are of different sizes.

Referring to the idea of RPN [18], the Siamese region proposal network takes the template branch and the search branch as the inputs of classification task and regression task. Siamese RPN obtains the position of the target in the search frame by calculating the correlation between the template branch and the search branch, the structure is shown in the orange part of Figure 2. Specifically, considering that the classification task and the regression task are asymmetric in the RPN structure, the template frame and the search frame undergo two convolution layers that are not shared but have the same number of channels. The two feature maps are correlated channel by channel. The results are fused by another convolution layer. The classification and regression branches are attached to the fusion layers; these two branches jointly notarize the range of the target in the search frame.

During training, the anchors are labeled as positive or negative. The method to generate an anchor was proposed in Faster R-CNN [18] and its shape is determined by size and ratio. In the object tracking task, the actual category of the target is not concerned, so the anchors are divided into foregrounds and backgrounds based on intersection-over-union (IoU) between the region proposal and ground-truth bounding-box. If an anchor possesses an IoU greater than 0.6 with ground-truth bounding-box, it will be designated as

positive. On the contrary, an anchor is designated as negative if it has an IoU less than 0.3 with ground-truth bounding-box. The shape of the target usually does not change in a single satellite video; however, in order to solve the size change problem of different kinds of targets in multiple videos, we adopted a single scale of 8 with 5 ratios of [0.33, 0.5, 1, 2, 3]. We utilize cross-entropy loss as the loss function of classification task. The form is as follows:

$$L_{cls}(p_i, p_i^*) = -log[p_i \cdot p_i^* + (1 - p_i)(1 - p_i^*)], \tag{4}$$

where $p_i$ represents the probability that an anchor is predicted to be the target, and $p_i^*$ is the value of the ground-truth bounding-box, represented as:

$$p_i^* = \begin{cases} 0, & negative \\ 1, & positive \end{cases}. \tag{5}$$

For the regression task, we utilize tuple $G = (G_x, G_y, G_w, G_h)$ to represent the center coordinates, width, and height of the ground truth bounding-box, and tuple $P = (P_x, P_y, P_w, P_h)$ to represent the center coordinates, width, and height of the predicted box. The normalized distance can be expressed as:

$$\begin{array}{cc} d_x = (G_x - P_x)/P_w & d_y = (G_y - P_y)/P_h \\ d_w = ln(G_w/P_w) & d_h = ln(G_h/P_h) \end{array}. \tag{6}$$

These four variables can be considered as a bounding-box regression from a predicted box to a nearby ground-truth box. The loss function of the regression task is defined as:

$$L_{reg}(G, P) = \sum_{i \in \{x,y,w,h\}} smooth_{L_1}(d_i), \tag{7}$$

in which

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & if|x| < 1 \\ |x| - 0.5 & otherwise \end{cases}. \tag{8}$$

Finally, we optimize the total loss function:

$$loss = L_{cls}(p_i, p_i^*) + \lambda \cdot L_{reg}(G, P), \tag{9}$$

where $\lambda$ is adopted to balance $L_{cls}$ and $L_{reg}$ items, and we set it to 1.2 in this paper.

Hence, the DSN can be trained end-to-end using stochastic gradient descent (SGD) after the feature extraction blocks. In the neural network structure, the concerns of the feature maps in different depths are also inconsistent. Features in shallow layers focus on the appearance of the target, such as shape and color, which help target positioning. Features in deep layers focus on the semantic information, which can resist the deformation, blur and other factors in the object's movement [16]. Therefore, combining information from hierarchical feature maps can help improve the performance of object tracking. This idea has been adopted in the Feature Pyramid Network (FPN) [39] method. However, we adopted a weighted sum of the separate classification and regression outputs from different blocks instead of aggregating hierarchical convolutional features. Specifically, we concentrate the Siamese RPN results obtained from $B_3$, $B_4$, and $B_5$ blocks by different weights, which are shown as the orange $\oplus$ in Figure 2. As mentioned at the beginning of this section, these three blocks have the same feature map sizes and channel numbers. This process can be expressed as follows:

$$\begin{array}{c} cls_{sum} = \sum_{i=3}^{5} \alpha \cdot cls_i \\ loc_{sum} = \sum_{i=3}^{5} \beta \cdot loc_i \end{array}, \tag{10}$$

where $cls_{sum}$ and $loc_{sum}$, respectively, represent the weighted sum results of the classification and location tasks, $i$ is the index of blocks, and $\alpha$ and $\beta$ represent the weight parameters, made available through the $softmax$ function during training. $cls_i$ and $loc_i$ contain results from an individual block.

During tracking, the template branch only calculates the first frame. In order to reduce the calculation cost, we cropped the center $7 \times 7$ regions as the feature map of the template branch to participate in calculations. The search branch operates in the subsequent frames of the video, so this process can be regarded as a one-shot local detection. These two branches are cross-correlated through Siamese RPN. The candidate region with the maximum response value is considered as the updated location of the tracking object.

### 3.2. Interframe Difference Centroid Inertia Motion Model

In satellite videos, the trajectory of a target is relatively simple, clear, and the target hardly undergoes serious deformation [9]. Utilizing these attributes to solve the problem of model drift, we proposed an interframe difference centroid inertia motion model, the mechanism is shown in Figure 3.

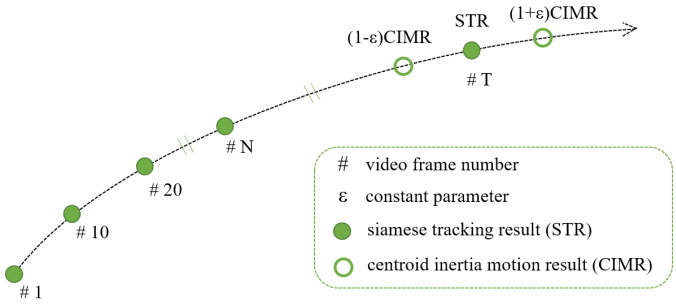



**Figure 3.** The mechanism of the proposed ID-CIM model.

The Siamese tracking result (STR) obtained by the deep neural network in the first $N$ frames are reliable, noting that $N$ is divisible by 10. We calculate the central coordinates of the bounding box obtained by DSN every 10 frames in the first $N$ frames. Then, we take the central coordinates of the bounding box in the currently participating frame minus that in the previous participating frame. We add up these Euclidean distance differences among the frames participated in the calculation. The sum of these differences is divided by $N$ and the result is the average of the center point difference of the first $N$ frames. The average differences distributed in the $x$ and $y$ coordinate axis can be marked as $\Delta\bar{x}$ and $\Delta\bar{y}$, which roughly reflect the trajectory and motion state of the target. Because the results are calculated by multiple frames in a uniform arrangement, it avoids contingency and improves robustness.

When calculating the centroid point coordinates of the target extent in frame $T$ ($T > N$), we first obtain the STR through the deep neural network, which is shown as solid point in frame $T$ of Figure 3. The centroid inertia motion result (CIMR) of frame $T$ is obtained by the following formulas:

$$\begin{aligned} X_{CIMR} &= X_{FTRP} + \Delta\bar{x} \\ Y_{CIMR} &= Y_{FTRP} + \Delta\bar{y}' \end{aligned} \tag{11}$$

where $X_{CIMR}$ and $Y_{CIMR}$ represent the CIMR of current frame, and $X_{FTRP}$ and $Y_{FTRP}$ represent the final tracking result of previous frame (FTRP). If $STR \in [(1-\varepsilon)CIMR, (1+\varepsilon)CIMR]$, we set $STR$ as the final tracking result of the current frame, otherwise, it is $CIMR$. We note that $\varepsilon$ is a constant parameter of the ID-CIM. We set it to 0.001 in this work. The general procedure of ID-CIM is summarized in Algorithm 1.

**Algorithm 1** The general procedure of ID-CIM.

**Require:** Variable frame number $t$, Fixed frame number $N$, $STR_{(t)}$ and $CIMR_{(t)}$ of frame $t$, Parameter $\varepsilon$;
**Ensure:** Final tracking result of frame $t$ ($FTR_{(t)}$);
**if** $t < N$ **then**
  Set $FTR_{(t)} = STR_{(t)}$
**else**
  **if** $(1-\varepsilon)CIMR_{(t)} < STR_{(t)} < (1+\varepsilon)CIMR_{(t)}$ **then**
    Set $FTR_{(t)} = STR_{(t)}$
  **else**
    Set $FTR_{(t)} = CIMR_{(t)}$
  **end if**
**end if**

### *3.3. The Method of Generating Satellite Video Object Tracking Training Datasets*

Abundant labeled training datasets promote the development of CNN-based deep models. However, in the field of satellite video single object tracking based on a Siamese network, there is no available public training dataset as far as we know. Therefore, we plan to generate training datasets suitable for satellite video single object tracking from the existing remote sensing image object detection datasets. Different from object tracking, object detection consists of two tasks: category recognition and coordinates positioning [40]. Many outstanding natural scene image object detection datasets, such as ImageNet [41] and COCO [42], have been proposed to facilitate this issue. In the field of remote sensing image object detection, DOTA [26] and DIOR [27] are two widely used large-scale datasets. Table 1 shows the basic information of these two datasets.

**Table 1.** The basic information of DOTA and DIOR datasets.

| Dataset | # Images | # Categories | # Instances | # Size | # Annotation Type | # Image Type |
|---|---|---|---|---|---|---|
| DOTA | 2806 | 15 | 188,282 | (800, 800)∼(4000, 4000) px | Oriented bounding box | Optical images |
| DIOR | 23,463 | 20 | 192,472 | (800, 800) px | Horizontal bounding box | Optical images |

DOTA is a large-scale dataset for object detection in aerial images, which consists of 2806 remote sensing images with pre-divided 1411 training images, 458 validation images, and 937 testing images, covering 15 categories. The sizes of these images range from 800 × 800 to 4000 × 4000 pixels. DOTA contains 188,282 fully annotated instances; each of them is labeled by an oriented bounding box. We convert the originally oriented bounding-boxes into horizontal bounding-boxes as ground truth. DIOR is another large-scale benchmark for object detection in optical remote sensing images, which contains 23,463 images with a size of 800 × 800 pixels, covering 20 categories. Each instance of DIOR dataset is labeled by a horizontal bounding box with a total number of 192,472. During the training of Siamese tracker, the category of target is completely ignored [6,17]. Therefore, we can generate the training datasets suitable for satellite video single object tracking with the help of numerous annotations contained in these two remote sensing object detection datasets; we achieve this through image clipping, instance scaling, etc.

The input for the Siamese network is a pair of images, one for template branch and the other for search branch, with different image sizes. In this paper, we set the input size of the template branch to 127 × 127 pixels and the search branch to 255 × 255 pixels. For each labeled sample, we construct the training image pairs by the following formula:

$$(w + (w + h)/2) \cdot (h + (w + h)/2) \cdot s^2 = S^2, \tag{12}$$

where $w$ and $h$, respectively, represent the width and height of the labeled bounding-box, $s$ represents the scale factor, and $S$ represents the input image sizes of the Siamese network, which equals 127 for the template branch. The search branch undergo the similar process. Each bounding-box takes its original center point as its processed center point. The $(w + h)/2$ term is the context margin added in the width and height axis in order to cover the bounding-box as completely as possible when cropping.

For the DOTA dataset, due to its large image size and too many objects on each image, we cropped these DOTA images into patches of $1000 \times 1000$ pixels with an overlap of 500 pixels. Finally, we obtained 23,573 training samples. For the DIOR dataset, we adopted its inherent 23,463 images, and finally obtained 23,463 training samples. Since each image contains multiple labeled bounding-boxes, a sample pair is randomly selected from the generated sample datasets as the training data source.

## 4. Experiments and Result Analysis

This section consists of four subsections. Section 4.1 describes the general settings of experiments, including experimental dataset description, evaluation criteria, and implementation details. Sections 4.2 and 4.3 are internal ablation experiments, which illustrate the effectiveness of the proposed training datasets generation scheme and model drift mitigation method. Lastly, Section 4.4 is an external ablation experiment, which indicates the effectiveness of ID-DSN by comparing the proposed framework in this work with other 11 state-of-the-art trackers.

### *4.1. Experimental Settings*

#### 4.1.1. Experimental Datasets Description

We performed the experiments on six satellite videos with a total of 2930 frames. The target is labeled by a horizontal bounding box in each frame. Taking into the slight deformation of the target appearance, and to enhance the robustness of the tracker at the same time, the extents of bounding boxes are different for the same target in different frames. The testing dataset thumbnails are shown in Figure 4. Video 1, 2, 4, 5, and 6 were acquired from "Jilin-1" satellites provided by Chang Guang Satellite Technology Co., Ltd. (Changchun, China). (http://mall.charmingglobe.com/Sampledata (accessed on 28 October 2019)), captured at 25 FPS. The ground sample distance (GSD) of the frame is 0.92 m. The frame is $3840 \times 2160$ pixels and approximately covers 2.7 km $\times$ 1.5 km. Video 3 was acquired from the International Space Station (ISS) provided by Deimos Imaging and UrtheCast (Vancouver, BC, Canada). (https://ieee-dataport.org/open-access/data-fusion-contest-2016-dfc2016 (accessed on 9 October 2019)) [43], captured at 30 FPS. The GSD of the frame is as fine as 1 m. The frame is $3840 \times 2160$ pixels and approximately covers 3.8 km $\times$ 2.1 km. The test sequences are labled by the attributes of small size (SS), residual shadow (RS), partial occlusion (PO), poor target-background discriminability (PTBD), shape deformation (SD), and poor general field illumination (PGFI), which represent the challenging aspects in satellite video object tracking task. The detailed descriptions of these experimental datasets are given in Table 2.

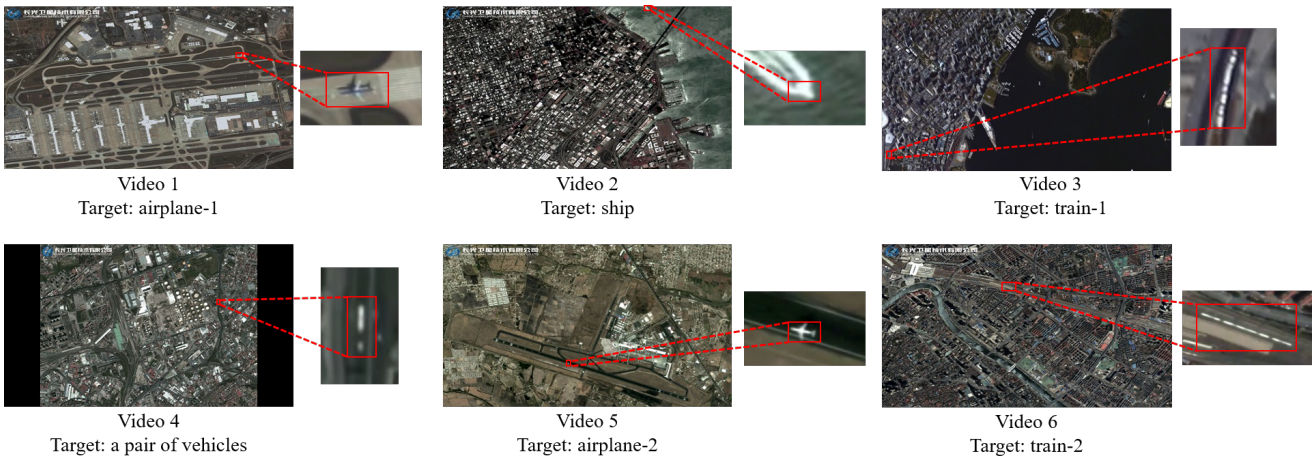



**Figure 4.** The testing datasets used in the experiments. There are six satellite videos, and the tracking objects consist of airplane, ship, train, and a pair of vehicles.

**Table 2.** The detailed descriptions of six experimental satellite videos.

| | Frame Size (px) | Resolution (m) | Target Size (px) | Target Category | Frame Number | FPS | Source | Attributes |
|---|---|---|---|---|---|---|---|---|
| video 1 | 3840 × 2160 | 0.92 | 48 × 35 | airplane | 568 | 25 | Jilin-1 | SS, RS PTBD, SD, PGFI |
| video 2 | 3840 × 2160 | 0.92 | 14 × 15 | ship | 500 | 25 | Jilin-1 | SS, PTBD, SD |
| video 3 | 3840 × 2160 | 1 | 29 × 79 | train | 362 | 30 | ISS | SD |
| video 4 | 3840 × 2160 | 0.92 | 10 × 30 | a pair of vehicles | 500 | 25 | Jilin-1 | SS, PO, SD |
| video 5 | 3840 × 2160 | 0.92 | 39 × 35 | airplane | 375 | 25 | Jilin-1 | SS, RS, SD, PGFI |
| video 6 | 3840 × 2160 | 0.92 | 154 × 63 | train | 625 | 25 | Jilin-1 | PTBD, SD |

#### 4.1.2. Evaluation Criteria

We adopted precision plot and success plot as evaluation criteria in our experiments, which are widely used in object tracking tasks. The precision plot reveals the percentage of frames in which the point distance (PD) between the target center location estimated by the tracker and the corresponding ground-truth center location is less than a range of thresholds. Different thresholds correspond to different percentages, so a precision plot can be obtained. The PD is defined as follows:

$$PD = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}, \tag{13}$$

where $(x_1, y_1)$ is the center point of the predicted target, and $(x_2, y_2)$ is the center point of the corresponding ground truth. The maximum threshold adopted in the precision plot equals 50 in this work.

For success plot, let $s_{pb}$ denotes the area of predicted bounding-box by tracker and $s_{gt}$ denotes the area of ground truth. An overlap degree (OD) can be defined by:

$$OD = \frac{s_{pb} \cap s_{gt}}{s_{pb} \cup s_{gt}}, \tag{14}$$

where $\cap$ and $\cup$ are the intersection and union of two areas, and their values are represented by the number of pixels in the corresponding area. When the OD of a frame is greater than the given threshold, it will be considered as a successful tracking frame. The ratios of the successful tracking frame at the thresholds ranged [0, 1] are plotted in success plots.

An effective tracker is manifested as large values with low thresholds in the precision plot and large values with high thresholds in the success plot. The quantized values are expressed in terms of the area under curves (AUC). In this paper, we adopted the AUC of the precision plot and the success plot to evaluate the performance of a tracker. Generally, a higher AUC indicates a better tracker. In addition, we adopted the FPS criterion to illustrate the speed of trackers. The higher the FPS value, the faster the tracking speed.

#### 4.1.3. Implementation Details

We implemented the proposed framework on the open source deep learning library Pytorch (https://pytorch.org/ (accessed on 28 May 2021)) and executed on a 64-bit Ubuntu 16.04 workstation with 8GB memory GeForce GTX1070Ti GPU. The network was trained end-to-end by Stochastic Gradient Descent for 30 epochs. Specifically, 15 epochs were used to train the backbone, and the remaining 15 epochs were used to train the Siamese RPN. We adopted ResNet50 as backbone for feature extraction, which was pre-trained on the ImageNet dataset. The batch size was set to 16. The learning rate, momentum, and weight decay were set to 0.005, 0.9, and 0.0001, respectively.

### *4.2. Ablation Experiment 1: The Validation of ID-DSN*

We adopted five training schemes to demonstrate the effectiveness of the tracking performance on the proposed ID-DSN. To improve the overall performance of the tracker, we also adopted the COCO [42] dataset and dealt with it as the guidelines described in Section 3.3. We adopted (a), (b), (c), etc., to represent each scheme in the tables of experimental results. The bold **P** and **S** represent the AUC of precision plot and the AUC of success plot, respectively. In each row, the bold number indicates the highest criterion, and the other tables are the same. The first row represents the dataset(s) used in the corresponding experiment, which are shown in Table 3.

**Table 3.** The comparison results of different training schemes for the proposed ID-DSN.

| | **P** | **S** | **FPS** |
|---|---|---|---|
| COCO (a) | **0.875** | 0.596 | 30.467 |
| DOTA+COCO (b) | 0.854 | 0.616 | 30.583 |
| DIOR+COCO (c) | 0.868 | **0.646** | **32.117** |
| DOTA+DIOR (d) | 0.854 | 0.603 | 30.775 |
| DOTA+DIOR+COCO (e) | 0.857 | 0.632 | 30.7 |

Comparing these five ablation experiments, as shown in Table 3, scheme (a) obtained the highest value for precision criterion (**P**). It was only trained with COCO dataset obtaining a precision value of 0.875, 0.007 higher than the second ranked scheme (c), which was jointly trained with the DIOR and COCO datasets, with a precision value of 0.868. However, in terms of the success criterion (**S**), scheme (c) obtained the highest value of 0.646 and higher than scheme (a) by 0.05. This was because the COCO dataset consists of 117,266 training samples, thus dominating the training in contrast to the two smaller remote sensing datasets, DOTA and DIOR. Generally, a natural scene image contains a few labeled instances, and the size of the target is larger than a target in remote sensing images. In addition, the overlap of bounding-boxes among the targets in natural scene images are smaller, which makes the tracker pay more attention to the tracking object and will not be affected by similar objects or backgrounds. Therefore, the precision criterion is higher when only trained with COCO dataset. However, the precision criterion describes the distance between the center point of predicted bounding-box and the center point of ground-truth, which cannot reflect the changes in the size and scale of the tracking object. The objects in satellite videos are small; hence, when trained only using the COCO dataset, a tracker cannot properly adapt to the change of scale. The categories contained in a remote sensing image dataset, such as ship, airplane, and vehicle, can help the tracker adapt to such scale changes. Therefore, the success criterion is significantly improved after adding remote sensing images for training.

Although schemes (b) and (c), in Table 3, have roughly the same number of training samples, the precision and success criteria obtained by joint training with DIOR dataset nevertheless, are higher than with the DOTA dataset, so we analyzed these two datasets. The image size in the DOTA dataset is larger and contains more instances. After the cropping operation, the DOTA dataset contained 23,573 images and 402,907 labeled instances.

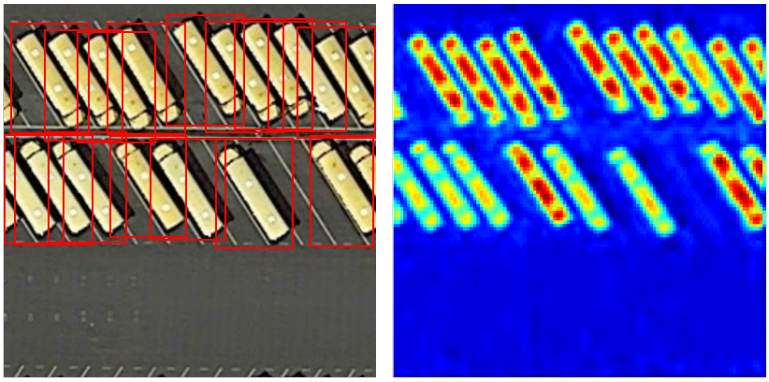
The DIOR dataset originally contained 23,463 images and 192,472 labeled instances. We defined the density of labeled instances as follows:

$$D_{li} = N_{ins} / N_{img}, \tag{15}$$

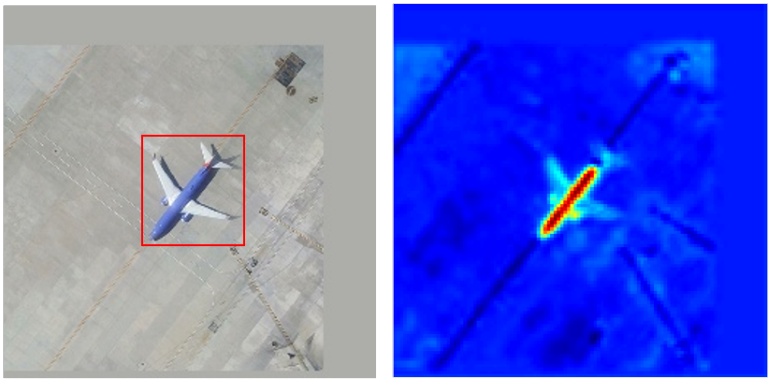
where the total number of labeled instances is $N_{ins}$ and $N_{img}$ represents the total number of images. Therefore, we can calculate that the $D_{li}$ of DOTA dataset is about 2.1 times that of DIOR dataset. However, densely clustered annotations are not suitable for tracking, especially when there is a high degree of similarity between the targets or backgrounds.

The results in scheme (d) were obtained by training two remote sensing datasets DOTA and DIOR. With the least amount of training samples, it obtained the lowest precision value of 0.854 and the penultimate success value of 0.603. It indicates that only using remote sensing datasets for training cannot obtain satisfactory precision and success values. However, compared with the results in scheme (a), scheme (d) still obtained a higher success value with fewer training samples of 47,036, which indicates that using remote sensing datasets for training can enhance the tracker's ability to adapt the scale and size changes of target, thus improving the success criterion of object tracking.

The results in scheme (e) were obtained by jointly training DOTA, DIOR and COCO datasets. Although this scheme has the largest amount of training samples, the precision and success values were not the highest. This phenomenon once again proves that highly overlapped samples will damage the performance of the tracker. Since this method adopts the DIOR dataset, it will "dilute" the excessively redundant annotations to a certain extent, so the results were still better than that in scheme (b). We selected samples from the DOTA and DIOR datasets and visualized feature maps from them, as shown in Figure 5.



(a) The sample from **DOTA** dataset and its feature map

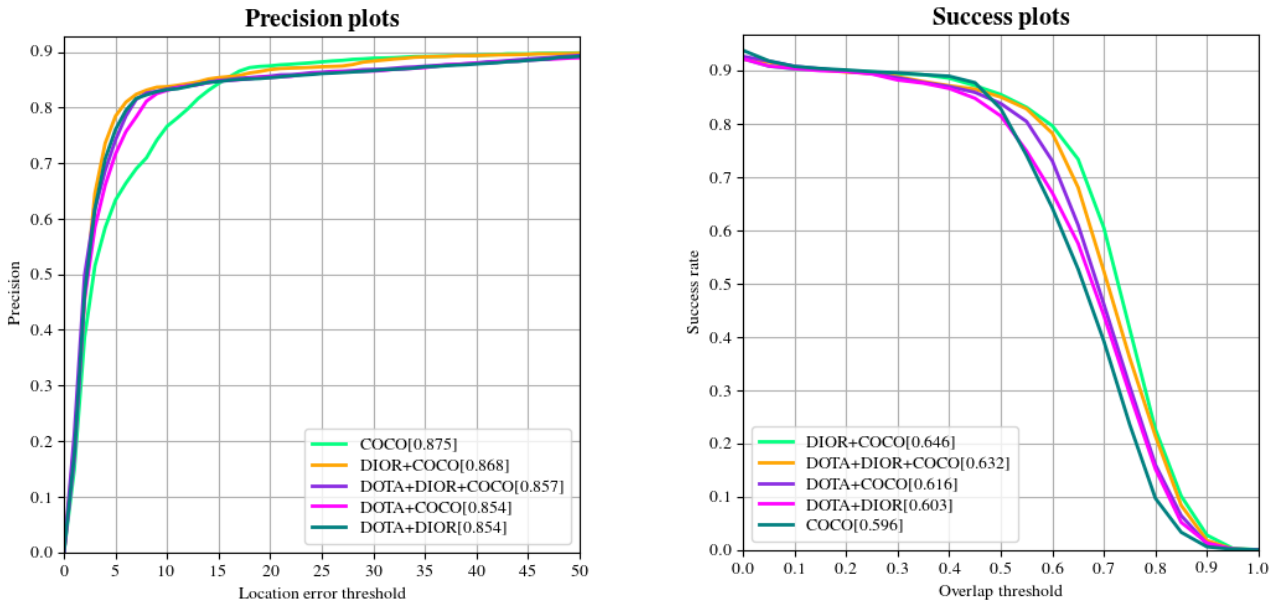


(b) The sample from **DIOR** dataset and its feature map

**Figure 5.** The samples from DOTA and DIOR datasets and their feature maps.

From Figure 5, we can see that the samples in the DOTA dataset have a high degree of overlap. The representations of these samples are similar on the feature map, thus hindering the improvements in the performance of the tracker. The sample distribution in DIOR dataset is relatively sparse, which can help the tracker focus on the current target and remain not affected by the surrounding targets. In addition, the DIOR dataset includes airplanes, ships, vehicles, and other such objects, which improves the capability of the tracker to adapt to changes in object scale. By sacrificing a little precision, the success value improved in this comparative experiment, illustrating the effectiveness of the proposed method for generating training samples in appropriate datasets. The method in scheme (c) served as our baseline.

We also calculated the tracking speed of the five schemes. Although they are on the same order of magnitude, the scheme (c) obtained the highest FPS value of 32.117, which can achieve real-time processing for the satellite videos used in this paper. The precision plots and success plots about the comparison results of different training schemes for the proposed ID-DSN are shown in Figure 6.




(**a**) The precision plots of different training schemes (**b**) The success plots of different training schemes

**Figure 6.** The precision plots and success plots of different training schemes for the proposed ID-DSN.

### *4.3. Ablation Experiment 2: The ID-CIM Mechanism for Alleviating Model Drift*

We conducted several ablation experiments to verify the effectiveness of the proposed ID-CIM mechanism for alleviating model drift. We obtained the average interval difference of the center point by calculating the coordinate difference of centroid every 10 frames in the first $N$ frames to reflect the motion state of the object and changed the value of $N$ to observe how it affects the tracking results. The relevant comparative experiments and results are shown in Table 4. The first row represents the datasets used in the training schemes, while the first column represents the assigned $N$ values. Row (a) did not adopt the ID-CIM mechanism.

When $N$ is equal to 140, the first training scheme (COCO) obtained a precision value of 0.909 and success value of 0.642, outperforming the training method without the ID-CIM mechanism, row (a) for the precision value of 0.875 and success value of 0.596. This indicates that our mechanism can prevent precision and success decline stemming from model drift. When $N$ is equal to 150 or 160, the results in row (e) or row (f) are still higher than row (a) when training with the COCO dataset, but not as effective as when $N$ is equal to 140. Therefore, when only training with the COCO dataset, the first 140 frames can

effectively reflect the motion state of the object, noting that only 14 frames were involved in the ID-CIM mechanism calculation. The fourth training scheme (DOTA+DIOR) presented a similar situation. When $N$ is equal to 140, it obtained a precision value of 0.856 and success value of 0.631, outperforming the training scheme without the ID-CIM mechanism, row (a) for the precision value of 0.854 and success value of 0.603. This once again proved that the ID-CIM mechanism can enhance the performance of the tracker.

**Table 4.** The influence of value $N$ on the experimental results of alleviating model drift.

| | | COCO | DOTA + COCO | DIOR + COCO | DOTA + DIOR | DOTA + DIOR COCO |
|---|---|---|---|---|---|---|
| NO ID-CIM (a) | **P** | 0.875 | 0.854 | 0.868 | 0.854 | 0.857 |
| | **S** | 0.596 | 0.616 | 0.646 | 0.603 | 0.632 |
| N = 120 (b) | **P** | 0.858 | 0.843 | 0.861 | 0.842 | 0.844 |
| | **S** | 0.616 | 0.605 | 0.676 | 0.594 | 0.624 |
| N = 130 (c) | **P** | 0.862 | **0.864** | **0.927** | 0.856 | **0.871** |
| | **S** | 0.625 | **0.644** | **0.694** | 0.627 | **0.668** |
| N = 140 (d) | **P** | **0.909** | 0.864 | 0.915 | **0.856** | 0.866 |
| | **S** | **0.642** | 0.645 | 0.688 | **0.631** | 0.665 |
| N = 150 (e) | **P** | 0.907 | 0.864 | 0.874 | 0.856 | 0.864 |
| | **S** | 0.638 | 0.644 | 0.68 | 0.629 | 0.666 |
| N = 160 (f) | **P** | 0.887 | 0.862 | 0.867 | 0.856 | 0.857 |
| | **S** | 0.631 | 0.642 | 0.67 | 0.627 | 0.658 |

When $N$ is equal to 130, the other three training schemes, namely combination of DOTA and COCO, combination of DIOR and COCO, and combination of DOTA, DIOR, and COCO, all met the optimal criteria. In addition, the third training scheme (DIOR+COCO) obtained the global highest values, with a precision value of 0.927 and a success value of 0.694, outperforming the training method without the ID-CIM mechanism, row (a) for the precision value of 0.868 and success value of 0.646, illustrating the effectiveness of the proposed ID-CIM for alleviating model drift. After adding the DIOR dataset for training, the global highest values in the third scheme are greater than the local optimal values in the first scheme, which once again proved the effectiveness of the method of generating satellite video object tracking training datasets based on the existing object detection datasets for remote sensing images. However, the precision decreases as $N$ increasing, because a continuously complex scenarios (e.g., the surrounding waves while tracking a ship in video 2 and the buildings while tracking a train in video 6) will affect the performance of the tracker, resulting a poor incipient tracking result.

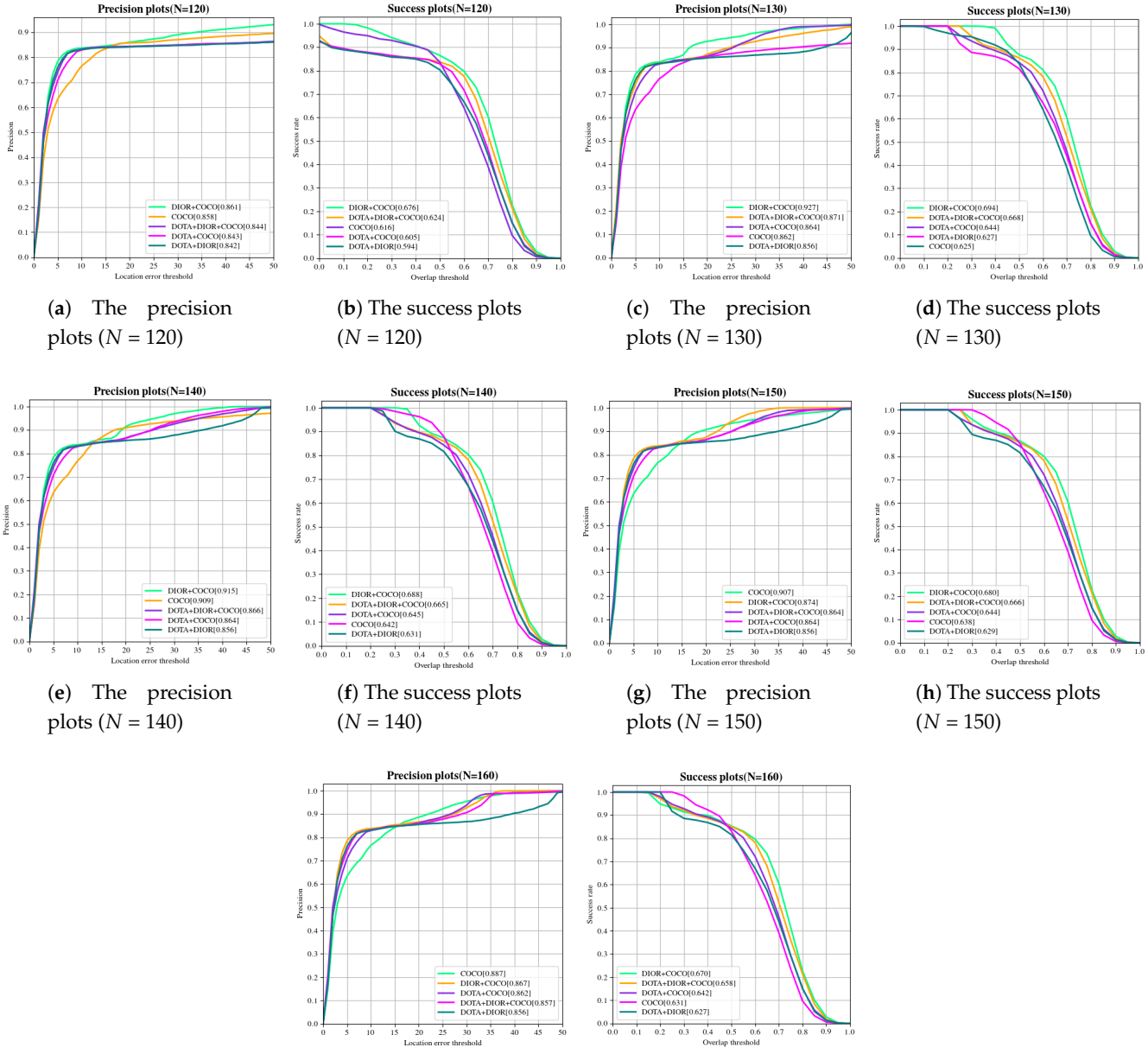
When $N$ is equal to 130, 140 and 150, the results of the second (DOTA+COCO), the fourth (DOTA+DIOR) and the fifth (DOTA+DIOR+COCO) training schemes are not significantly different, and they all tend to be the local optimal values. It illustrates that although the extensive annotations contained in the DOTA dataset cannot make the tracker obtain the global optimal values, they can improve the robustness of the tracker and help to adapt to the changes of the target scale.

We did not display the FPS criteria because ID-CIM is a post-process mechanism that computes the Siamese tracking results in numerical terms without involving image manipulation, thus not adding extra time-cost. Therefore, we still employed the corresponding FPS criteria shown in Table 3. In addition, the precision plots and success plots when $N$ is assigned different values are shown in Figure 7.

### *4.4. Ablation Experiment 3: Compared ID-DSN With the State-of-the-Art Trackers*

To verify the effectiveness and robustness of the proposed ID-DSN framework, we compared it with the other 11 state-of-the art trackers with different backbones. Our framework was not fine-tuned on six satellite videos and adopted the highest precision and success results in Table 4, it was trained with DIOR and COCO datasets and was equipped with the ID-CIM mechanism when N is equal to 130. We also showed the results

of not using ID-CIM mechanism under this training scheme. The other tested trackers were implemented in their original environments without any additions. The network architectures and the comparison results are shown in Table 5. The numbers contained in the "Backbones" column indicate the depth of the network layers. We adopted the bold **P** and **S** represent the AUC of precision plot and the AUC of success plot, and the FPS criterion to illustrate the speed of trackers. The top four results are shown in bold **red**, **orange**, **green**, and **blue**, respectively.




(**a**) The precision plots ($N$ = 120)

(**b**) The success plots ($N$ = 120)

(**c**) The precision plots ($N$ = 130)

(**d**) The success plots ($N$ = 130)

(**e**) The precision plots ($N$ = 140)

(**f**) The success plots ($N$ = 140)

(**g**) The precision plots ($N$ = 150)

(**h**) The success plots ($N$ = 150)

(**i**) The precision plots ($N$ = 160)

(**j**) The success plots ($N$ = 160)

**Figure 7.** The precision plots and success plots when $N$ is assigned different values.

**Table 5.** The network details and comparison results of the trackers.

| Trackers | Methods | Features | Backbones | Scheme for Alleviating Model Drift | P | S | FPS |
|---|---|---|---|---|---|---|---|
| KCF [12] | CF-based | HOG Features | - | - | 0.826 | 0.595 | 18.231 |
| DCFNet [44] | CF-based | CNN Features | conv1 from VGG [44] | - | 0.804 | 0.586 | 12.383 |
| ECO [15] | CF-based | CNN Features | ResNet18 with VGG-m conv1 layer [15] | - | **0.842** | **0.596** | 4.904 |
| SiamFC [17] | DL-based | CNN Features | CIResNet22 [45] | - | 0.679 | 0.518 | **61.533** |
| | | | CIResIncep22 [45] | - | 0.841 | 0.567 | **53.833** |
| | | | CIResNeXt22 [45] | - | **0.845** | **0.603** | **46.45** |
| ATOM [46] | DL-based | CNN Features | ResNet18 [20] | - | 0.598 | 0.414 | 13.953 |
| DiMP [47] | DL-based | CNN Features | ResNet18 [20] | - | 0.576 | 0.427 | 16.689 |
| | | | ResNet50 [20] | - | 0.592 | 0.45 | 14.822 |
| SiamRPN [6] | DL-based | CNN Features | CIResNet22 [45] | - | 0.72 | 0.485 | **114.867** |
| SiamRPN++ [16] | DL-based | CNN Features | ResNet50 [20] | - | 0.776 | 0.566 | 31.033 |
| Ours (without ID-CIM) | DL-based | CNN Features | ResNet50 [20] | - | **0.868** | **0.646** | 32.117 |
| Ours (with ID-CIM) | DL-based | CNN Features | ResNet50 [20] | ✓ | **0.927** | **0.694** | 32.117 |

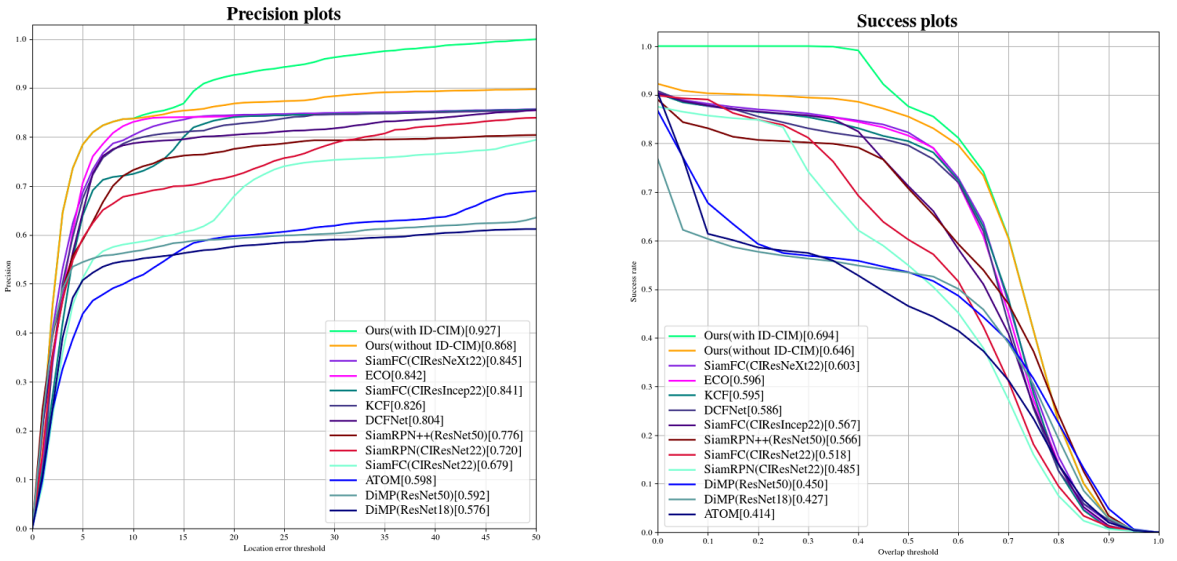
As shown in Table 5, the proposed ID-DSN (with ID-CIM mechanism) obtained the highest values, with a precision value of 0.927 and success value of 0.694. Without the ID-CIM mechanism, our ID-DSN still obtained a precision value of 0.868 and success value of 0.646, ranking the second place. In addition to our ID-DSN, other ResNet-based trackers, namely ATOM (ResNet18), DiMP (ResNet18), DiMP (ResNet50), and SiamRPN++ (ResNet50), did not obtain satisfactory results. However, some shallow network-based trackers, like DCFNet and ECO, performed well. It illustrated that, in satellite video, the target with small size cannot be effectively represented in the deep layers. The shallow layers show advantages because they focus on the appearance and shape features. The three CF-based methods also obtained high scores because the features they adopted (e.g., HOG features and conv1 features form the VGG Network) are more conducive for representing appearance and shape information, enabling the trackers to capture small objects in a satellite video.

The performance of the trackers with Cropping-Inside Residual (CIR) structure, which removes the padding in residual units [45], were more effective than the trackers based on the ResNet structure on the whole. The third-ranked SiamFC (CIResNeXt22) adopted a 22-layer backbone with no padding and obtained a precision value of 0.845 and success value of 0.603, outperforming the ResNet-based trackers (e.g., ATOM, DiMP, SiamRPN++). This indicated that padding creates position bias, and will degrade accuracy in object tracking tasks. Therefore, our future work will pay more attention to the features extracted from the shallow network and eliminate the impact of padding. At the same time, we will apply these two optimization strategies to satellite video object tracking tasks.

As to the tracking speed criterion, only the CIResNet22-based SiamRPN achieved more than 100 FPS, running at 114.867 FPS. Our ID-DSN obtained an FPS value of 32.117, which can achieve real-time processing for the satellite videos used in this paper. However, compared with CIResNet22-based SiamRPN, the significant difference in FPS value encouraged us to improve the tracker's speed in the following work. Notably, the three CF-based trackers were relatively slow in tracking speed, thus losing the huge advantage in natural scene object tracking tasks. We deemed that the overlarge remote sensing image size, complex backgrounds and abundant information make satellite video object tracking task more difficult, thus greatly reducing the processing speed. The precision plots and success plots of the ablation experiments with the state-of-the-art trackers are shown in Figure 8.

We also visualized some tracking examples and scale adaptation results in six videos, as shown in Figure 9. The yellow, red, green, and blue bounding-boxes represent the results of ground-truth, ID-DSN (with ID-CIM mechanism), SiamFC (CIResNeXt22), and ECO, respectively. From Figure 9, we can see that all tested trackers can trace the target with simple backgrounds, as in video 1, video 4, and video 5. In video 2, the SiamFC (CIResNeXt22) tracker eventually failed to track the ship due to the interference of the surrounding waves. In video 3, the ECO tracker failed to adapt to the shape changes of

the train and could only track a part of train's extents. In video 6, SiamFC (CIResNeXt22) and ECO completely failed to track the target. The tracking target in this video was a train; however, there was also building very similar to the target in the tracking area. Through the ID-CIM mechanism, our ID-DSN effectively alleviates model drift and yields effective tracking performance.




(**a**) The precision plots of the compared experiments (**b**) The success plots of the compared experiments

**Figure 8.** The precision plots and success plots of the compared experiments with the state-of-the-art trackers.

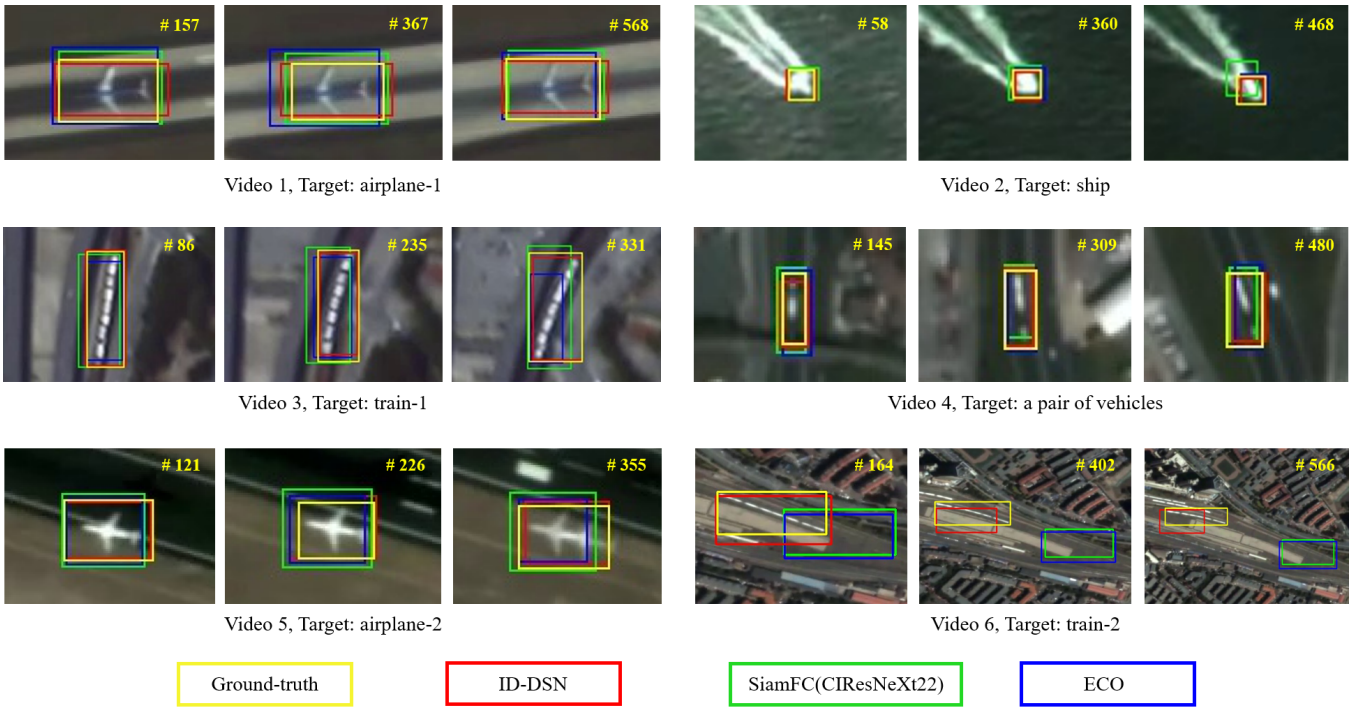



**Figure 9.** The visualized tracking results of ID-DSN, Fully-Convolutional Siamese Networks (SiamFC) (CIResNeXt22; CIR = Cropping-Inside Residual) and Efficient Convolution Operator (ECO) trackers with corresponding ground-truth. The yellow # at the top right of the image represents the video frame number.

## 5. Discussion

By analyzing the tracking results of these six satellite videos, we find that ID-DSN can fully adapt to the attributes of the small size and various scales (e.g., a ship with a size of 14 × 15 pixels and a train with a size of 154 × 63 pixels) of the ground objects. The ground targets in satellite videos are rigid objects, their shapes hardly undergo rapid shifts in single-target locations and their trajectories are approximately smooth curves in a short time, which help the tracker solve the problem of misidentification caused by similar objects. However, in the process of processing satellite video image sequences by a tracker, due to the high speed of ground target or the factors, such as poor general field illumination, a residual shadow (ghost shadow) phenomenon will occur, as shown in Figure 10.

In the above situations, ID-DSN can adaptively expand the tracking range to capture the objects more accurately. When there exists partial occlusion or the object is poorly distinguishable from the background, ID-DSN cannot completely solve the size adaptability, as shown in Figure 11.


(**a**) Video 1, Target: airplane-1


(**b**) Video 5, Target: airplane-2

**Figure 10.** ID-DSN still maintains relatively accurate tracking results in the case of residual shadow.


(**a**) Occlusion causes only a part of the object to be tracked


(**b**) Poor target-background discriminability causes loss of object size matching

**Figure 11.** Partial occlusion and poor target-background discriminability cause loss of object size matching.

In video 4, the tracking object is a pair of vehicles and partially occluded by the bridge from frame 142 to frame 191. ID-DSN can only track the unoccluded part. In video 6, ID-DSN can track the train target; however, it gradually loses the size matching degree in the later frames. The reasons for this phenomenon are that, on the one hand, the background areas on the satellite video sequence images are more complicated, and there exists backgrounds similar to the target in the search area; on the other hand, the train with a size of 154 × 63 pixels in this video is beyond the range of template frame with a size

of 127 × 127 pixels, causing the tracker to extract incomplete information. In addition, the ID-CIM mechanism focuses on exploring the movement pattern of the target's centroid but does not solve small deformation of the object size during tracking. In future works, we will expand the scopes of the template frame and search frame to adapt to the scale in the satellite video single object tracking task and simultaneously integrate the object size change and target's centroid movement pattern into the ID-CIM mechanism to achieve more accurate tracking results.

## 6. Conclusions

In this paper, an effective ID-DSN framework is proposed to solve the satellite video object tracking problem. It consists of a deep Siamese network (DSN) for obtaining the incipient tracking result and an interframe difference centroid inertia motion model (ID-CIM) for alleviating model drift. We also adopted the existing remote sensing image DOTA and DIOR object detection datasets to construct satellite video object tracking training datasets. Through the introduction of a variety of small objects, such as airplanes, ships, and vehicles, we made the tracker adapt to scale changes.

Ablation experiments were performed on six high-resolution satellite videos to verify the effectiveness of the proposed framework. We also compared the proposed ID-DSN with other 11 state-of-the-art trackers, including different networks and backbones. Our ID-DSN obtained the most effective performance in both the precision and success criterion. Our proposed tracker obtained a FPS value of 32.117 and delivered real-time processing of the satellite videos used in this paper. In the future, we will conduct research in the following areas:

1. Generalization of the tracker. The tracking performance in satellite videos from different sources under the same training mechanism will be studied.
2. Domain adaption ability of the tracker. The tracking performance on target categories that are not present in the training datasets but present in the testing datasets will be studied.
3. Improvement of network structure. We will pay more attention to features extracted from the shallow network and eliminate the impact of padding, and further will apply these two optimization strategies to satellite video single object tracking tasks. In addition, we will further improve the tracking speed to achieve high performance in both tracking accuracy and speed.

**Author Contributions:** X.Z. guided the algorithm design. K.Z. and G.C. designed the whole framework and experiments. K.Z. wrote the paper. G.C., X.T., and P.L. help organize the paper and performed the experimental analysis. H.W. provided the datasets of the experiments. X.C., Y.Z., and Z.L. contributed to the data processing and discussion of the design. K.Z. drafted the manuscript, which was revised by all authors. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by China Postdoctoral Science Foundation (No.2020M680109) and the Fundamental Research Funds for the Central Universities.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Acknowledgments:** The authors would like to thank Deimos Imaging and Chang Guang Satellite Technology Co., Ltd. for acquiring and providing the satellite videos used in this paper. The authors also would like to thank Gui-Song Xia from State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University for providing the awesome remote sensing object dataset DOTA, and Gong Cheng, Junwei Han from School of Automation, Northwestern Polytechnical University, Xi'an for providing the awesome remote sensing object dataset DIOR. The author would like to thank Supercomputing Center of Wuhan University.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Lee, K.; Hwang, J. On-Road Pedestrian Tracking Across Multiple Driving Recorders. *IEEE Trans. Multimed.* **2015**, *17*, 1429–1438. [CrossRef]
2. Liu, L.; Xing, J.; Ai, H.; Ruan, X. Hand posture recognition using finger geometric feature. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 565–568.
3. Tang, S.; Andriluka, M.; Andres, B.; Schiele, B. Multiple People Tracking by Lifted Multicut and Person Re-identification. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3701–3710. [CrossRef]
4. Zhang, G.; Vela, P.A. Good features to track for visual SLAM. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
5. Du, B.; Cai, S.; Wu, C. Object Tracking in Satellite Videos Based on a Multiframe Optical Flow Tracker. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3043–3055. [CrossRef]
6. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High Performance Visual Tracking With Siamese Region Proposal Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
7. Zhu, Z.; Wu, W.; Zou, W.; Yan, J. End-to-End Flow Correlation Tracking with Spatial-Temporal Attention. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 548–557. [CrossRef]
8. Wu, Y.; Lim, J.; Yang, M.H. Object Tracking Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1834. [CrossRef]
9. Guo, Y.; Yang, D.; Chen, Z. 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.* **2019**, *12*, 3538–3551. [CrossRef]
10. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550. [CrossRef]
11. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Discriminative Scale Space Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1561–1575. [CrossRef]
12. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 583–596. [CrossRef] [PubMed]
13. Danelljan, M.; Khan, F.S.; Felsberg, M.; van de Weijer, J. Adaptive Color Attributes for Real-Time Visual Tracking. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, 23–28 June 2014; pp. 1090–1097. [CrossRef]
14. Danelljan, M.; Robinson, A.; Khan, F.S.; Felsberg, M. Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking. *arXiv* **2016**, arXiv:1608.03773.
15. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ECO: Efficient Convolution Operators for Tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 6931–6939. [CrossRef]
16. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 4282–4291. [CrossRef]
17. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H.S. Fully-Convolutional Siamese Networks for Object Tracking. In Proceedings of the Computer Vision—ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–16 October 2016; pp. 850–865. [CrossRef]
18. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
19. Zhu, Z.; Wang, Q.; Li, B.; Wu, W.; Yan, J.; Hu, W. Distractor-aware Siamese Networks for Visual Object Tracking. *arXiv* **2018**, arXiv:1808.06048.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
21. Li, D.; Wang, M.; Dong, Z.; Shen, X.; Shi, L. Earth observation brain (EOB): an intelligent earth observation system. *Geo-Spat. Inf. Sci.* **2017**, *20*, 134–140. [CrossRef]
22. d'Angelo, P.; Kuschk, G.; Reinartz, P. Evaluation of Skybox Video and Still Image products. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2014**, *XL1*, 95–99. [CrossRef]
23. Aoran, X.; Zhongyuan, W.; Lei, W.; Yexian, R. Super-Resolution for "Jilin-1" Satellite Video Imagery via a Convolutional Network. *Sensors* **2018**, *18*, 1194.
24. Jiaqi, W.U.; Zhang, G.; Wang, T.; Jiang, Y. Satellite Video Point-target Tracking in Combination with Motion Smoothness Constraint and Grayscale Feature. *Acta Geod. Cartogr. Sin.* **2017**, *46*, 1135.
25. Hu, Z.; Yang, D.; Zhang, K.; Chen, Z. Object Tracking in Satellite Videos Based on Convolutional Regression Network With Appearance and Motion Features. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 783–793. [CrossRef]
26. Xia, G.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.J.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983. [CrossRef]

27. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
28. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014. [CrossRef]
29. Zagoruyko, S.; Komodakis, N. Learning to compare image patches via convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; pp. 4353–4361. [CrossRef]
30. Shao, J.; Du, B.; Wu, C.; Zhang, L. Tracking Objects From Satellite Videos: A Velocity Feature Based Correlation Filter. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7860–7871. [CrossRef]
31. Shao, J.; Du, B.; Wu, C.; Zhang, L. Can We Track Targets From Space? A Hybrid Kernel Correlation Filter Tracker for Satellite Video. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8719–8731. [CrossRef]
32. Han, X.; Zhong, Y.; Zhang, L. An Efficient and Robust Integrated Geospatial Object Detection Framework for High Spatial Resolution Remote Sensing Imagery. *Remote Sens.* **2017**, *9*, 666. [CrossRef]
33. Zhang, X.; Xia, G.S.; Lu, Q.; Shen, W.; Zhang, L. Visual object tracking by correlation filters and online learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 77–89. [CrossRef]
34. Guan, M.; Wen, C.; Shan, M.; Ng, C.; Zou, Y. Real-Time Event-Triggered Object Tracking in the Presence of Model Drift and Occlusion. *IEEE Trans. Ind. Electron.* **2019**, *66*, 2054–2065. [CrossRef]
35. Luo, S.; Li, B.; Yuan, X. An Anti-Drift Background-Aware Correlation Filter for Visual Tracking in Complex Scenes. *IEEE Access* **2019**, *7*, 185857–185867. [CrossRef]
36. Huang, Z.; Yu, Y.; Xu, M. Bidirectional Tracking Scheme for Visual Object Tracking Based on Recursive Orthogonal Least Squares. *IEEE Access* **2019**, *7*, 159199–159213. [CrossRef]
37. Basso, G.F.; De Amorim, T.G.S.; Brito, A.V.; Nascimento, T.P. Kalman Filter with Dynamical Setting of Optimal Process Noise Covariance. *IEEE Access* **2017**, *5*, 8385–8393. [CrossRef]
38. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, 2–4 May 2016.
39. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
40. Zhang, X.; Zhu, K.; Chen, G.; Tan, X.; Zhang, L.; Dai, F.; Liao, P.; Gong, Y. Geospatial Object Detection on High Resolution Remote Sensing Imagery Based on Double Multi-Scale Feature Pyramid Network. *Remote Sens.* **2019**, *11*, 755. [CrossRef]
41. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A.C.; Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
42. Lin, T.; Maire, M.; Belongie, S.J.; Bourdev, L.D.; Girshick, R.B.; Hays, J.; Perona, P.; Ramanan, D.; Doll'ar, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. *arXiv* **2014**, *arXiv:1405.0312*.
43. 2016 IEEE GRSS Data Fusion Contest. Available online: http://www.grss-ieee.org/community/technical-committees/data-fusion (accessed on 9 October 2019).
44. Wang, Q.; Gao, J.; Xing, J.; Zhang, M.; Hu, W. DCFNet: Discriminant Correlation Filters Network for Visual Tracking. *arXiv* **2017**, arXiv:1704.04057.
45. Zhang, Z.; Peng, H. Deeper and Wider Siamese Networks for Real-Time Visual Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
46. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ATOM: Accurate Tracking by Overlap Maximization. *arXiv* **2018**, arXiv:1811.07628.
47. Bhat, G.; Danelljan, M.; Gool, L.V.; Timofte, R. Learning Discriminative Model Prediction for Tracking. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, 27 October–2 November 2019; pp. 6181–6190. [CrossRef]