remote sensing

MDPI

Article

Single Object Tracking in Satellite Videos Based on Feature Enhancement and Multi-Level Matching Strategy

Jianwei Yang [1,2,3], Zongxu Pan [1,2,3,*], Yuhan Liu [1,2], Ben Niu [1,2,3] and Bin Lei [1,2,3]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; yangjianwei20@mails.ucas.ac.cn (J.Y.)
[2] Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Chinese Academy of Sciences, Beijing 100190, China
[3] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: zxpan@mail.ie.ac.cn

Abstract: Despite significant advancements in remote sensing object tracking (RSOT) in recent years, achieving accurate and continuous tracking of tiny-sized targets remains a challenging task due to similar object interference and other related issues. In this paper, from the perspective of feature enhancement and a better feature matching strategy, we present a tracker SiamTM specifically designed for RSOT, which is mainly based on a new target information enhancement (TIE) module and a multi-level matching strategy. First, we propose a TIE module to address the challenge of tiny object sizes in satellite videos. The proposed TIE module goes along two spatial directions to capture orientation and position-aware information, respectively, while capturing inter-channel information at the global 2D image level. The TIE module enables the network to extract discriminative features of the targets more effectively from satellite images. Furthermore, we introduce a multi-level matching (MM) module that is better suited for satellite video targets. The MM module firstly embeds the target feature map after ROI Align into each position of the search region feature map to obtain a preliminary response map. Subsequently, the preliminary response map and the template region feature map are subjected to the Depth-wise Cross Correlation operation to get a more refined response map. Through this coarse-to-fine approach, the tracker obtains a response map with a more accurate position, which lays a good foundation for the prediction operation of the subsequent sub-networks. We conducted extensive experiments on two large satellite video single-object tracking datasets: SatSOT and SV248S. Without bells and whistles, the proposed tracker SiamTM achieved competitive results on both datasets while running at real-time speed.

Keywords: satellite video; object tracking; siamese network; feature enhancement; matching strategy

Citation: Yang, J.; Pan, Z.; Liu, Y.; Niu, B.; Lei, B. Single Object Tracking in Satellite Videos Based on Feature Enhancement and Multi-Level Matching Strategy. *Remote Sens.* **2023**, *15*, 4351. https://doi.org/10.3390/rs15174351

Academic Editor: Teng Huang, Qiong Wang and Yan Pang

Received: 21 July 2023
Revised: 29 August 2023
Accepted: 1 September 2023
Published: 4 September 2023

1. Introduction

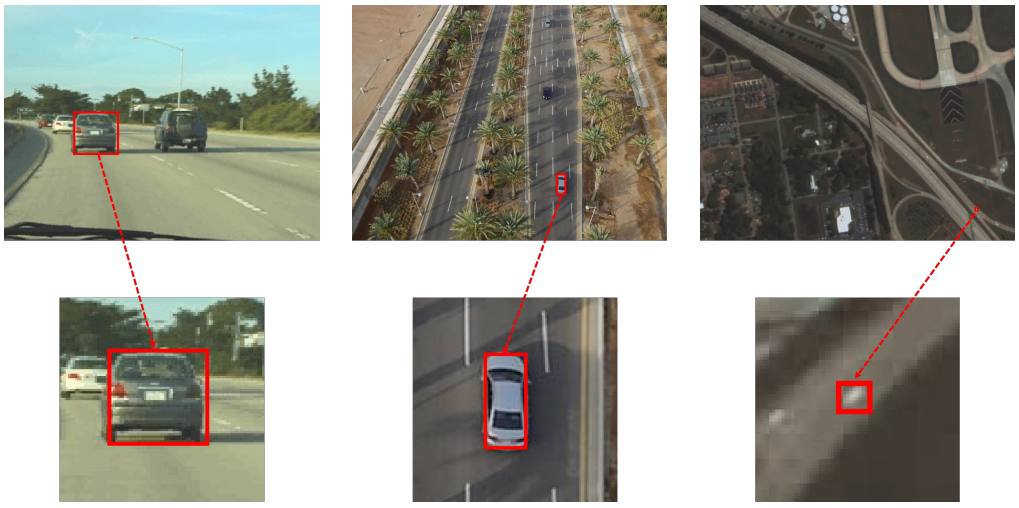
Single object tracking (SOT) refers to the task of predicting the position and size of a given target in the subsequent frames of a video, given its initial information in the first frame [1]. It is an important research area in computer vision and has been utilized in various applications, such as motion object analysis [2], automatic driving [3], and human–computer interactions [4]. As deep learning demonstrates outstanding performance in various visual fields, more and more researchers have introduced deep learning into the field of single object tracking, achieving many astonishing results. Tao et al. [5] proposed the SINT network, the first to apply the Siamese network to single-object tracking fields. SINT learns a matching function through a Siamese network using the first frame of target information as a template, and calculates the matching score between the template and all subsequent frames sampled. The highest score represents the position of the target in the current frame. SiamFC [6] abstracts the tracking process into a similarity learning problem. By learning a function $f(Z, X)$ to compare the similarity between the template

image Z and the search image X, the target's location can be predicted based on the position with the highest score on the response map. Building upon SiamFC, SiamRPN [7] introduces the RPN module from Faster R-CNN [8], which eliminates the time-consuming multi-scale testing process, further improving performance and speeding up the system. Li et al. [9] proposed the SiamRPN++ network, which introduces the ResNet-50 [10] network to improve feature extraction capability. Additionally, they proposed Depth-wise Cross Correlation to replace the previous Up-Channel Cross Correlation, drastically reducing the number of parameters and enhancing the overall stability of the training. Xu et al. [11] proposed an anchor-free tracking algorithm called SiamFC++. The algorithm enhances the original SiamFC tracker with added position regression, quality score, and multiple joint training losses, resulting in a significantly improved tracking performance. SiamCAR [12] is also a novel anchor-free fully convolutional Siamese tracking network, which decomposes visual tracking tasks into two sub-problems: pixel-wise classification and target bounding box regression, and solves end-to-end visual tracking problems in a pixel-wise manner. Guo et al. [13] proposed a simple and perceptible Siamese graph network, SiamGAT, for generic object tracking. The tracker establishes part-to-part correspondences between the target and search regions using a full bipartite graph and applies a graph attention mechanism to propagate target information from template features to search features. Yang et al. [14] proposed a dedicated Siamese network, SiamMDM, designed for single object tracking in satellite videos. This network addresses the challenge of weak features exhibited by typical targets in satellite videos by incorporating feature map fusion and introducing a dynamic template branch. Additionally, the network suggests an adaptive fusion of both motion model predictions and Siamese network predictions to alleviate issues commonly encountered in satellite videos, such as partial or full occlusions.

As Transformer [15] has shown great potential in the field of object detection, more and more researchers are trying to introduce Transformer into the field of object tracking. Due to the utilization of only spatial features in Siamese algorithms, they may not be particularly suitable for scenarios involving target disappearance or significant object variations. To address this limitation, Yan et al. [16] proposed the incorporation of a transformer architecture, which combines spatial and temporal characteristics, effectively resolving the issue of long-range interactions in sequence modeling. Chen et al. [17] pointed out that correlation operation is a simple way of fusion, and they proposed a Transformer tracking method based on the attention fusion mechanism called TransT. SparseTT [18] has designed a sparse attention mechanism that allows the network to focus on target information in the search area, and proposed a Double-Head approach to improve classification and regression accuracy. The ToMP [19] tracker also replaces traditional optimization-based model predictors with transformers. This tracker incorporates two novel encoding methods that include both target position information and range information. In addition, a parallel two-stage tracking method is proposed to decouple target localization and bounding box regression, achieving a balance between accuracy and efficiency. In order to fully leverage the capabilities of self-attention, Gui et al. [20] introduced a novel tracking framework known as MixFormer, which deviates from traditional tracking paradigms. Additionally, they proposed the MAM module, which employs attention mechanisms to perform feature extractions and feature interactions simultaneously. This renders MixFormer remarkably concise without the need for additional fusion modules. Ye et al. [21] introduced OSTrack, a concise and efficient one-stream one-stage tracking framework. This tracker leverages the prior knowledge of similarity scores obtained in the early stages and proposes an in-network early candidate elimination module, thereby reducing inference time. To address the issue of inhibited performance improvements due to independent correlation calculations in attention mechanisms, AiATrack [22] introduces an Attention in Attention (AiA) module. This module enhances appropriate correlations and suppresses erroneous correlations by seeking consensus among all relevant vectors. SwinTrack [23] employs Transformer for feature extraction and fusion, enabling full interaction between the template region and the search region for tracking. Moreover, SwinTrack extensively

investigates various strategies for feature fusion, position encoding, and training loss to enhance performance.

In recent years, with the advancement of remote sensing and image processing technology [24–26], the resolution of satellite video has been continuously improving, enabling the tracking of trains, ships, airplanes, and even ordinary cars from a remote sensing perspective. Compared with general optical object tracking, remote sensing single object tracking (RSOT) [27–30] faces several challenges that lead to significant performance degradation when directly applying deep learning-based single-object tracking methods to RSOT. Taking the common vehicle target in the tracking dataset as an example, we conducted a comparative analysis involving vehicle targets captured in natural image scenes, aerial views from unmanned aerial vehicles (UAV), and satellite perspectives. The results are depicted in Figure 1. The vehicle depicted in Figure 1a is selected from the Car24 video sequence in the OTB [31] dataset. In the displayed image frame, the vehicle target occupies 1848 pixels, accounting for 2.41% of the entire image. Despite the limited resolution of this image, the rich feature information at the rear of the vehicle provides sufficient discriminative cues for the tracker's decision-making process. The car depicted in Figure 1b is extracted from the car4 video sequence in the UAV123 [32] dataset. The car target occupies a total of 1560 pixels, which accounts for approximately 0.17% of the entire image. Despite the diminished proportion of the vehicle target within the entire image when viewed from the aerial perspective of a UAV, the contour of the vehicle target remains distinctly discernible. The car in Figure 1c is selected from the car_04 video sequence in the SatSOT [33] dataset. In complete contrast to the previous two images, the car target in this image occupies only 90 pixels, taking up only 0.01% of the whole image.



(**a**) Car in OTB100 dataset (**b**) Car in UAV123 dataset (**c**) Car in SatSOT dataset

Figure 1. Vehicle targets from different shooting angles. From left to right: car in natural image scene; car in drone aerial perspective; car in satellite capture perspective. Compared to the previous two, the proportion of the cars in the satellite images is very low, and the cars themselves are tiny in size, lacking distinct and identifiable features.

The comparison of the three images in Figure 1 reveals that the minuscule size of the vehicle targets in satellite videos results in a reduced amount of feature information. Therefore, it is essential to employ corresponding feature enhancement techniques to bolster the features of small objects in satellite videos. Additionally, it is unreasonable to rely solely on the Depth-wise Cross Correlation matching approach from the general object tracking domain, given the aforementioned scarcity of features in small objects. In summary, applying methods from the natural tracking field to the RSOT field primarily confronts two critical issues, as follows.

(1) **Weak target feature:** Due to the altitude of the satellite and the spatial resolution of satellite videos, the target in the satellite video usually occupies only a tiny percentage of the entire image [34]. Compared to generic optical targets, targets to be tracked in satellite video are generally too small in size, resulting in insufficient feature information for trackers to exploit [35].

(2) **Inappropriate matching strategy:** The common Depth-wise Cross Correlation finds the best match within the search map based on the target appearance and texture information provided by the template map. However, due to the tiny size of objects in satellite videos, which usually exist in clusters or lines in images and lack noticeable contour features, the Depth-wise Cross Correlation matching strategy is not fully applicable to RSOT.

In summary, how to effectively suppress the background information around the target and enhance the intensity of the target's own features has become an unavoidable problem in the study of target tracking in the RSOT field. At the same time, it is essential to design a matching method suitable for target tracking in satellite videos. Based on the above thinking, we have designed a tracker SiamTM specifically for satellite video target tracking based on feature enhancement and coarse-to-fine matching strategies.

The main contributions of this paper can be summarized as follows.

- Firstly, we propose a novel target information enhancement module that can capture the direction and position-aware information from both the horizontal and vertical dimensions, as well as inherent channel information from a global perspective of the image. The target information enhancement module embeds position information into channel attention to enhance the feature expression of our proposed SiamTM algorithm for small targets in satellite videos.
- Secondly, we have designed a multi-level matching module that is better suited for satellite video targets' characteristics. It combines coarse-grained semantic abstraction information with fine-grained location detail information, effectively utilizing template information to accurately locate the target in the search area, thereby improving the network's continuous tracking performance of the target in various complex scenarios.
- Finally, extensive experiments have been conducted on two large-scale satellite video single-object tracking datasets, SatSOT and SV248S. The experimental results show that the proposed SiamTM algorithm achieved state-of-the-art performance in both success and precision metrics, while having a tracking speed of 89.76 FPS, exceeding the standard of real-time tracking.

2. Related Work

In this chapter, we will introduce relevant works from two aspects. Firstly, we will introduce some typical feature enhancement methods in single-object tracking. Secondly, we will discuss the contributions of previous research in changing the matching methods between the template feature map and the search feature map.

2.1. Feature Enhancement Methods in Single Object Tracking

As one of the most crucial steps in a Siamese-based tracker, the quality of the feature maps extracted from the backbone directly influences the tracking performance and robustness. Therefore, numerous researchers have been exploring methods to obtain feature maps with richer feature information and emphasize focus on key regions. In early Siamese-based trackers (e.g., SINT [5] and SiamFC [6]), they employed a modified version of AlexNet [36] as the feature extraction network. Due to the shallow architecture of AlexNet, it lacked strong feature representation capabilities, leading to limited performance in early trackers. To further enhance network performance, subsequent works such as SiamRPN++ [9] and SiamDW [37] introduced a deeper backbone such as ResNet [10] to replace the original shallow AlexNet network. As a result of ResNet's simplicity and powerful feature extraction capabilities, it has become the default backbone extraction network in Siamese-based trackers [38]. Given a fixed backbone, exploring methods to further enhance the network's

feature representation capabilities has naturally become a topic of great interest among researchers.

Fan et al. [39] proposed a multi-stage tracking framework called Siamese Cascaded Region Proposal Network (C-RPN), which consists of a series of cascaded RPNs in the Siamese network from deep high-level to shallow low-level layers. By introducing a novel Feature Transfer Block (FTB), C-RPN effectively utilizes multi-level features for each RPN, thereby enhancing its ability to leverage both high-level semantic information and low-level spatial information. Yu et al. [40] introduced a novel Siamese attention mechanism that computes deformable self-attention features and cross-attention features. The deformable self-attention features capture rich contextual information in the spatial domain and selectively enhance the interdependencies between channel features. On the other hand, the cross-attention features aggregate and communicate abundant information between the template and search regions, thereby improving the discriminative power of the features. Cao et al. [41] proposed an Attentional Aggregation Network (AAN) that leverages attention mechanisms to enhance the expressive power of features. The AAN utilizes both a Self-Attention Aggregation Network (Self-AAN) and a Cross-Attention Aggregation Network (Cross-AAN) to aggregate attention. By incorporating self-attention and cross-attention mechanisms, the AAN effectively captures the dependencies and relationships between different regions of the input, leading to improved feature representation. Similarly, Xie et al. [42] introduced a target-dependent feature network. By incorporating deep cross-image feature correlations into multiple layers of the feature network, this novel approach effectively suppresses non-target features and possesses the ability to extract instance-variant features. Chan et al. [1] first proposed a fine feature aggregation module to integrate low-level and high-level features for a more robust feature representation. They then utilized a Compound Attention Module to independently encode the local key information of template features and the global contextual information of search features.

In addition to the aforementioned methods for enhancing the feature extraction capabilities of trackers in the general object tracking domain, researchers have also proposed targeted approaches to address the inherent challenges posed by weak target features, susceptibility to background interference, and occlusion in satellite videos.

Cao et al. [43] proposed the utilization of a three-branch Siamese network structure. In addition to using the first frame as the template branch, they also incorporated a template branch based on the previous frame. Furthermore, in order to fully leverage deep and shallow features, multiple attention mechanisms were introduced after various stages of the backbone network to achieve significant representation of object features. Song et al. [29] proposed an attention-based tracker using a Siamese network architecture. By jointly optimizing multiple attention modules, they achieved information filtering and focused on key regions, thereby enhancing robustness to weak features and background noise. Zhang et al. [44] proposed an architectural framework designed specifically for satellite video object tracking, known as ThickSiam. ThickSiam replaces the original residual modules in the backbone network with Thickened Residual Blocks, aiming to extract robust semantic features. Nie et al. [45] proposed the information compensation module called Dim-Aware to enhance the representation of object features. This module utilizes high-frequency and crucial information to enhance the localization of small objects.

2.2. Matching Methods between Template Region and Search Region

Since the target tracking model structure based on the Siamese network was determined as the basic framework, the matching method between the target region and search region has undergone multiple iterations and improvements. In the early tracker SiamFC, the Siamese network used the naive correlation matching method. SiamFC considered the feature map corresponding to the search region as a convolutional kernel and performed a correlation operation on the feature map corresponding to the template region, obtaining a single-channel response map. The target position was determined based on the location of the maximum value in the response map. Subsequently, in the SiamRPN

network, the authors proposed an Up-Channel Cross Correlation matching method to facilitate the subsequent operation of the classification and regression headers. However, Up-Channel Cross Correlation can lead to excessive parameterization. Therefore, in the SiamPRN++ network, the authors proposed the Depth-wise Cross Correlation matching method to address the abovementioned issue. It dramatically simplifies parameterization, balances the training of both branches, and stabilizes the training process, resulting in better network convergence. Alpha-Refine [46] believes that the key to improving the fine-tuning performance is to extract and maintain detailed spatial information as much as possible. Therefore, the network proposed uses a pixel-wise correlation instead of traditional correlation operations and deep correlation for high-quality feature representation, ensuring that each correlation map encodes local information of the target and avoids an extremely large correlation window. PG-Net [47] proposed that traditional similarity measurement methods introduce a lot of background noise during the tracking prediction process. Therefore, a pixel-to-global matching method is proposed to reduce the impact of noise. Zhou et al. [48] proposed a fine-grained saliency mining module to capture local saliency and a saliency-association modeling module to associate the captured salient regions and learn the global correlation between the target template and search image for state estimation. Zhang et al. [49] believe that a single matching operator is difficult to ensure stable tracking in challenging environments. Therefore, the authors proposed six different matching operators to replace the traditional cross-correlation operation. These operators are combined to explore complementary features, and a structural search method is used to select the most suitable combination of operators.

3. Proposed Approach

In this section, we introduce the proposed tracker SiamTM network in detail. Firstly, in Section 3.1, we expound the proposed SiamTM single-object tracking network from a holistic perspective. In order to extract more valuable and discriminative features from the template feature map and the search feature map, we creatively introduce a target information enhancement (TIE) module, and a detailed explanation of this module is presented in Section 3.2. Furthermore, we propose a multi-level matching (MM) module that integrates target information into the search feature map to improve tracking performance in Section 3.3.

3.1. Overall Architecture

The overall framework of the proposed SiamTM network is shown in Figure 2. The SiamTM network consists of three subnetworks: the feature extraction subnetwork, the feature enhancement and matching subnetwork, and the target prediction subnetwork. Similar to SiamCAR [12], the feature extraction network of SiamTM comprises two parts, the template branch and the search branch. The size of the template image is $3 \times 127 \times 127$, while that of the search image is $3 \times 255 \times 255$. Among them, the first dimension represents the number of channels in the image, and the last two dimensions represent the height and width of the image. The template region image and the search region image are fed into the modified ResNet50 [9] network with weight parameter sharing at the same time. After feature extraction, we obtain a $256 \times 15 \times 15$ template feature map and a $256 \times 31 \times 31$ search feature map. To enhance localization and discern foreground from background more effectively, we leverage the features extracted from the final three residual blocks of the backbone in reference to SiamCAR. Extensive literature [9,12,50] has substantiated that the joint utilization of low-level and high-level feature maps significantly contributes to improved accuracy in tracking. This is because low-level features encompass a multitude of informative cues facilitating precise positioning, such as the target's edge details and color attributes. On the other hand, high-level features encompass a greater wealth of semantic information that aids in effectively differentiating between foreground and background.

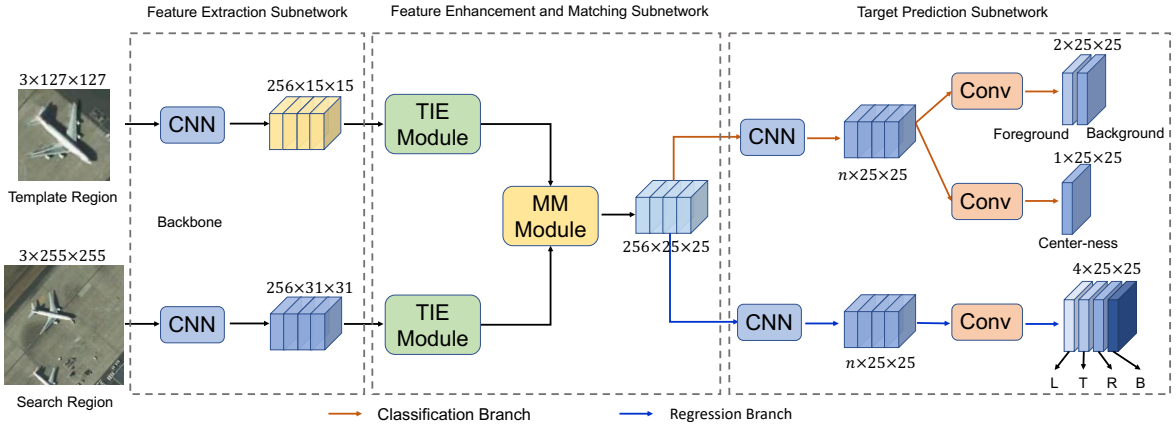


Figure 2. Architecture of the proposed SiamTM network. The SiamTM network consists of three subnetworks: the feature extraction subnetwork, the feature enhancement and matching subnetwork, and the target prediction subnetwork.

In order to extract more beneficial and distinguishable features from the original feature maps, and thereby facilitate the subsequent processing of the network, the template region feature map and the search region feature map are separately fed into the TIE module. The above operations outputs new feature maps that focus more on the information of the target itself, while maintaining the same size as the original feature maps. Compared to targets in natural images, targets in satellite videos are smaller in size and are more easily interfered by other objects in the background. In response to these characteristics of satellite video targets, a multi-level matching (MM) module was designed. The new template feature map and the search feature map are fed into the MM matching module. First, the template feature is embedded into each position of the search feature map to obtain a more accurate center position of the target, resulting in a preliminary rough response map. Then, through coarse-grained semantic abstraction information matching, the target outline is determined, resulting in a more refined final response map after matching.

Finally, the response map is fed into the target prediction subnetwork. In this subnetwork, the classification branch is used to distinguish the foreground and background in the current frame and to perform center-ness calculation, while the regression branch is used to determine the predicted bounding box. For the classification branch, it first performs a feature transformation on the response map $F_{response}$ output from the Feature Enhancement and Matching Subnetwork through a four-layer CNN structure to get the feature map F_{cls}. Each layer in the CNN structure consists of a Convolution layer, a GroupNorm layer, and a ReLU layer successively. The feature map F_{cls} is subsequently channel transformed by two independent single-layer Convolution layers to get the output feature maps $R_{cls} \in \mathbb{R}^{2\times25\times25}$ and $R_{cen} \in \mathbb{R}^{1\times25\times25}$. R_{cls} is used to differentiate the foreground from the background of the input image, whereas R_{cen} denotes the center-ness score of each position. Similar to the classification branch, the regression branch first performs a feature transformation on $F_{response}$ through an identical but independent CNN structure to obtain the feature map $F_{reg} \in \mathbb{R}^{n\times25\times25}$. Subsequently, a channel transformation is performed on F_{reg} through a Convolution layer to obtain the output of the regression branch $R_{reg} \in \mathbb{R}^{4\times25\times25}$. Each point in the feature map R_{reg} respectively represents the distance from the corresponding position to the four sides of the bounding box in the search region.

3.2. Target Information Enhancement Module

Compared to objects in natural scenarios, objects in remote sensing images are smaller and contain less feature information [51,52]. Therefore, how to effectively extract distinctive features from objects in remote sensing images has become one of the critical issues affecting subsequent tracking performance. Studies [53] on lightweight networks have shown that channel attention can significantly improve model performance. However, channel

attention often ignores the crucial positional information in visual tasks that capture target structures [54]. Therefore, it is necessary to consider how to embed positional information into channel attention. Based on the above problems and corresponding considerations, we designed a feature enhancement module to capture inter-channel information at the global 2D image level, while also capturing direction and position-aware information along the two spatial directions. The TIE module is shown in Figure 3.

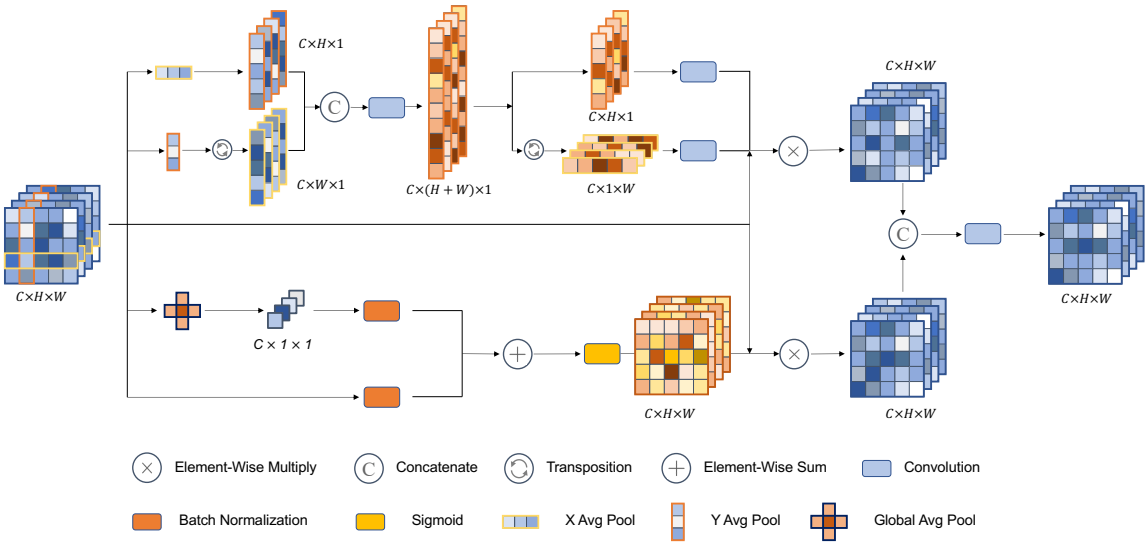


Figure 3. Architecture of the Target Information Enhancement Module. The meanings of each symbol are shown below the image.

Given a feature map $X \in \mathbb{R}^{C\times H\times W}$, first, a one-dimensional feature-encoding operation is performed on each channel using pooling kernels of size $(H, 1)$ and $(1, W)$ along the horizontal and vertical directions, respectively. The output expression of the c-th channel with a height of h is shown below:

$$e_c^h = \frac{1}{W} \sum_{0\le i\le W} x_c(h, i), \tag{1}$$

where c and h represent the channel and height of the current operation respectively, while W represents the width of the image.

Similarly, the output expression of the c-th channel with a width of w is shown as follows:

$$e_c^w = \frac{1}{H} \sum_{0\le j\le H} x_c(j, w), \tag{2}$$

where H represents the height of the image.

Then, concatenate the feature map $X_{horizontal} \in \mathbb{R}^{C\times H\times 1}$ obtained through vertical average pooling with the feature map $X_{vertical} \in \mathbb{R}^{C\times W\times 1}$ obtained through horizontal average pooling and transpose and transform it into an intermediate feature map $X_{inter} \in \mathbb{R}^{C\times (H+W)\times 1}$ through a 1×1 convolution kernel. The intermediate feature map retains spatial information from the original feature map in both vertical and horizontal directions and also captures relationships between channels.

After that, the intermediate feature map X_{inter} is divided into two separate tensors, $X'_{horizontal} \in \mathbb{R}^{C\times H\times 1}$ and $X'_{vertical} \in \mathbb{R}^{C\times 1\times W}$, along the spatial dimension. After undergoing a 1×1 convolution operation, attention weights W_h and W_w are obtained. The final output expression for the 1D part is

$$Y_1 = X \times C_1(X'_{horizontal}) \times C_1(X'_{vertical}), \tag{3}$$

where $C_1(\cdot)$ denotes the 1×1 convolution layer, $C_1(X'_{horizontal})$ and $C_1(X'_{vertical})$ denote the attention weights along the horizontal direction and the attention weights along the vertical direction, respectively.

In addition, in order to capture more distinctive target information from a global perspective and highlight more effective target features, we started from the 2D level of the image and performed global average pooling on the original feature map to obtain the feature map $X_{global} \in \mathbb{R}^{C\times1\times1}$. We also generated another intermediate feature map, whose output expression is:

$$Y_2 = \sigma(X_{global} \oplus X) \otimes X, \tag{4}$$

where $\oplus$ means the broadcasting addition, $\otimes$ means the element-wise multiplication, and σ denotes the Sigmoid operation.

Finally, we concatenate and reduce the one-dimensional feature map Y_1 and the two-dimensional feature map Y_2 obtained from feature encoding operations, and obtain the final output feature map.

3.3. Multi-Level Matching Module

Feature matching in Siamese networks refers to integrating feature maps obtained from the template branch and the search branch, calculating the similarity between each region of the template feature map and the search feature map, and finally, outputting a response map. The output response map is then sent to subsequent target prediction subnetworks for classification, regression, and other operations. As one of the most essential steps in the single-object tracking network, the quality of the feature matching directly determines the tracking performance. In existing works, since SiamRPN++ introduced Depth-wise Cross Correlation to the Siamese tracking network, most tracking networks have used this correlation operation as their feature matching subnetwork. Few networks modify the feature matching module for the characteristics of remote sensing targets. However, the Depth-wise Cross Correlation uses the target template as a spatial filter to convolve over the search area, emphasizing coarse-grained semantic abstractions such as target contours and appearance, while ignoring position information. Therefore, we propose a multi-level matching (MM) module, and the architecture of the MM module is shown in Figure 4.

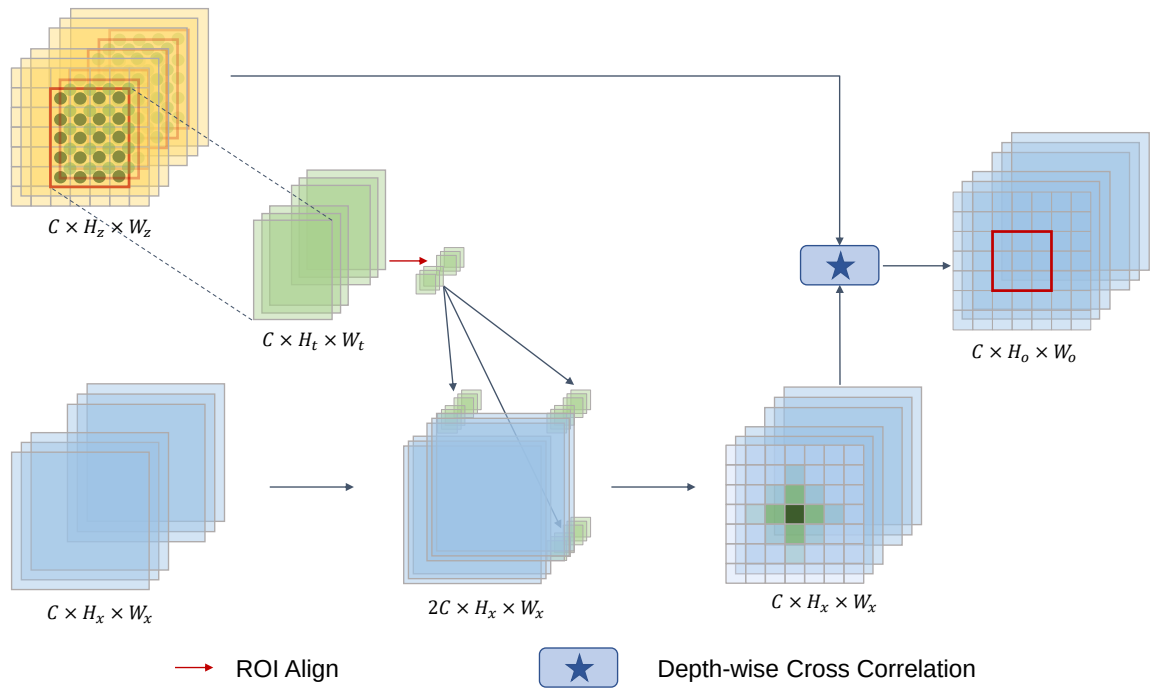


Figure 4. Architecture of the Multi-level Matching Module.

In order to achieve a more precise localization and tracking of tiny-sized objects in remote sensing videos, the proposed MM Module employs a coarse-to-fine feature matching fashion. Initially, the template feature map, denoted as X_z, is utilized to extract the segment containing solely the features of the target, followed by the application of the ROI Align operation. The resultant feature map X_t^p is then embedded into the feature map of the search region, denoted as X_s, yielding an initial response map. Subsequently, a Depth-wise Cross Correlation operation is conducted between the template feature map X_z and the preliminary response map X_s^*, yielding the ultimate response map. Within this multi-level matching mechanism, each position within the preliminary response map X_s^* assimilates information from the template features, thus rendering higher response values for pixels corresponding to the object's location within the response map. Moreover, the Depth-wise Cross Correlation operation imparts further constraints upon the object's contours. The MM Module aims to enhance the performance of the tracker by employing a two-step approach: initially coarsely localizing the target's position and subsequently refining the determination of the target's bounding box. This methodology leads to the acquisition of a more refined response map, consequently improving the tracking performance. The specific steps of the MM Module are delineated as follows.

Firstly, the feature map of the template region and the feature map of the search region, both of which have undergone feature extraction and enhancement, are, respectively, represented as $X_z \in \mathbb{R}^{C\times H_z\times W_z}$ and $X_s \in \mathbb{R}^{C\times H_x\times W_x}$. Then, according to the label information, extract the feature map that only contains the information of the target itself from the template area feature map, represented by $X_t \in \mathbb{R}^{C\times H_t\times W_t}$, ignoring the interfering background information in the template area feature map. Next, we use ROI Align to transform feature map $X_t \in \mathbb{R}^{C\times H_t\times W_t}$ containing only target information into feature map $X_t^p \in \mathbb{R}^{C\times 1\times 1}$. Feature map X_t^p will be embedded into the corresponding position of each pixel in the search area feature map, resulting in a 2C feature map $X_s^* \in \mathbb{R}^{2C\times H_x\times W_x}$. This ensures that every position in the feature map contains target features for later processing. To efficiently control the channel dimension of the feature map and avoid complex matrix operations, we use a 1×1 convolution to reduce the dimensionality of the feature map and obtain a response map with only C channels. After these operations, the network obtains a preliminary matching result, which is more focused on the localization of the target center rather than confirming the outline of the target rectangle.

Subsequently, a Depth-wise Cross Correlation operation is performed between the template area feature map and the preliminarily matched response map, which emphasizes more on locating the target bounding box, and the final response map obtained is fed into the target prediction subnetwork. Through this multi-level matching method, more precise positioning of the center point location and target rectangle box prediction can be achieved. The entire formula expression for the MM module is shown as follows.

$$F_{response} = X_z \star Concat(ROIAlign_1(X_t), X_s), \tag{5}$$

where $F_{response}$ denotes the final output response map, $\star$ denotes the Depth-wise Cross Correlation, $Concat$ denotes the Concatenation operation described above to embed X_t^p into X_s, and $ROIAlign_1$ denotes the process of converting the feature map X_t to the feature map X_t^p of size 1×1.

4. Evaluation

To test the performance of the proposed SiamTM tracker, a series of comparative and ablation experiments were carried out. The following section consists of three parts around the experimental content. Section 4.1 introduces the details of the experimental setup, including the dataset used in the experiment and evaluation methods. Section 4.2 analyzes in detail the role of each module in the proposed method through ablation experiments. In addition, in Section 4.3, we compare the proposed method with 12 other state-of-the-art object tracking algorithms to demonstrate the superiority of our algorithm.

4.1. Experimental Setup

4.1.1. Introduction to Experimental Datasets

In this paper, we conducted experiments using two large-scale satellite video single-object tracking datasets named SatSOT [33] and SV248S [55]. We performed an ablation study on the SatSOT dataset to analyze the effectiveness of each module. Furthermore, we compared the proposed SiamTM tracker with 12 other single-object tracking methods on both SatSOT and SV248S tracking datasets to validate the effectiveness and superiority of the proposed method.

SatSOT is a dedicated dataset focused on satellite video single-object tracking. The dataset consists of 105 satellite video sequences comprising a total of 27,664 frames and covering four typical moving targets in satellite videos, namely vehicles, trains, airplanes, and ships. The vehicle targets appear in 65 video sequences, train targets appear in 26 video sequences, while airplane and ship targets appear in 9 and 5 video sequences, respectively. The average length of the videos in the SatSOT dataset is 263 frames, with over 70% of the bounding box sizes in the sequence being less than 1000 pixels. The average number of pixels occupied by cars is the least, at only 112.4 pixels, while trains occupy the most, a whopping 39,566.3 pixels. To indicate the characteristics and challenges faced by each satellite video sequence and help better analyze the strengths and weaknesses of trackers, the dataset lists 11 challenge attributes in the satellite videos, and the corresponding abbreviations and definitions for each attribute are shown in Table 1.

Table 1. The 11 challenge attributes in the SatSOT dataset with their abbreviations and definitions.

Attribute	Abbreviation	Definition
Background Clutter	BC	The background has similar appearance as the target.
Illumination Variation	IV	The illumination of the target region changes significantly.
Low Quality	LQ	The image is low quality and the target is difficult to be distinguished.
Rotation	ROT	The target rotates in the video.
Partial Occlusion	POC	The target is partially occluded in the video.
Full Occlusion	FOC	The Target is fully occluded in the video.
Tiny Object	TO	At least one ground truth bounding box has less than 25 pixels.
Similar Object	SOB	There are objects of similar shape or same type around the target.
Background Jitter	BJT	Background jitter brings by the shaking of satellite camera.
Aspect Ratio Change	ARC	The ratio of the bounding-box aspect ratio of the first and the current frame is outside the range [0.5, 2].
Deformation	DEF	Non-rigid object deformation.

SV248S is a satellite video single-object tracking dataset composed of 248 video sequences. The dataset focuses on tiny objects in satellite videos. It includes four categories of objects: ships, vehicles, large vehicles, and airplanes, but no trains, which is slightly different from the SatSOT dataset. In SV248S, 202 video sequences track vehicles, 37 video sequences track large vehicles, 6 video sequences track airplanes, and the remaining 3 video sequences track ships. Compared to the SatSOT dataset, the video sequences in SV248S have a longer average length, with all sequences ranging from 500 frames to 753 frames. Additionally, the targets in the SV248S dataset are smaller, with an average of less than 65.6 pixels occupying the frames in over 81.45% of all video sequences. The largest target type airplane only occupies an average of approximately 2284.8 pixels. With sufficient video sequence numbers and a focus on tiny targets in satellite videos, the SV248S dataset is suitable for the comprehensive evaluation of a tracker's performance on satellite video targets, especially tiny and weak targets.

4.1.2. Evaluation Criteria

To quantitatively analyze the results of satellite video single-object tracking, we have adopted two standard evaluation methods: success score and precision score. In addition, we have introduced the FPS indicator to evaluate the inference speed of each tracker.

The success score represents the Intersection over Union (IoU) between the predicted bounding boxes and the ground truth bounding boxes, and is represented by the following formula.

$$S = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|}, \tag{6}$$

where B_p and B_{gt} represent the predicted bounding box and the ground truth bounding box, respectively, while $\cap$ and $\cup$ represent intersection and union. $|\cdot|$ represents the total number of pixels in the corresponding area. For a specific frame in the tracking task, if the IoU between the predicted box and the actual bounding box in the current frame exceeds a certain threshold $t_s \in (0, 1]$, it is considered that the target is successfully tracked in the current frame. Otherwise, the tracker is considered to have failed in tracking the target in the current frame.

The precision score represents the center location error (CLE) between the center point of the predicted bounding box and the center point of the ground truth bounding box, expressed as follows:

$$CLE = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}, \tag{7}$$

where (x_1, y_1) represents the center coordinate of the predicted bounding box, and (x_2, y_2) represents the center coordinate of the ground truth bounding box. For a specific frame, if the CLE between the predicted bounding box and the ground truth bounding box is less than a certain threshold $t_{CLE} \in [0, +\infty)$, it is considered that the tracker has successfully tracked the object in the current frame. Otherwise, it is considered that the tracker has failed to follow the target in the current frame accurately. In natural image single-object tracking datasets such as OTB [31] and GOT10k [56], the threshold is usually set to $t_{CLE} = 20$ pixels. However, typical targets in remote sensing images, such as vehicles, ships, and airplanes, are smaller in size. In order to accurately measure the true tracking performance of trackers, we set the threshold to $t_{CLE} = 5$ pixels in this paper.

Frames per second (FPS) measures how many frames the tracker can complete its tracking task per second. The higher the number, the more capable the tracker is of processing tasks per second, and the better the real-time performance of the tracker. We use this metric to reflect and compare the inference speeds of different trackers.

4.1.3. Implementation Details

We implemented the proposed tracking network SiamTM on the open-source deep learning library PyTorch 2.0.0, and implemented it on a 64-bit Ubuntu 20.04 workstation with 24 GB memory GeForce RTX4090 GPU. We used a modified ResNet-50 as the same in [9] as the backbone for the Siamese network feature extraction subnetwork, which was pre-trained on the ImageNet [57] dataset. The size of the initial frame template was set to 127×127 pixels, and the search region was set to 255×255 pixels. During the training phase, the batch size was set to 48, and the network was trained end-to-end for 20 epochs using the SGD method. The starting learning rate, momentum, and weight decay for SGD were set to 0.005, 0.9, and 0.0001, respectively. Consistent with SiamCAR [12], during the first 10 epochs, the parameters of the feature extraction subnetwork of the Siamese network were completely frozen, and only the following part was trained. In the subsequent 10 epochs, the last three blocks of the ResNet-50 were unfrozen for joint training. We trained our SiamTM network using data from VISO [58].

4.2. Ablation Study

The proposed SiamTM method consists of two novel components: (1) the TIE module, and (2) the MM module. To validate the effectiveness of each module in the SiamTM method and clarify the roles each module plays in improving tracking performance, we conducted a series of ablation experiments on the SatSOT dataset.

The performance of four trackers with different combinations of proposed modules is presented in Table 2. In this context, "T" abbreviates the TIE module, while "M" stands for the MM module.

Table 2. Ablation experiments of the proposed tracker on the SatSOT dataset.

Trackers	TIE Module	MM Module	Prec. (%)	Succ. (%)	Speed (FPS)
Baseline	-	-	56.4	44.6	158.41
SiamCAR+T	✓	-	60.2	46.6	113.65
SiamCAR+M	-	✓	59.5	46.7	109.17
SiamCAR+T+M	✓	✓	60.8	47.5	89.76

The performance comparison of SiamCAR, SiamCAR+T, SiamCAR+M, and the proposed SiamTM tracker can be readily observed in Table 2. Notably, the Precision and Success scores of the SiamTM tracker surpass those of SiamCAR, SiamCAR+T, and SiamCAR+M in a pronounced manner. In the case of the SiamCAR+T tracker, the utilization of the TIE module enables the tracker to effectively filter out pertinent feature information on top of the original feature extraction. Consequently, compared to the Baseline tracking network, there has been an improvement of 3.8 percentage points in precision score and 2.0 percentage points in success score. Furthermore, for the SiamCAR+M tracker, a more suitable MM module specifically designed for satellite video target tracking has been employed. The MM module allows for improved matching between the template region feature map and the search region feature map, resulting in a more accurate response map. In terms of the obtained results, compared to the baseline SiamCAR tracking network, the tracker SiamCAR+M exhibits an increase of 3.1 percentage points in precision score and 2.1 percentage points in success score. The SiamTM tracker proposed in this study introduces a TIE module to enhance the features of both the template region and the search region from both a 1D directional aspect and a 2D global aspect. Additionally, a MM module specifically tailored for satellite video targets is utilized. Effectively combining the proposed two modules can enhance the tracker's tracking performance in terms of feature extraction and feature matching dimensions. Ultimately, this approach achieves competitive results on the SatSOT dataset, with a precision score of 60.8 and a success score of 47.5. However, there is no such thing as a free lunch, and as a consequence, the inference speed of the tracker decreased by 68.64 FPS compared to the baseline, reaching 89.76 FPS. Considering that this speed is still significantly higher than the standard threshold of real-time tracking at 24 FPS, it is acceptable to trade off some inference speed in exchange for improved tracking performance.

In order to qualitatively analyze the results before and after the addition of the modules, the classification (Cls) maps of the baseline SiamCAR tracker and the proposed SiamTM tracker are visualized, and the visualization results are shown in Figure 5.

Comparing the last two columns of Figure 5, it is easy to see that the proposed SiamTM tracker is more capable of suppressing similar distractors in the search area map when distinguishing between foreground and background than the baseline SiamCAR tracker, while SiamTM is more focusing on the information of the object itself. In summary, thanks to the TIE module and MM module, the SiamTM tracker shows excellent robustness and tracking performance in the face of Tiny Object and Similar Object challenge attributes.

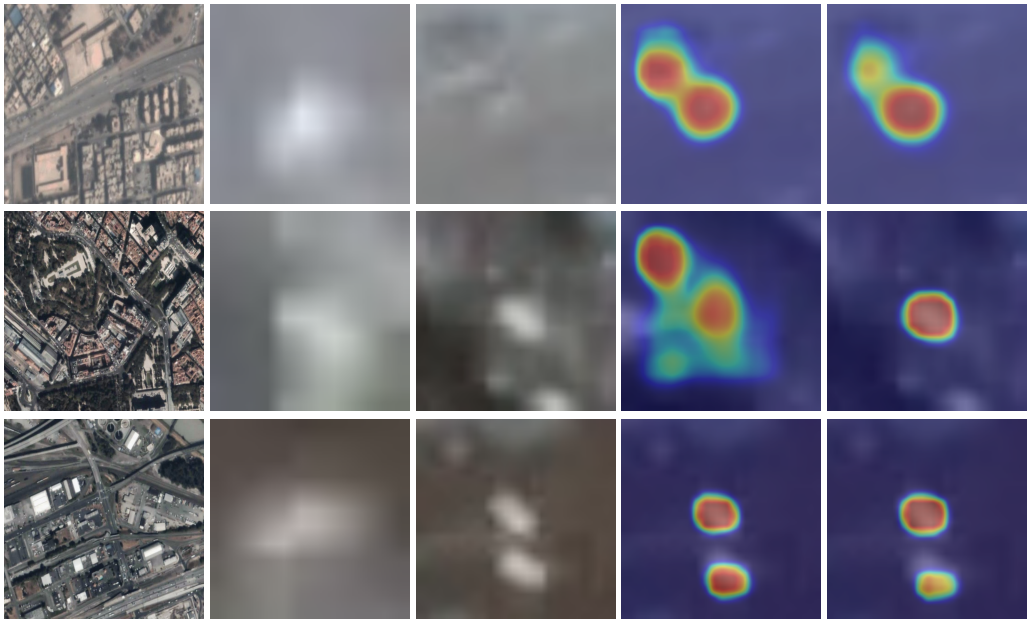


Figure 5. Comparison of classification (Cls) map between baseline and our proposed method. The first column is the original image, the second and third columns are the template region image and the search region image, respectively. The fourth and fifth columns are the Cls maps corresponding to the baseline method and the proposed method, respectively.

4.2.1. Target Information Enhancement Module

The TIE module integrates 1D information along the spatial direction dimension as well as 2D information across the image global dimension. To validate the effectiveness of fusing the 1D and 2D information, we conducted an ablation study on SatSOT dataset. We compared four different scenarios: solely employing the baseline model SiamCAR, employing solely 1D information on top of the baseline, employing solely 2D information on top of the baseline, and employing the proposed TIE module that integrates 1D and 2D information on top of the baseline. We evaluated the precision score and success score, and the experimental results are presented in Table 3.

Table 3. Performance comparison of TIE module with different structures.

Baseline	1D	1D†	2D	Prec. (%)	Succ. (%)	Speed (FPS)
✓	-	-	-	56.4	44.6	158.41
✓	✓	-	-	58.1	45.5	131.48
✓	-	✓	-	57.7	45.0	142.94
✓	-	-	✓	57.4	44.4	151.93
✓	✓	-	✓	60.2	46.6	113.65

In Table 3, "✓" indicates that the module is used in the tracker, while "-" indicates that the module is not used. From Table 3, it can be observed that the baseline SiamCAR exhibits the highest inference speed, reaching 158.41 FPS. However, it achieves the lowest levels of precision score (56.4%) and success rate score (44.6%) among the four comparisons. After incorporating only 1D or 2D information on top of the baseline, the precision score of the tracking network has shown improvements compared to the baseline. Specifically, the inclusion of 1D information improves the tracker's precision score by 1.7 percentage points compared to the baseline, while the inclusion of 2D information improves the tracker's precision score by 1.0 percentage points compared to the baseline. As for the Success metric, after adding 1D information, the tracker improves by 0.9 percentage points compared to the baseline, while the tracker with 2D information declines by 0.2 percentage points compared to the baseline, remaining almost equal to the baseline. When only adding

2D information, the network emphasizes discriminative features and the center of the target, so the tracker rises in the Precision metric. However, the Success metric places emphasis on the degree of overlap between the predicted bounding box and ground truth bounding box, which is more dependent on the shape and size of the box, and the 2D information contributes less to this. Hence, the tracker remains essentially unchanged in the Success metrics. To provide a more comprehensive demonstration of the effectiveness of our proposed module in exploiting 1D information, we conducted a comparison between two methodologies: one involves encoding spatial information along the horizontal and vertical directions using an intermediate feature map X_{inter}, followed by attention map generation (represented as "1D" in Table 3), while the other directly perform two groups of 1×1 convolution operations on $X_{horizonal}$ and $X_{vertical}$ to generate the attention weight (represented as "1D†" in Table 3). As can be seen from Table 3, compared to the baseline, directly performing two groups of 1×1 convolution operations on $X_{horizonal}$ and $X_{vertical}$ to generate the attention weight is still effective in improving the overall tracking performance of the tracker. The use of the "1D†" attention generation method resulted in a 1.3 percentage point improvement in precision score and a 0.4 percentage point improvement in success score. However, better results can be achieved by first encoding the spatial information along both the horizontal and vertical directions through an intermediate feature map X_{inter}. The use of "1D" attention generation method resulted in a 1.7 percentage point improvement in precision score and a 0.9 percentage point improvement in success score. This demonstrates that the approach of first encoding spatial information together and then generating attention maps is superior in terms of effectiveness to the approach of directly generating attention maps. However, this enhancement comes at the expense of varying degrees of reduction in inference speed. When applying the TIE module on the baseline, this module concatenates and reduces the obtained 1D and 2D information along the channel dimension, resulting in a feature map enriched with enhanced features. With the integration of the target information enhancement module, the tracking network exhibits a significant improvement of 3.8 percentage points in precision metric and 2.0 percentage points in success metric, demonstrating the effectiveness of the proposed target information enhancement module.

4.2.2. Multi-Level Matching Module

As previously mentioned, the MM Module is composed of a cascaded Concatenation matching operation and a Depth-wise Cross Correlation operation. To compare the performance difference between using the MM Module and a single matching operation, we conducted ablation experiments on the matching operation component of the network. The experimental results are presented in Table 4.

Table 4. Performance comparison of different matching methods.

Concatenation	Depth-Wise Cross Correlation	Multi-Level Matching Module	Prec. (%)	Succ. (%)	Speed (FPS)
✓	-	-	57.9	42.0	111.13
-	✓	-	56.4	44.6	158.41
-	-	✓	59.5	46.7	109.17

In Table 4, "✓" indicates the matching method used in the tracker. As indicated in Table 4, employing a MM module results in a precision score of 59.5% and success score of 46.7% for the tracking network, which stand as the highest values among the compared methods. Both the Concatenation matching operation and the Depth-wise Cross Correlation matching operation possess distinct advantages. The accuracy score attained by the Concatenation matching operation amounts to 57.9%, demonstrating a superiority of 1.5% over the Depth-wise Cross Correlation approach in this evaluation metric. This advantage primarily arises from the fact that the Concatenation matching operation embeds the infor-

mation solely contained within the target itself into each position of the search feature map. As a result, the tracking network exhibits an enhanced ability to determine the target's location. From another perspective, Depth-wise Cross Correlation employs the template feature map as a convolution kernel, conducting correlation operation on the search feature map. This process effectively constrains the target range by utilizing information such as the target's contour and texture, enabling the tracking network to predict the target's bounding box with higher success, which is not achievable by concatenation matching operation. As a result, the Depth-wise Cross Correlation operation outperforms the Concatenation matching operation by a margin of 2.6 percentage points in terms of success score. The MM module combines the Concatenation matching operation with the Depth-wise Cross Correlation operation. Firstly, it embeds the target's own information into each position of the search feature map using concatenation, resulting in an intermediate response map. Subsequently, the Depth-wise Cross Correlation operation is performed on the template feature map and the search feature map to better delineate the target's range, yielding the final feature map. The results of ablation experiments validate the effectiveness of the proposed MM Module.

4.3. Comparison with State-of-the-Art

4.3.1. Evaluated Trackers

To validate the competitiveness of our algorithm SiamTM, we compared the proposed SiamTM algorithm with 12 state-of-the-art target tracking algorithms, namely SiamFC [6], SiamRPN [7], ATOM [59], SiamRPN++ [9], SiamRPN++_lt [9], SiamMask [60], DiMP18 [61], DiMP50 [61], SiamCAR [12], PrDiMP18 [62], ToMP50 [19], and ToMP101 [19].

4.3.2. Overall Performance on SatSOT and SV248S

Tables 5 and 6 present the quantitative results of precision score, success score, and speed for the proposed SiamTM tracker and the selected trackers on the SatSOT and SV248S datasets, respectively. The precision score and success score of the SiamTM tracker outperform those of the selected comparative trackers on both two satellite video single-object tracking datasets. These findings demonstrate the effectiveness of the proposed SiamTM tracker in satellite video object tracking. In terms of speed, the top three performers on both datasets are SiamRPN, SiamFC, and ATOM. Due to the presence of the TIE module and the MM module, the speed of SiamTM on SatSOT and SV248S is 89.76 FPS and 72.50 FPS, respectively, which shows a slight decrease compared to the Baseline SiamCAR. However, it still surpasses the real-time tracking benchmark of 24 FPS. This trade-off, sacrificing some speed in exchange for higher tracking performance, is considered acceptable.

Taking the SV248S dataset as an example, we examine the performance of various trackers on this dataset. In terms of precision metric, the SiamTM tracker achieves the highest score of 75.3%, surpassing the second-ranked SiamCAR by an improvement of 5.2 percentage points and the third-ranked SiamRPN++ by an improvement of 9.7 percentage points. This remarkable performance is primarily attributed to the superior matching paradigm of the MM module. After enhancing the extracted feature maps with the TIE module, the MM module first obtains a rough intermediate response map through Concatenation. This response map emphasizes the proximity between the centers of objects rather than contour matching. Subsequently, the final response map is obtained based on Depth-wise Cross Correlation on this intermediate response map. By adopting this approach of initial center-point coarse matching followed by contour fine matching, SiamTM achieves reduced centroid error between the predicted target bounding boxes and the ground truth, as well as a higher overlap rate between contours. Furthermore, it is noteworthy that the ToMP series, which is based on Transformer for tracking, has achieved state-of-the-art performance in natural scene tracking. However, its results in satellite video object tracking are rather mediocre. Unlike objects in natural scenes, targets in satellite videos exhibit lower inter-class separability [63]. Therefore, although ToMP has been trained extensively on

natural scene images, directly applying it to RSOT is inappropriate, as experiment results have also confirmed.

Table 5. Overall performance on SatSOT. The best three performances are respectively highlighted with red, green, and blue colors. (For features, ConvFeat/CF:Convolutional Feature, TF:Transformer).

Trackers	Features	Backbone	Prec. (%)	Succ. (%)	Speed (FPS)
SiamFC	ConvFeat	AlexNet	49.8	41.3	298.3
SiamRPN	ConvFeat	AlexNet	49.4	38.5	436.6
ATOM	ConvFeat	ResNet-18	52.9	42.4	249.2
SiamRPN++	ConvFeat	ResNet-50	55.4	41.5	172.0
SiamRPN++_lt	ConvFeat	ResNet-50	52.5	38.4	148.2
SiamMask	ConvFeat	ResNet-50	55.6	40.2	173.5
DiMP18	ConvFeat	ResNet-18	52.8	42.6	115.9
DiMP50	ConvFeat	ResNet-50	51.3	42.0	92.7
SiamCAR	ConvFeat	ResNet-50	56.4	44.6	158.41
PrDiMP18	ConvFeat	ResNet-18	46.0	39.7	69.5
ToMP50	CF+TF	ResNet-50	49.2	38.8	58.1
ToMP101	CF+TF	ResNet-101	46.7	36.9	52.4
SiamTM(Ours)	ConvFeat	ResNet-50	60.8	47.5	89.76

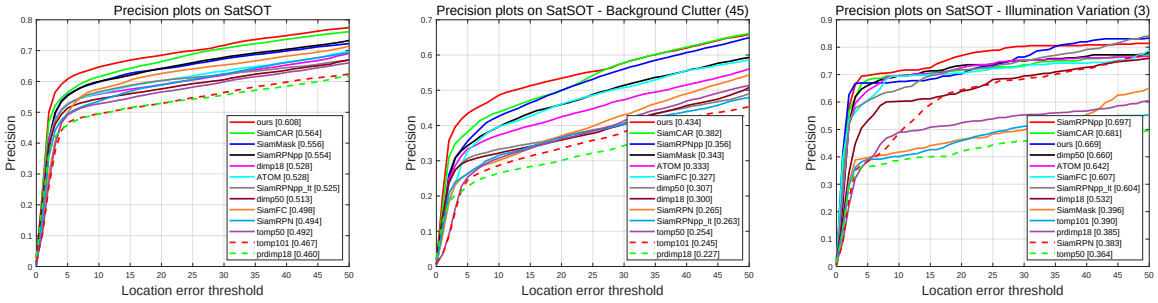
Table 6. Overall performance on SV248S. The best three performances are respectively highlighted with red, green, and blue colors. (For features, ConvFeat/CF:Convolutional Feature, TF:Transformer).

Trackers	Features	Backbone	Prec. (%)	Succ. (%)	Speed (FPS)
SiamFC	ConvFeat	AlexNet	63.4	39.4	402.0
SiamRPN	ConvFeat	AlexNet	34.4	14.7	625.6
ATOM	ConvFeat	ResNet-18	62.8	36.4	262.0
SiamRPN++	ConvFeat	ResNet-50	65.6	40.5	180.7
SiamRPN++_lt	ConvFeat	ResNet-50	56.8	21.5	167.6
SiamMask	ConvFeat	ResNet-50	55.9	21.9	179.6
DiMP18	ConvFeat	ResNet-18	58.3	35.3	111.5
DiMP50	ConvFeat	ResNet-50	61.4	36.7	82.6
SiamCAR	ConvFeat	ResNet-50	70.1	44.8	176.16
PrDiMP18	ConvFeat	ResNet-18	57.6	36.4	68.2
ToMP50	CF+TF	ResNet-50	38.7	16.5	65.4
ToMP101	CF+TF	ResNet-101	37.1	16.0	56.0
SiamTM(Ours)	ConvFeat	ResNet-50	75.3	48.7	72.50

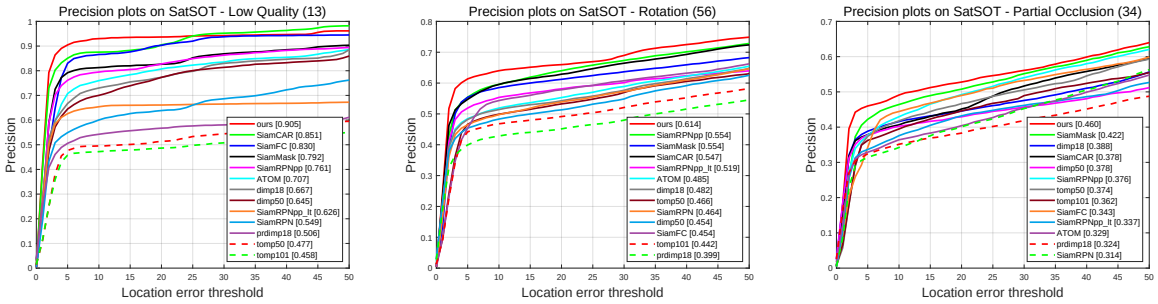
4.3.3. Attribute-Based Evaluation

To analyze the tracking performance of the trackers on different challenge attributes, we plotted the precision score plots and success score plots of the trackers on each challenge attribute of the SatSOT dataset, respectively. The plotting results are shown in Figures 6 and 7.

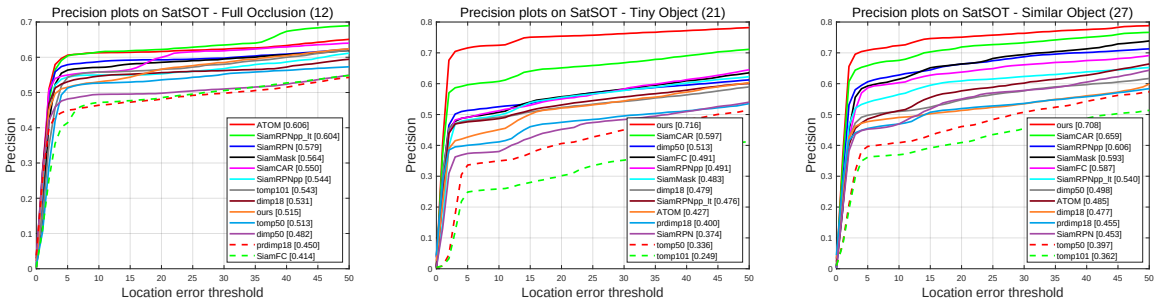
In terms of precision score, the proposed SiamTM tracker outperforms the other 12 tracking algorithms in eight challenge attributes, including the overall attribute. Specifically, SiamTM achieves the highest performance in the overall attribute, Background Clutter, Low Quality, Rotation, Partial Occlusion, Tiny Object, Similar Object, and Deformation. Notably, SiamTM delivers exceptional performance in the Tiny Object challenge attribute. It achieves a precision score of 71.6%, which is 11.9 percentage points higher than the second-ranked SiamCAR tracker in this attribute. Figure 6 clearly demonstrates that, compared to other trackers, the SiamTM tracker excels in exploring and enhancing feature information within small targets, resulting in superior tracking of such objects. This improvement is particularly significant for satellite video tracking datasets predominantly composed of small-sized objects.



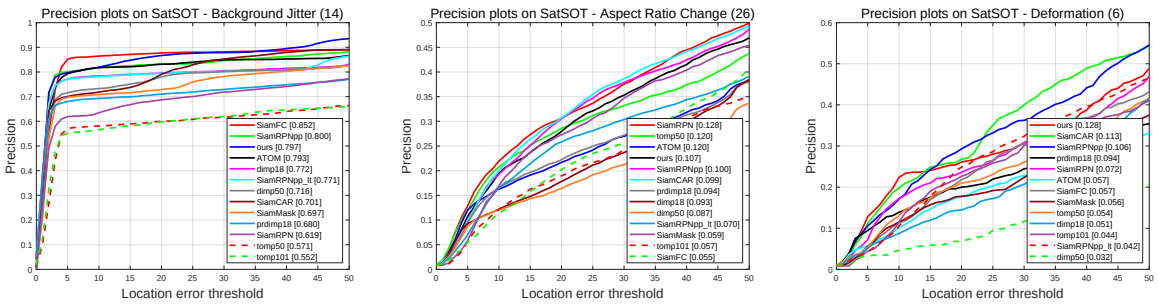
(**a**) Overall precision plot (**b**) Precision plot of BC (**c**) Precision plot of IV



(**d**) Precision plot of LQ (**e**) Precision plot of ROT (**f**) Precision plot of PO



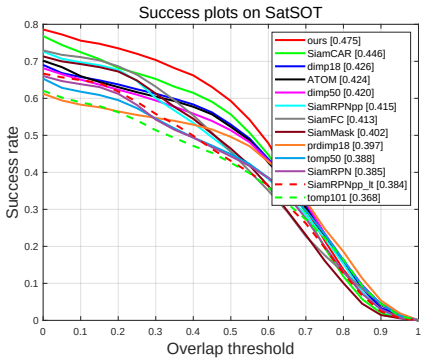
(**g**) Precision plot of FO (**h**) Precision plot of TO (**i**) Precision plot of SO



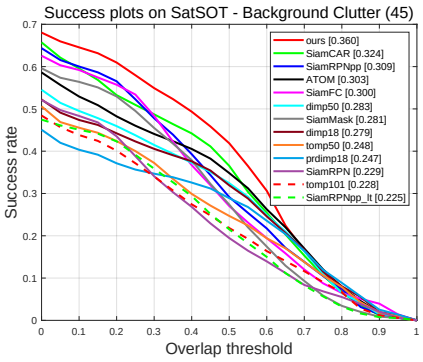
(**j**) Precision plot of BJ (**k**) Precision plot of ARC (**l**) Precision plot of DEF

Figure 6. Precision plots of 13 trackers under overall attributes and different challenge attributes.

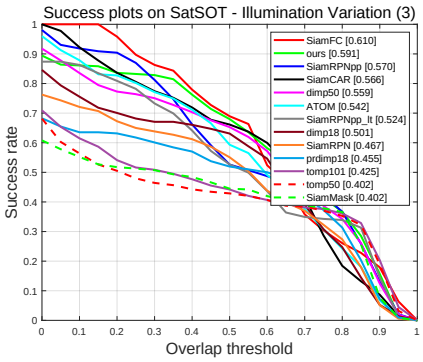
Regarding success score, compared to the other 12 tracking algorithms, the proposed SiamTM tracker achieves the highest success scores in 9 out of the 12 challenge attributes. For the remaining three challenge attributes, namely Illumination Variation, Full Occlusion, and Scale Variation, the trackers with the highest scores are SiamFC, ATOM, and ToMP50, respectively. Similar to the precision score metric, SiamTM outperforms the second-ranked tracker by 5.6 percentage points in the Tiny Object (TO) attribute. Hence, the success score plots across various challenge attributes demonstrate that the SiamTM tracker can deliver a robust tracking performance in scenes with different challenges.



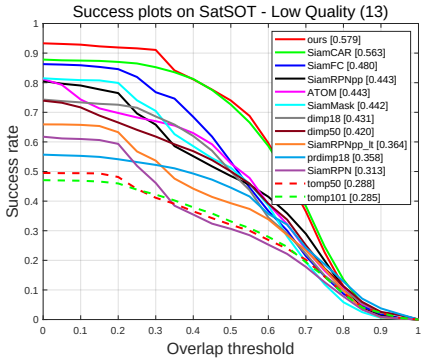
(**a**) Overall success plot



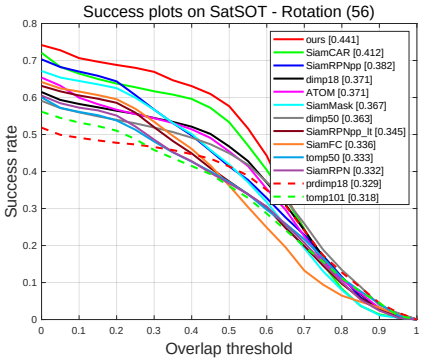
(**b**) Success plot of BC



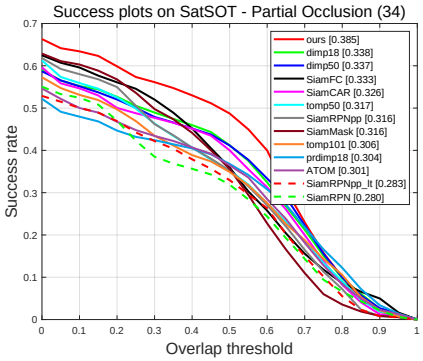
(**c**) Success plot of IV



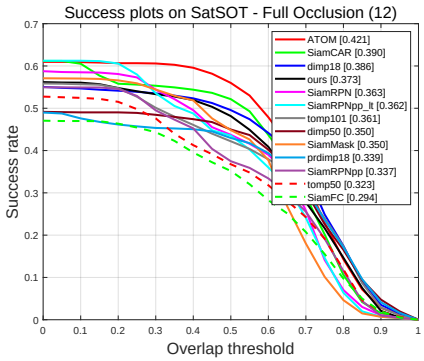
(**d**) Success plot of LQ



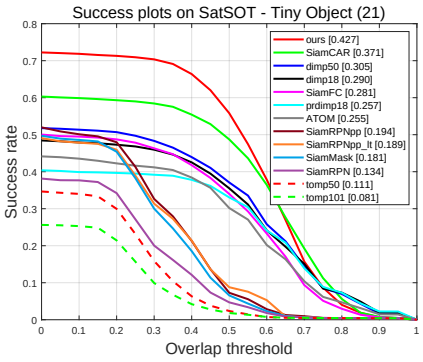
(**e**) Success plot of ROT



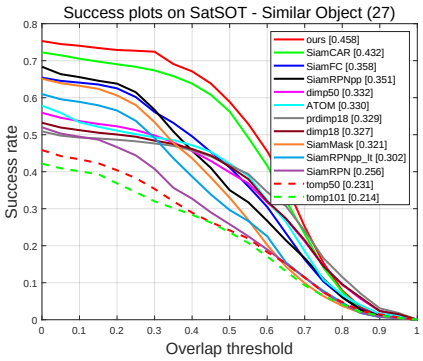
(**f**) Success plot of PO



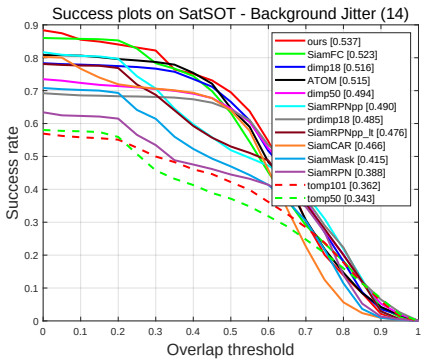
(**g**) Success plot of FO



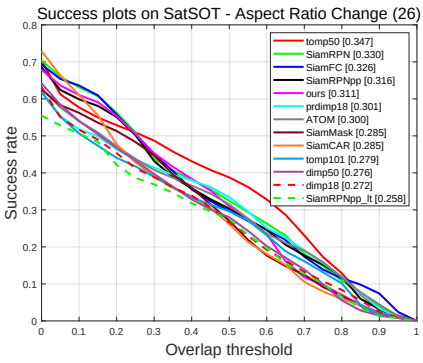
(**h**) Success plot of TO



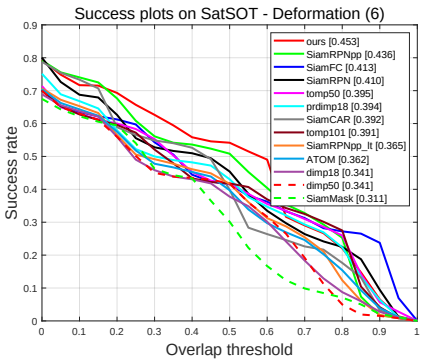
(**i**) Success plot of SO



(**j**) Success plot of BJ



(**k**) Success plot of ARC



(**l**) Success plot of DEF

Figure 7. Success plots of 13 trackers under overall attributes and different challenge attributes.

4.3.4. Visual Analysis

To qualitatively analyze the tracking performance of the proposed SiamTM tracker, we visualized its tracking results alongside the tracking results of four other trackers that

achieved the highest accuracy scores on the SatSOT dataset, as well as the ground truth. The visualization results are shown in Figure 8. In order to comprehensively analyze the tracking performance of the trackers under different challenging attributes, we selected four video sequences from SatSOT, each with distinct challenging attributes. From top to bottom in Figure 8, the sequences are Car_06, Car_07, Car_27, and Car_45.

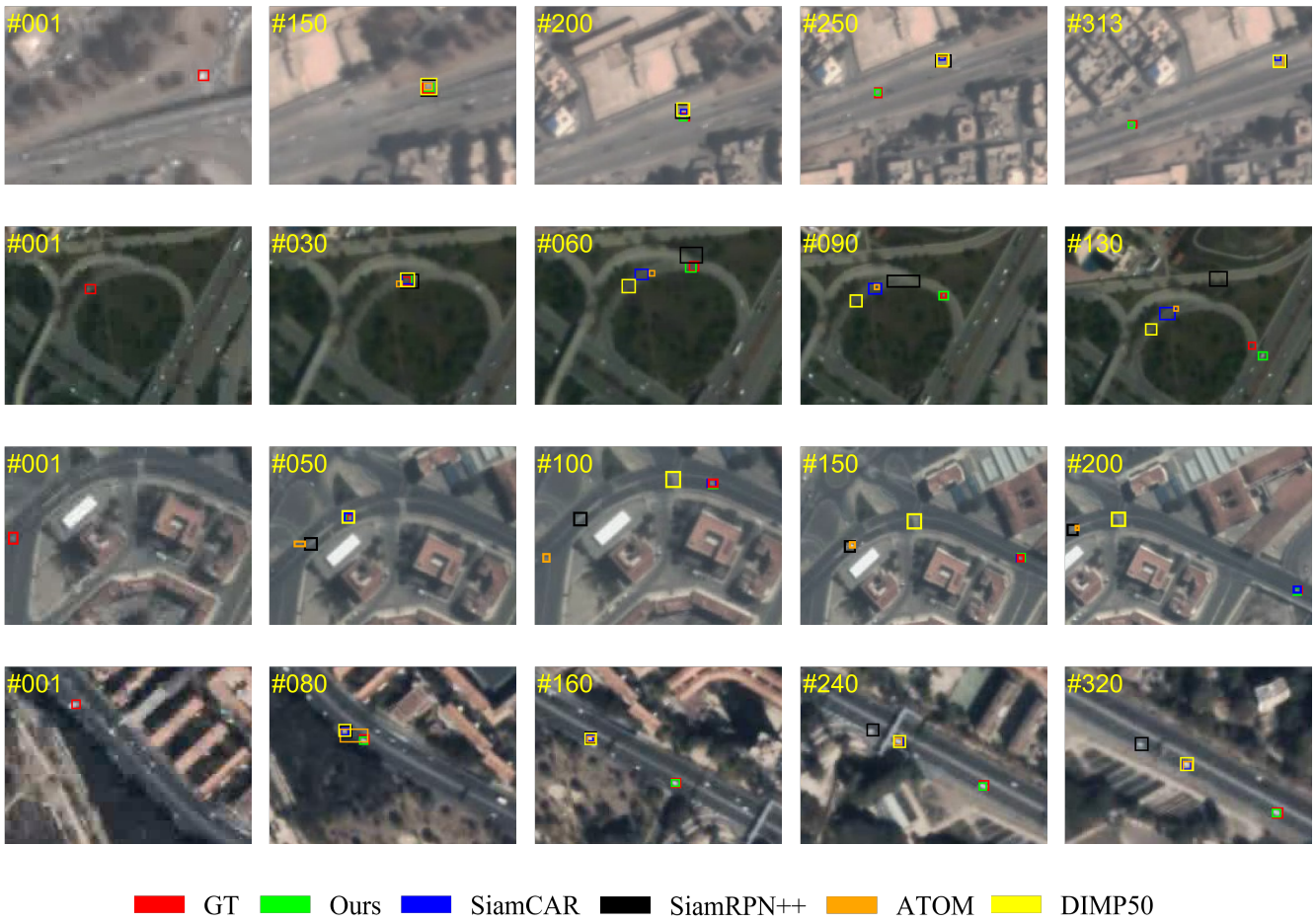


Figure 8. Qualitative comparison of five trackers and groundtruth in different videos with different challenging attributes.

The Car_06 sequence exhibits the attributes of low quality (LQ) and tiny object (TO), allowing us to visually compare the tracking performance of different trackers in terms of tracking low-quality and small objects. In the early frames of the Car_06 sequence, the target itself has relatively prominent features, enabling clear distinction between the target and its surrounding background. As a result, all selected trackers demonstrate good tracking performance in frames 1, 150, and 200, with SiamTM yielding tighter predicted bounding boxes. In the later frames of the Car_06 sequence, the discriminative features between the target and the background gradually diminish, limiting the information provided by the target itself. From frames 250 and 313, it can be observed that while other trackers experience tracking drift, the proposed SiamTM tracker maintains correct and consistent tracking. These results indicate that the SiamTM tracker exhibits robust tracking capability, allowing it to track small targets in low-quality satellite images.

The video sequences named Car_07 and Car_27 face similar types of challenges. Both sequences primarily encounter challenges such as background clutter (BC), rotation (ROT), target occlusion (TO), and low quality (LQ). Throughout the Car_07 video image sequence, the tracked target consistently resides in a complex environment with low

lighting conditions and undergoes rotation, which significantly amplifies the difficulty of tracking small targets in such environments. When observing small targets in this scenario, it was noticed at the 30th frame that the ATOM tracker exhibited tracking drift during the early stages of the video image sequence. At the 60th and 90th frames of the video, all other trackers, except for the proposed SiamTM tracker, encountered tracking drift or failures. Similarly, when examining frames 50, 100, 150, and 200 of the Car_27 image sequence, it becomes evident that SiamRPN++, ATOM, and DIMP experienced tracking failures at various stages, while only SiamCAR and the proposed SiamTM achieved consistent and accurate target tracking. These findings indicate that the SiamTM tracker, relative to other trackers, better handles the challenges of tracking small targets amidst complex backgrounds and rotation.

In the video sequence Car_45, in addition to the common challenge attributes found in the previous three video sequences (TO, BC, ROT, LQ), it also exhibits unique attributes such as partial occlusion (POC), similar object (SOB), and background jitter (BJT). Upon observing frames 80 and 160 of Car_45, it can be observed that due to the presence of similar objects in proximity, both SiamCAR and SiamRPN experienced tracking drift, while DIMP50 also encountered difficulties. However, SiamTM maintained robustness and achieved accurate tracking of the vehicle target. Even in the presence of short-term occlusion (when the vehicle passes under a bridge), SiamTM is able to predict the target's position accurately once the occlusion ends. This demonstrates that the SiamTM tracker exhibits excellent tracking performance in complex scenarios involving multiple challenges, such as partial occlusion, similar object, and background jitter.

The above visualization results demonstrate that the proposed SiamTM tracker is capable of effectively handling various challenge attributes in satellite video tracking scenarios, particularly excelling in challenges such as tiny object (TO) and low quality (LQ). This provides strong evidence of the effectiveness of SiamTM in satellite video object tracking.

4.3.5. Failure Case Analysis

While our approach has yielded highly competitive results in single-object tracking on satellite videos, inevitably, our proposed SiamFM method has encountered tracking drift or failure in some video sequences. Taking the second video sequence in Figure 8 as an example, during the initial stages of tracking, our tracker successfully distinguishes the target of interest from the surrounding background interference, thanks to the assistance provided by the TIE module and MM module. This enables the tracker to maintain continuous tracking of the target. However, at frame 130 of the video sequence, the tracker experienced tracking drift, where the predicted results deviated towards a similar distractor located very close to the target's position. The reason for the tracking drift is as follows: due to various factors such as the satellite video's resolution and background jitter, the target's distinguishable features become scarce at frame 130 of the video sequence and are almost overwhelmed by the background information. Conversely, within the same search area, a distractor appears that is very similar to the target. This distractor possesses highly similar features to the target's template in the first frame. For the proposed tracker SiamFM, which relies on the appearance features of the target for foreground-background discrimination, it is unable to discern the distinction between the target and the interfering object in such scenarios. Consequently, it erroneously assigns a higher classification confidence to the interfering object. This, in turn, leads to the tracking drifting from the correct target to the surrounding similar distractor. In the future, integrating feature enhancement with solutions such as super-resolution or motion-aware approaches holds promise for addressing tracking failure cases of this nature.

5. Conclusions

In this paper we present SiamTM, a single-object tracking algorithm based on the target information enhancement (TIE) module and the multi-level matching (MM) module,

especially for typical moving targets in satellite video. First, to address the problem of indistinct features caused by the tiny size of objects in satellite video, we propose a TIE module to extract more effective and discriminative features from the original feature maps corresponding to the images, laying a good foundation for subsequent matching and target prediction tasks. Furthermore, In order to solve the problem that the matching method of the common Siamese network algorithm cannot fully release the tracker performance and better match the template feature map and the search feature map, we propose a MM module that is more suitable for satellite video objects. We conduct comprehensive experiments on two dedicated satellite video single-object tracking datasets, namely SatSOT and SV248S. The results of ablation experiments show that the proposed two modules proposed in this article effectively improve the accuracy of single-object tracking in satellite videos. When compared with other 12 competitive methods, our proposed method SiamTM achieved state-of-the-art tracking results. The propoed SiamTM method selects and enhances discriminative features by means of attention, while acquiring response maps by moving from coarse to fine, which inevitably increases the computational complexity of the tracker. In the future, we will further improve the proposed method by reducing the computational complexity and maintaining the accuracy and robustness of the method for satellite video object tracking.

Author Contributions: Conceptualization, J.Y. and Z.P.; methodology, J.Y., Z.P and Y.L.; resources, Z.P. and Y.L.; writing—original draft preparation, J.Y. and B.N.; and supervision, Z.P. and B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Youth Innovation Promotion Association, CAS under number 2022119.

Data Availability Statement: The SatSOT dataset is available at http://www.csu.cas.cn/gb/jggk/kybm/sjlyzx/gcxx_sjj/sjj_wxxl/ (accessed on 17 July 2023). The SV248S dataset is available at https://github.com/xdai-dlgvv/SV248S (accessed on 17 July 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chan, S.; Tao, J.; Zhou, X.; Bai, C.; Zhang, X. Siamese implicit region proposal network with compound attention for visual tracking. *IEEE Trans. Image Process.* **2022**, *31*, 1882–1894. [CrossRef] [PubMed]
2. 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]
3. Lee, K.H.; Hwang, J.N. On-road pedestrian tracking across multiple driving recorders. *IEEE Trans. Multimed.* **2015**, *17*, 1429–1438. [CrossRef]
4. Dong, X.; Shen, J.; Wu, D.; Guo, K.; Jin, X.; Porikli, F. Quadruplet network with one-shot learning for fast visual object tracking. *IEEE Trans. Image Process.* **2019**, *28*, 3516–3527. [CrossRef] [PubMed]
5. Tao, R.; Gavves, E.; Smeulders, A.W. Siamese instance search for tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1420–1429. [CrossRef]
6. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In Proceedings of the Computer Vision–ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–10 and 15–16 October 2016; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 850–865. [CrossRef]
7. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8971–8980. [CrossRef]
8. Ren, S.; He, K.; Girshick, R.; 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]
9. 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/CVF Conference On Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4282–4291. [CrossRef]
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. [CrossRef]

11. Xu, Y.; Wang, Z.; Li, Z.; Yuan, Y.; Yu, G. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In Proceedings of the AAAI Conference On Artificial Intelligence, Online, 7–12 February 2020; Volume 34, pp. 12549–12556. [CrossRef]
12. Guo, D.; Wang, J.; Cui, Y.; Wang, Z.; Chen, S. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 6269–6277. [CrossRef]
13. Guo, D.; Shao, Y.; Cui, Y.; Wang, Z.; Zhang, L.; Shen, C. Graph attention tracking. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9543–9552. [CrossRef]
14. Yang, J.; Pan, Z.; Wang, Z.; Lei, B.; Hu, Y. SiamMDM: An Adaptive Fusion Network with Dynamic Template for Real-time Satellite Video Single Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–19. [CrossRef]
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. [CrossRef]
16. Yan, B.; Peng, H.; Fu, J.; Wang, D.; Lu, H. Learning spatio-temporal transformer for visual tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 10448–10457. [CrossRef]
17. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 8126–8135. [CrossRef]
18. Fu, Z.; Fu, Z.; Liu, Q.; Cai, W.; Wang, Y. SparseTT: Visual Tracking with Sparse Transformers. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence—IJCAI-22, Vienna, Austria, 23–29 July 2022; pp. 905–912. [CrossRef]
19. Mayer, C.; Danelljan, M.; Bhat, G.; Paul, M.; Paudel, D.P.; Yu, F.; Van Gool, L. Transforming model prediction for tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8731–8740. [CrossRef]
20. Cui, Y.; Jiang, C.; Wang, L.; Wu, G. Mixformer: End-to-end tracking with iterative mixed attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13608–13618. [CrossRef]
21. Ye, B.; Chang, H.; Ma, B.; Shan, S.; Chen, X. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022; pp. 341–357. [CrossRef]
22. Gao, S.; Zhou, C.; Ma, C.; Wang, X.; Yuan, J. Aiatrack: Attention in attention for transformer visual tracking. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022; pp. 146–164. [CrossRef]
23. Lin, L.; Fan, H.; Zhang, Z.; Xu, Y.; Ling, H. Swintrack: A simple and strong baseline for transformer tracking. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16743–16754. [CrossRef]
24. Wang, J.; Shao, Z.; Huang, X.; Lu, T.; Zhang, R.; Li, Y. From artifact removal to super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
25. Wang, J.; Shao, Z.; Huang, X.; Lu, T.; Zhang, R.; Cheng, G. Pan-sharpening via deep locally linear embedding residual network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
26. Wang, J.; Shao, Z.; Huang, X.; Lu, T.; Zhang, R. A Dual-Path Fusion Network for Pan-Sharpening. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
27. 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]
28. Feng, J.; Zeng, D.; Jia, X.; Zhang, X.; Li, J.; Liang, Y.; Jiao, L. Cross-frame keypoint-based and spatial motion information-guided networks for moving vehicle detection and tracking in satellite videos. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 116–130. [CrossRef]
29. Song, W.; Jiao, L.; Liu, F.; Liu, X.; Li, L.; Yang, S.; Hou, B.; Zhang, W. A joint siamese attention-aware network for vehicle object tracking in satellite videos. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
30. Lin, B.; Bai, Y.; Bai, B.; Li, Y. Robust Correlation Tracking for UAV with Feature Integration and Response Map Enhancement. *Remote Sens.* **2022**, *14*, 4073. [CrossRef]
31. Wu, Y.; Lim, J.; Yang, M.H. Online object tracking: A benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2411–2418. [CrossRef]
32. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for uav tracking. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 445–461. [CrossRef]
33. Zhao, M.; Li, S.; Xuan, S.; Kou, L.; Gong, S.; Zhou, Z. SatSOT: A benchmark dataset for satellite video single object tracking. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
34. Wu, D.; Song, H.; Fan, C. Object Tracking in Satellite Videos Based on Improved Kernel Correlation Filter Assisted by Road Information. *Remote Sens.* **2022**, *14*, 4215. [CrossRef]
35. Wu, J.; Pan, Z.; Lei, B.; Hu, Y. FSANet: Feature-and-spatial-aligned network for tiny object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
36. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]

37. Zhang, Z.; Peng, H. Deeper and wider siamese networks for real-time visual tracking. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4591–4600.
38. Javed, S.; Danelljan, M.; Khan, F.S.; Khan, M.H.; Felsberg, M.; Matas, J. Visual object tracking with discriminative filters and siamese networks: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 6552–6574. [CrossRef]
39. Fan, H.; Ling, H. Siamese cascaded region proposal networks for real-time visual tracking. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7952–7961.
40. Yu, Y.; Xiong, Y.; Huang, W.; Scott, M.R. Deformable siamese attention networks for visual object tracking. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 6728–6737.
41. Cao, Z.; Fu, C.; Ye, J.; Li, B.; Li, Y. SiamAPN++: Siamese attentional aggregation network for real-time UAV tracking. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Prague, Czech Republic, 27 October–1 October 2021; pp. 3086–3092.
42. Xie, F.; Wang, C.; Wang, G.; Cao, Y.; Yang, W.; Zeng, W. Correlation-aware deep tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8751–8760.
43. Cao, J.; Song, C.; Song, S.; Xiao, F.; Zhang, X.; Liu, Z.; Ang, M.H., Jr. Robust object tracking algorithm for autonomous vehicles in complex scenes. *Remote Sens.* **2021**, *13*, 3234. [CrossRef]
44. Zhang, X.; Zhu, K.; Chen, G.; Liao, P.; Tan, X.; Wang, T.; Li, X. High-resolution satellite video single object tracking based on thicksiam framework. *GISci. Remote Sens.* **2023**, *60*, 2163063. [CrossRef]
45. Nie, Y.; Bian, C.; Li, L. Object tracking in satellite videos based on Siamese network with multidimensional information-aware and temporal motion compensation. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
46. Yan, B.; Zhang, X.; Wang, D.; Lu, H.; Yang, X. Alpha-refine: Boosting tracking performance by precise bounding box estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5289–5298. [CrossRef]
47. Liao, B.; Wang, C.; Wang, Y.; Wang, Y.; Yin, J. Pg-net: Pixel to global matching network for visual tracking. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 429–444. [CrossRef]
48. Zhou, Z.; Pei, W.; Li, X.; Wang, H.; Zheng, F.; He, Z. Saliency-associated object tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9866–9875. [CrossRef]
49. Zhang, Z.; Liu, Y.; Wang, X.; Li, B.; Hu, W. Learn to match: Automatic matching network design for visual tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13339–13348. [CrossRef]
50. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R. Siamese box adaptive network for visual tracking. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6668–6677. [CrossRef]
51. Chen, S.; Wang, T.; Wang, H.; Wang, Y.; Hong, J.; Dong, T.; Li, Z. Vehicle tracking on satellite video based on historical model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 7784–7796. [CrossRef]
52. Wang, Z.; Yang, J.; Pan, Z.; Liu, Y.; Lei, B.; Hu, Y. APAFNet: Single-Frame Infrared Small Target Detection by Asymmetric Patch Attention Fusion. *IEEE Geosci. Remote Sens. Lett.* **2022**, *20*, 1–5. [CrossRef]
53. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141. [CrossRef]
54. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722. [CrossRef]
55. Li, Y.; Licheng, J.; Huang, Z.; Zhang, X.; Zhang, R.; Song, X.; Tian, C.; Zhang, Z.; Liu, F.; Shuyuan, Y.; et al. Deep learning-based object tracking in satellite videos: A comprehensive survey with a new dataset. *IEEE Geosci. Remote Sens. Mag.* **2022**. [CrossRef]
56. Huang, L.; Zhao, X.; Huang, K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1562–1577. [CrossRef]
57. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
58. Yin, Q.; Hu, Q.; Liu, H.; Zhang, F.; Wang, Y.; Lin, Z.; An, W.; Guo, Y. Detecting and tracking small and dense moving objects in satellite videos: A benchmark. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–18. [CrossRef]
59. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. Atom: Accurate tracking by overlap maximization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4660–4669. [CrossRef]
60. Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; Torr, P.H. Fast online object tracking and segmentation: A unifying approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1328–1338. [CrossRef]
61. Bhat, G.; Danelljan, M.; Gool, L.V.; Timofte, R. Learning discriminative model prediction for tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6182–6191. [CrossRef]

62. Danelljan, M.; Gool, L.V.; Timofte, R. Probabilistic regression for visual tracking. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 7183–7192. [CrossRef]
63. Zhang, T.; Zhang, X.; Zhu, P.; Jia, X.; Tang, X.; Jiao, L. Generalized few-shot object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 353–364. [CrossRef]