symmetry

MDPI

Article

Spectral-Spatial Feature Enhancement Algorithm for Nighttime Object Detection and Tracking

Yan Lv [1], Wei Feng [2,*], Shuo Wang [2], Gabriel Dauphin [3], Yali Zhang [2] and Mengdao Xing [4]

[1] The Optoelectronic Information Department, School of Optoelectronic Engineering, Xidian University, Xi'an 710071, China
[2] The Department of Remote Sensing Science and Technology, School of Electronic Engineering, Xidian University, Xi'an 710071, China
[3] The Laboratory of Information Processing and Transmission, L2TI, Institut Galilée, University Paris XIII, 75013 Villetaneuse, France
[4] The Academy of Advanced Interdisciplinary Research, Xidian University, Xi'an 710071, China
* Correspondence: wfeng@xidian.edu.cn

Abstract: Object detection and tracking has always been one of the important research directions in computer vision. The purpose is to determine whether the object is contained in the input image and enclose the object with a bounding box. However, most object detection and tracking methods are applied to daytime objects, and the processing of nighttime objects is imprecise. In this paper, a spectral-spatial feature enhancement algorithm for nighttime object detection and tracking is proposed, which is inspired by symmetrical neural networks. The proposed method consists of the following steps. First, preprocessing is performed on unlabeled nighttime images, including low-light enhancement, object detection, and dynamic programming. Second, object features for daytime and nighttime times are extracted and modulated with a domain-adaptive structure. Third, the Siamese network can make full use of daytime and nighttime object features, which is trained as a tracker by the above images. Fourth, the test set is subjected to feature enhancement and then input to the tracker to obtain the final detection and tracking results. The feature enhancement step includes low-light enhancement and Gabor filtering. The spatial-spectral features of the target are fully extracted in this step. The NAT2021 dataset is used in the experiments. Six methods are employed as comparisons. Multiple judgment indicators were used to analyze the research results. The experimental results show that the method achieves excellent detection and tracking performance.

Keywords: target detection; target tracking; transfer learning; nighttime

Citation: Lv, Y.; Feng, W.; Wang, S.; Dauphin, G.; Zhang, Y.; Xing, M. Spectral-Spatial Feature Enhancement Algorithm for Nighttime Object Detection and Tracking. *Symmetry* **2023**, *15*, 546. https://doi.org/10.3390/sym15020546

Academic Editors: Jeng-Shyang Pan and Sergei D. Odintsov

Received: 19 January 2023
Revised: 10 February 2023
Accepted: 15 February 2023
Published: 17 February 2023

1. Introduction

Target tracking is an application of visual algorithm research with great practical significance [1]. The long-term single-target tracking algorithm is an important direction in this field [2]. Video target detection is one of the hot spots, which is a technology for locating and classifying targets from video scenes [3]. With the popularity and continuous iteration of UAVs, the range, payload and survivability of UAVs have been further improved, which can provide longer-term surveillance and higher-precision reconnaissance [4]. A rich library of object-tracking datasets that are publicly available has been built for researchers [5–7].

Previously, the solutions of related algorithms were mainly based on traditional filtering [8–10]. In recent years, the performance of long-term single-target tracking algorithms based on deep learning is gradually catching up with traditional filtering methods [11–13]. With the rise of deep learning and the rapid improvement of GPU (Graphics Processing Unit) computing power, existing target tracking algorithms have shown high success rates and accuracy rates [14,15]. Industrial products are used in many practical scenarios and show good performance [16]. However, if the captured image dataset is directly applied to target detection, it will encounter the problem of decreased detection accuracy due to

motion blur, video out-of-focus, etc. [17]. Moreover, although these methods have shown their positive effects in related studies, they focus on tracking targets based on favorable lighting conditions [18,19]. However, nighttime and bad weather account for a considerable proportion of the whole year, and target tracking at night also has strong practical application significance. Therefore, it is necessary to further study the tracking algorithm for targets at night.

Compared with daytime-based target recognition, detecting and tracking nighttime targets is somewhat difficult. Mainly at night, the brightness of the target image is low, and the image is blurred. Detection targets are affected by low contrast, low illumination, low color saturation, and noise in the image [20,21]. These differences lead to differences in feature distributions between daytime and nighttime images.

To enhance the features of night objects, a spectral-spatial feature enhancement algorithm for nighttime object detection and tracking (SFDT) is proposed in this paper. The features of nighttime targets are enhanced by integrating the features of daytime targets. However, due to cross-domain differences, current trackers generalize poorly to nighttime scenes [22,23], which seriously hinders the performance improvement of nighttime tracking algorithms. To improve this cross-domain gap and the difficulty of performance degradation, this algorithm is devoted to solving the cross-domain object tracking problem. Then the detection and tracking model can adapt to the working environment at night under the general conditions of the day.

There are three main contributions of this work:

- A novel algorithmic framework is proposed for nighttime object detection and tracking tasks. We perform a feature-enhancing preprocessing operation on nighttime object images. Object features in images at night are more prominent to improve the accuracy of object detection.
- Introduce the concept of domain adaptation in transfer learning to create a day-night discriminator, which can align the target features of day and night and narrow the domain gap between them.
- Low-light enhancement and Gabor filtering are performed on the dataset to enhance the features, and the spectral and spatial features are fully utilized to improve the tracking performance.

The major parts of this paper are concluded as follows. Section 2 introduces the development of object detection methods and transfer learning. Section 3 describes the proposed methodology in detail. Section 4 shows the results of the experiments. Section 5 analyzes the advantages and disadvantages of the algorithm. Section 6 draws the conclusion.

2. Related Work

2.1. Target Detection

Object detection has been a research hotspot in the field of intelligent algorithms [24]. According to the contents of different studies, we can divide the target detection field into the target detection of single-frame static images and the target detection with video timing information [25]. From the perspective of algorithm development, the target detection field can be divided into target detection based on traditional methods and target detection based on deep learning [26]. The following focuses on object detection based on deep learning algorithms.

In the Large-Scale Visual Recognition Challenge (ILSVRC) [27] competition in 2012, Krizhevsky et al. [28] built a deeper convolutional neural network, and the image classification accuracy came out on top in the competition. Subsequently, the method of selecting regions by sliding window, which is commonly used in algorithms, is gradually replaced by advanced candidate region generation methods, such as constrained parametric min-cuts (CPMC) [29], Selective Search [30] and Multiscale combinatorial grouping (MCG) [31]. The target detection algorithm has entered a new stage. Compared with the traditional sliding window method, the emerging candidate region generation method can make full use of the texture, color and other information of the image, and merge regions of different scales

by calculating the similarity [32]. While the number of candidate regions is reduced, the quality of the candidate regions is improved. Ross Girshick et al. [33] combined the Region Proposal with the convolutional neural network (CNN)[34] and proposed a region-based convolutional neural network Region-Convolutional Neural Network (R-CNN) detection algorithm. The algorithm achieves leading results on the ImageNet dataset. However, this algorithm has the disadvantage of low detection speed. In order to improve the detection efficiency, He et al. [35] proposed the Spatial Pyramid Pooling Network (SPP-Net). The algorithm performs feature extraction on the data only once. Ross Girshick et al. [36] proposed Fast R-CNN. In this algorithm, a Region of Interest (RoI) pooling layer is adopted to implement feature mapping. The multi-task loss function is also introduced, which further improves the accuracy of target detection. Ren et al. [37] proposed a Faster R-CNN on the basis of Fast R-CNN. The innovation lies in the use of a Region Proposal Network (RPN).

The above target detection method based on candidate regions is called a two-stage detection method. Joseph Redmom et al. [38] proposed the You Only Look Once (YOLO) algorithm, which is called a one-stage detection method because the candidate region does not need to be generated. Instead, the classification and regression of candidate boxes are performed directly on the original image [39]. Target detection using YOLO is a current hotspot. Papers [40,41] all employed YOLO to track the targets.

A Siamese network is a special type of neural network and is one of the simplest and most commonly used one-shot learning algorithms [42]. Compared with CNN networks, Siamese networks require less training data. Therefore, Siamese networks can perform better in detection if there is insufficient prior data. This method was proposed by Sint [43] and SiamFC [44]. B.Li et al. [37] added the Region Proposal Network (RPN) to the Siamese framework for the purpose of object detection. A deeper backbone network and feature aggregation structure are applied by the SiamRPN++ algorithm [45], and the tracking accuracy is further improved. To alleviate the problem of algorithmic complexity caused by the introduced hyperparameters, the AnclFree method [46–48] computes a per-pixel regression to predict the offset at each pixel. To further improve detection and tracking, transformer is introduced into the Siamese framework [48,49] to model global information.

2.2. Domain Adaptation

Domain adaptation is a branch of Transfer Learning. Introducing domain adaptation into the algorithm can achieve the purpose of narrowing the differences in the characteristics of different domains [50]. The principle is to map data features from different domains (such as two different datasets) to the same feature space. Then data from other domains (source domains) are leveraged to enhance training in the target domain. Usually, there is rich labeled prior information in the source domain, which contains samples different from the test samples; while the samples with the same nature as the test samples are in the target domain, which does not contain any information or contains a small amount of prior information [51]. The source domain and the target domain often belong to the same type of domain, but the properties of the samples in them are different. Domain adaptation is widely used in the field of object classification. Y. Chen et al. [52] constructed a domain-adaptive object detection algorithm that improved the domain movement problem. Yu et al. [53] proposed a new model Faster Multi-Domain Net. Domain adaptation components are designed to obtain more general characteristics. An adaptive spatial pyramid pooling layer is implemented to reduce model complexity and speed up tracking. The framework proposed by Jihoon Moon et al. [54] employs an Incremental Mean Subspace Computation (ICMS) technique to solve the Online Unsupervised Domain Adaptation (OUDA) problem. Debaditya Acharya et al. [55] combined hierarchical edge maps and semantic segmentation for domain adaptation to achieve the purpose of single-image localization of 3D models.

3. Proposed Method

Spectral characteristics state that any object in nature has its own law of electromagnetic radiation. Spatial features are texture features obtained by Gabor filtering [56]. We use the method of spectral-spatial features to further improve the accuracy of detection and tracking. The SFDT algorithm is dedicated to enhancing the features of nighttime targets for easy extraction. Daytime target features and nighttime target features are modulated to enhance the detectability of nighttime features. The method consists of four steps. First, the target domain images without labels are preprocessed, which includes low-light enhancement, object detection and dynamic programming. Then the patches at nighttime and the target images at daytime pairs are obtained. Secondly, a feature extractor is applied to obtain daytime and nighttime features. In this step, adversarial learning is introduced to reduce the difference between them. Third, the Siamese network is trained on the feature sets, as the object detector and tracker (SDT), which is inspired by symmetry theory. Fourth, the test data are preprocessed (including low-light enhancement and Gabor filtering) and used as the input of SDT to obtain the final detection and tracking results. To clearly illustrate the algorithm flow, the overall description of the SFDT is shown in Algorithm 1

Algorithm 1 Spectral-spatial feature enhancement algorithm for nighttime object detection and tracking

Input: Target dataset (datasets for the target domain, the source domain, and the test)
1. The target domain data is preprocessed, including low-light enhancement, object detection, and dynamic programming with (1)–(3).
2. Source and target domain data features are extracted with (4)–(5)
3. Source and target domain data features are modulated by domain adaptive structure with (6)–(9).
4. Siamese network is trained to get the tracker head, and the loss function is obtained.
5. The test data is preprocessed, including low-light enhancement and Gabor filter with (1)–(3) and (11)–(12).
6. The feature-enhanced test data is detected and tracked by SDT with (13).
Output: Object detection maps and location data

3.1. Preprocessing

3.1.1. Low Light Enhancement

Images acquired at night generally suffer from low light and low visibility. To avoid the phenomenon that the images cannot provide useful information for subsequent object detection, we apply the low-light enhancement method to "illuminate" the objects [57]. This method achieves illumination enhancement by adjusting the pixmap curve.

First, the pixel values of all points are normalized. Then the red, green and blue three-channel pixels of the image are mapped to the enhancement curve (*EC*) as follows:

$$EC(I(x);\gamma) = I(x) + \gamma I(x)(1 - I(x)) \tag{1}$$

where x represents the coordinate point of a pixel, and *I(X)* is the pixel value of the point; γ is a variable with a value range of $[-1, 1]$, which is used to control the exposure level of the image and adjust the tone; *EC(I(x);*γ*)* is the output enhancement pixel.

Second, to deal with challenging low-light conditions, the higher-order *EC* curve with an iterative function is defined as follows:

$$EC_k(x) = EC_{k-1}(x) + \gamma_n EC_{k-1}(x)(1 - EC_{k-1}(x)) \tag{2}$$

where k is the order, which is the number of iterations to control the curvature. The exact value of k is 8 in the research.

Finally, to obtain the mapping curve with optimal effect, a deep learning-based parameter estimation network is proposed. An image is input, and the network will generate a pixel-based curve parameter map as output.

3.1.2. Video Object Detection

An algorithm based on salient object detection is applied to detect the aforementioned low-light enhanced dataset. In order to take full advantage of the context-sensitive information, every three consecutive frames are grouped as input. At the same time, dense salient features can be obtained. ResNet-101 [58], as a feature extractor, can generate feature maps with four spatial resolutions. To preserve the spatial structure, spatial pyramid layers are introduced to replace the last two layers of the network. Finally, the decoder aggregates temporal and spatial features into spatiotemporal features to generate salient predictions.

3.1.3. Dynamic Programming

To generate motion sequences of objects in videos, a dynamic programming method is introduced [59]. We constructed bounding rectangles for the above targets as candidate boxes. Based on the fact that the trajectories of moving objects in adjacent frames are smooth, the key of the proposed method is to encourage smooth trajectories. Therefore, the reward of the trajectory between candidate boxes in two adjacent frames is computed to remove unreliable boxes. The reward for dynamic programming $\mathcal{R}$ is defined as follows:

$$\mathcal{R} = (\frac{l_{i,m} - l_{j,n}}{w_{j,n}})^2 + (\frac{t_{i,m} - t_{j,n}}{h_{j,n}})^2 + (log(\frac{w_{i,m}}{w_{j,n}}))^2 + (log(\frac{h_{i,m}}{h_{j,n}}))^2 \quad (3)$$

where $[l_{i,m}, t_{i,m}, w_{i,m}, w_{i,m}]$ and $[l_{j,n}, t_{j,n}, w_{j,n}, w_{j,n}]$ represent the information of the two boxes; i and j represent the labels of the i-th and j-th frames; m and n indicate the box indexes; l is the left coordinate, t is the top coordinate; w and h denote the width and height of the box.

3.2. Dat-Net

To modulate the features of the source domain and target domain for good detection of nighttime objects, we apply an adaptive-based nighttime tracking network (DAT-Net) [60]. It contains the following modules.

3.2.1. Feature Extractor

The applied Siamese network contains two sub-networks and two branches. One branch is the template branch, and the input is the cropped image T of the previous frame; the other is the search branch, and the input is the data S of the current frame. Then feature maps $\varpi(T)$ and $\varpi(S)$ are generated, represented as follows [60]:

$$\varpi(T) = \sum_{k=p}^{N} F_k(T) \quad (4)$$

$$\varpi(S) = \sum_{k=p}^{N} F_k(S) \quad (5)$$

where F_k indicates the extracted features from the k-th block of N backbones in all. The data used by the tracker are usually the features of the last p to N blocks. This backbone in the Siamese Neural Network is recorded as FTS-101 [61].

3.2.2. Transformer Adaptive Structure

Aiming at the difficulty that object features are inconspicuous and hard to be detected at nighttime, we introduce a domain-adaptive structure in transfer learning. By modulating the characteristics of the daytime target and the nighttime target, nighttime tracking will be more effective. Specifically, since the transformer structure is good at learning long-range

target relations [62], an adaptive layer based on the transformer structure is constructed in this paper.

The transformer applies global information, the timing information in the video cannot be reflected. Therefore, we add the sequence code P to the features obtained above. The transformer structure consists of an encoder and a decoder; these two parts have multiple self-attention structures. Multihead self-attention (MSA) is conducted as:

$$\widehat{\omega(T)}' = MSA(P + \omega(T)) + P + \omega(T) \tag{6}$$

$$\widehat{\omega(S)}' = MSA(P + \omega(S)) + P + \omega(S) \tag{7}$$

$$\widehat{\omega(T)} = LN(FFN(Mod(LN(\widehat{\omega(T)}'))) + \widehat{\omega(T)}') \tag{8}$$

$$\widehat{\omega(S)} = LN(FFN(Mod(LN(\widehat{\omega(S)}'))) + \widehat{\omega(S)}') \tag{9}$$

where $\widehat{\omega(T)}'$ and $\widehat{\omega(S)}'$ are the intermediate variables, FFN represents the feedforward network, and LN represents the normalization operation. Mod is a modulation layer in [18].

3.2.3. Tracker Head

The above-mentioned feature sets $\widehat{\omega(T)}$ and $\widehat{\omega(S)}$ modulated by the transformer bridge layer are subjected to a cross-correlation operation. Then the similarity map is generated as input to the tracker head to predict object locations.

3.2.4. Feature Discrimination Structure

For the above-modulated feature set $\widehat{\omega(T)}$, we need to correctly distinguish the features from daytime or nighttime to achieve the alignment of day/night features. The day-night feature discriminator D consists of two transformer layers and a gradient reversal layer (GRL) [63]. $\widehat{\omega(T)}$ is first performed as a Softmax operation and then put into the GRL layer; finally, it goes through two transformer layers. In this way, the function of distinguishing features is realized and the final output is the predicted class c.

3.2.5. Loss Function

The loss of the entire network consists of three parts: classification loss ℓ_{clc} [64], regression loss ℓ_{reg} and domain adaptive loss ℓ_{adp}. The application of classification loss and regression loss functions guarantees the superior tracking performance of the tracker; here the tracking loss is consistent with the baseline tracker. The adaptive loss function is used to ensure the modulation effect of source and target domain features. The total loss function [60] is:

$$\ell = \ell_{clc} + \ell_{reg} + \mu\ell_{adp} \tag{10}$$

where μ is a weighting coefficient with a value of 0.01.

3.2.6. Gabor Filter

Increasing the number of features in the image is beneficial for better detection of the target. In this algorithm, we introduce texture features as representatives of spatial features. Therefore, the combination of spatial features and spectral features is used as the basis for detecting targets, which will help improve the accuracy of detection. While local features can well represent local changes and capture the existence of small objects better than global features, the Gabor wavelet feature based on biological characteristics is one of the more successful local features [56,65]. The two-dimensional Gabor filter is a complex exponential function modulated by a Gaussian function with strong spatial position and direction selectivity, which can effectively extract the direction features of multiple scales of

the image. Therefore, the Gabor filter has significant advantages in extracting the texture and orientation of local features of the image and plays a vital role in the subsequent analysis, processing and identification. To enhance data features, the data are preprocessed by low-light enhancement and Gabor filter based on paper [60]. Gabor filters are adopted to extract texture features from low-light enhanced test data. However, too many features may cause the opposite effect and increase the complexity of the algorithm. In this paper, only one texture feature is adopted. The size of the Gabor filter is set to 9, and the angle is set to $\frac{\pi}{4}$ clockwise.

$$Test' = EC(Test) \tag{11}$$

$$TF = Gabor(Test') \tag{12}$$

$$Map = SDT(TF) \tag{13}$$

where *Test* is the original test set. *Test*′ is the dataset after low-light enhancement. *Gabor* means Gabor filter. *Map* is the final result.

The effect of the images enhanced by low light and the Gabor filter is shown in Figure 1.

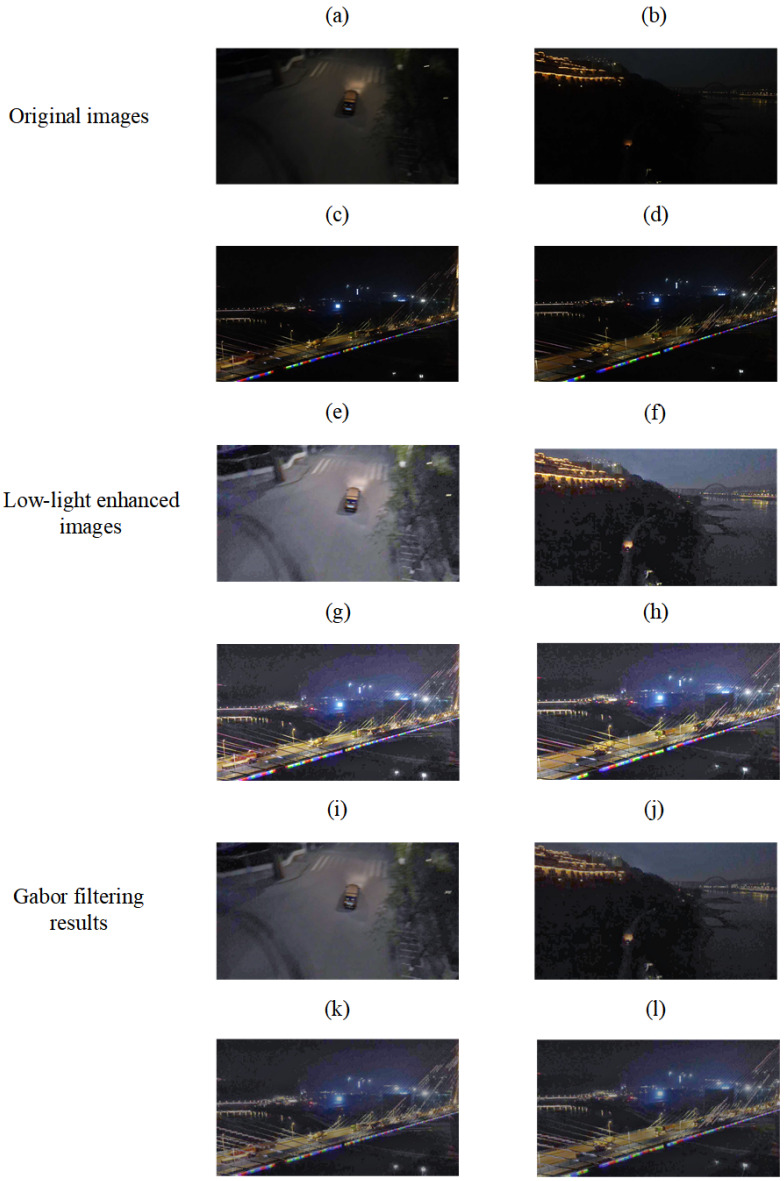


Figure 1. Original Images (**a**–**d**), low-light enhanced images (**e**–**h**) and filtered images (**i**–**l**). Four pictures are a group, (**a**,**e**,**i**), (**b**,**f**,**j**), (**c**,**g**,**k**) and (**d**,**h**,**l**) respectively correspond to the same frame.

4. Experimental Results

4.1. Datasets

The training sets of the Siamese network are the VID dataset [66], GOT-10K dataset [6] and NAT2021 benchmark [60].

The VID dataset contains a total of 5354 video sequences with 30 target categories set. The training set contains 3862 video sequences. Each frame of an image in the training set corresponds to a manual annotation file in xml format, which contains information such as the target ID, target category and target frame.

GOT-10K is a target tracking dataset released by the Chinese Academy of Sciences. The dataset contains more than 10,000 video sequences with more than 1.5 million manually annotated target boxes. The object class contains a total of 563 objects, including five sub-object classes: animal, vehicle, person, passive motion object and object part.

In order to provide an evaluation of long-term tracking performance, we further build a long-term tracking subset, namely NAT2021-L-test, consisting of 23 sequences that are longer than 1400 frames. The NAT2021 dataset is a sequence of nighttime video images containing a test set and an unlabeled training set. It consists of multiple objects (cars, people, buses, trucks, buildings, etc.) or activities (skating, running, cycling, etc.). The training set contains 1400 unlabeled sequences. Different from the training dataset, the NAT2021L of NAT2021 is the test set in these experiments. This is a set of long-tracking datasets.

4.2. Evaluation Metrics

In this paper, we adopt three indices, Intersection over Union (IoU), success rate and precision, as the result evaluation criteria.

- Intersection over Union

Intersection over Union(IoU) [67] represents the degree of overlap between the location of the predicted box and the location of the ground truth box. The larger the IoU, the better the tracker effect. This evaluation criterion can reflect the change in the tracking target scale and directly determine whether the target tracking task fails.

$$IoU = \frac{R_G^k \cap R_O^k}{R_G^k \cup R_O^k} \tag{14}$$

where R_G^k and R_O^k denote the area contained by the ground-truth tracking box and the area contained by the predicted tracking box in the k-th frame, respectively.

- Success rate (SR)

The success rate of object tracking is defined by IoU. If the IoU of a frame is greater than the threshold (T_h= 0.5), the tracking is considered successful. Therefore, the success rate is the percentage of successfully tracked frames out of the total frames. The specific calculation formula is as follows:

$$SR = \frac{Number(IoU \geq T_h)}{N} \tag{15}$$

where $Number(IoU \geq T_h)$ represents the number of frames detected successfully; N represents the total number of frames.

SR will change as the threshold changes. Therefore, we find the success rate curves by plotting the threshold and its corresponding accuracy into a curve.

- Precision

The D is defined as the distance between the center position of the box output by the model and the center position of the ground truth box. The distance is determined

by Euclidean. Precision is the percentage of video frames where D is less than a given threshold. Generally, the threshold is set to 20 pixels.

4.3. Overall Performance

4.3.1. Comparison Algorithms

For a more in-depth analysis of trackers in nighttime tracking, five novel trackers were introduced as comparison algorithms. They are D3S [68], HiFT [18], Ocean [69], UpdateNet [70], SiameseNet [46] and UDAT [60]. They are used for evaluation on the NAT2020L dataset.

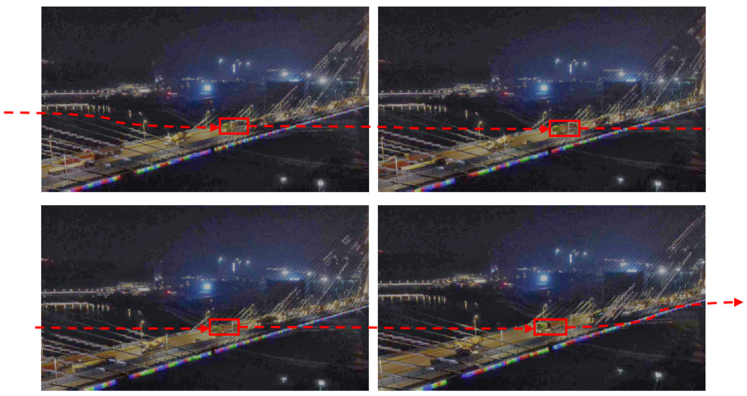
4.3.2. Parameter Settings

The Adam optimizer is applied to train the discriminator. The base learning rate is set to 0.005 and decayed according to a multivariate learning rate strategy with a power of 0.8. The base learning rate of the bridging layer is 0.005 and is optimized with a baseline tracker [63].

4.3.3. Experimental Results

As one of the most common scenarios in object tracking, long-term tracking involves several challenging properties. We tested the feature-augmented NAT2021L dataset with the SDT. The SFDT framework is conducted using PyTorch on an Intel(R) Xeon(R) Silver 4216 CPU. The scores of all methods are shown in Table 1. In order to show the detection performance of the proposed algorithm more clearly, the detection maps are shown in Figure 2. Here are two groups of consecutive frames taken from the test dataset.

(a)



(b)

Figure 2. Tracking maps obtained by the proposed algorithm. (**a**,**b**) are two groups of schematic diagrams, respectively.

Table 1. Performance of trackers on NAT2021L.

Trackers	D3S	HiFT	Ocean	UpdateNet	SiameseNet	UDAT	SFDT
Precision	0.483	0.441	0.460	0.412	0.483	0.506	**0.524**
SR	0.327	0.292	0.313	0.245	0.372	0.376	**0.401**

D3S [71], HiFT [18], Ocean [69], UpdateNet [72], SiameseNet [46], UDAT [60].

5. Discussion

1. The results in Table 1 show that SFDT has the best performance. It demonstrates that the algorithm can achieve competitive long-term tracking performance and significantly improve the tracking performance of the tracker. Figure 2 shows the detection and tracking results of the algorithm proposed in this paper. It can be seen that the proposed method can effectively track the target.
2. In particular, compared with the UDAT algorithm without preprocessing the test set, SFDT performed better than UDAT on both precision and SR. The accuracy is improved by 2%, and the SR is improved by 3%. This means that the data enhanced by lighting and texture features are more suitable for tracking. Under this condition, the network can detect the location of the target more accurately.
3. Paper [60] points out that the low-light enhanced test set is not conducive to object tracking. As shown in Figure 1, the low-light enhanced images are too bright, and the image's details are lost. We speculate that overexposure leads to a decrease in tracking performance. The addition of texture feature enhancement made image details more obvious. Moreover, the brightness of the images after Gabor filtering becomes lower, which makes up for the loss of the previous step. However, the image brightness is still brighter than the original image, which is good for detection and tracking.
4. As shown in Figure 2, two groups of consecutive frames are detected. In the first group of frames, there is only one vehicle as the target, and it was successfully detected; in the second group of frames, there are multiple vehicles in the background as interference, and the vehicles that appear continuously are still successfully tracked by the proposed algorithm. The tracking performance of the proposed algorithm is well illustrated.
5. Although the proposed algorithm is superior to the comparison algorithms in tracking and monitoring, the computational complexity of the algorithm increases because the preprocessing step is improved in this algorithm. A texture feature extraction method based on the Gabor filter is adopted. The processing time for the same dataset is longer than the proposed algorithm in the paper [60]. There is no evaluation for real-time object detection and tracking performance. We will investigate this in future work.

6. Conclusions

In order to enhance the features of nighttime objects and reduce the impact of insufficient lighting, the nighttime training set is preprocessed, including low-light enhancement, object detection and dynamic programming. To make nighttime targets easier to be detected, daytime datasets are processed together through feature extractors. Adversarial learning and domain adaptive structure are used to reduce the gap between features. Low-light enhancement and Gabor filtering are performed on the test data to enhance the characteristics. Finally, the test dataset is processed through the Siamese network to get the results. The experimental results show that the method proposed in this article performs well on multiple evaluation indicators and is superior to the compared algorithms. The proposed algorithm can improve the detection and tracking accuracy of objects. This fully demonstrates the effectiveness of day/night feature modulation and spatial-spectral feature combination and can be extended to other similar application fields.

Author Contributions: Conceptualization, Y.L. and W.F.; methodology, Y.L. and W.F.; software, Y.L. and Y.Z.; validation, Y.L., S.W. and G.D.; writing—original draft preparation, Y.L., W.F. and S.W.; writing—review and editing, Y.L., W.F., Y.Z. and S.W.; visualization, Y.Z.; supervision, M.X.; project administration, M.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Natural Science Foundation of China (Nos. 62201438, 61772397, 12005169); Basic Research Program of Natural Sciences of Shaanxi Province (No. 2021JC-23); Yulin Science and Technology Bureau Science and Technology Development Special Project (No. CXY 2020-094); Shaanxi Forestry Science and Technology Innovation Key Project (No. SXLK2022-02-8); Philosophy and Social Science Research Project of Shaanxi Province (No.20222HZ1759).

Data Availability Statement: The data presented in this study are openly available at https://vision4robotics.github.io/NAT2021/ accessed on 14 August 2022.

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

References

1. Peng, F.; Xu, Q.; Li, Y.; Zheng, M.; Su, H. Improved Kernel Correlation Filter Based Moving Target Tracking for Robot Grasping. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12. [CrossRef]
2. Liu, C.; Ibrayim, M.; Hamdulla, A. Multi-Feature Single Target Robust Tracking Fused with Particle Filter. *Sensors* **2022**, *22*, 1879. [CrossRef] [PubMed]
3. Uzair, M.; Brinkworth, R.S.; Finn, A. Bio-inspired video enhancement for small moving target detection. *IEEE Trans. Image Process.* **2020**, *30*, 1232–1244. [CrossRef] [PubMed]
4. Abro, G.E.M.; Zulkifli, S.A.B.M.; Masood, R.J.; Asirvadam, V.S.; Laouti, A. Comprehensive Review of UAV Detection, Security, and Communication Advancements to Prevent Threats. *Drones* **2022**, *6*, 284. [CrossRef]
5. Fan, H.; Bai, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Huang, M.; Liu, J.; Xu, Y.; et al. Lasot: A high-quality large-scale single object tracking benchmark. *Int. J. Comput. Vis.* **2021**, *129*, 439–461. [CrossRef]
6. 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]
7. Real, E.; Shlens, J.; Mazzocchi, S.; Pan, X.; Vanhoucke, V. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5296–5305.
8. Mahfouz, S.; Mourad-Chehade, F.; Honeine, P.; Farah, J.; Snoussi, H. Target tracking using machine learning and Kalman filter in wireless sensor networks. *IEEE Sens. J.* **2014**, *14*, 3715–3725. [CrossRef]
9. Zhu, S.; Chen, C.; Li, W.; Yang, B.; Guan, X. Distributed optimal consensus filter for target tracking in heterogeneous sensor networks. *IEEE Trans. Cybern.* **2013**, *43*, 1963–1976. [CrossRef]
10. Zhan, R.; Wan, J. Iterated unscented Kalman filter for passive target tracking. *IEEE Trans. Aerosp. Electron. Syst.* **2007**, *43*, 1155–1163. [CrossRef]
11. Hao, J.; Zhou, Y.; Zhang, G.; Lv, Q.; Wu, Q. A review of target tracking algorithm based on UAV. In Proceedings of the 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS), Shenzhen, China, 25–27 October 2018; IEEE: Piscatway, NJ, USA, 2018; pp. 328–333.
12. Guo, H.; Li, W.; Zhou, N.; Sun, H.; Han, Z. Research and Implementation of Robot Vision Scanning Tracking Algorithm Based on Deep Learning. *Scanning* **2022**, *2022*, 3330427. [CrossRef]
13. Ding, Q.; Ding, Z. Machine learning model for feature recognition of sports competition based on improved TLD algorithm. *J. Intell. Fuzzy Syst.* **2021**, *40*, 2697–2708. [CrossRef]
14. Hossain, S.; Lee, D.j. Deep learning-based real-time multiple-object detection and tracking from aerial imagery via a flying robot with GPU-based embedded devices. *Sensors* **2019**, *19*, 3371. [CrossRef] [PubMed]
15. Leclerc, M.; Tharmarasa, R.; Florea, M.C.; Boury-Brisset, A.C.; Kirubarajan, T.; Duclos-Hindié, N. Ship classification using deep learning techniques for maritime target tracking. In Proceedings of the 2018 21st International Conference on Information Fusion (FUSION), Cambridge, UK, 10–13 July 2018; IEEE: Piscatway, NJ, USA, 2018; pp. 737–744.
16. Yang, B.; Cao, X.; Yuen, C.; Qian, L. Offloading optimization in edge computing for deep-learning-enabled target tracking by internet of UAVs. *IEEE Internet Things J.* **2020**, *8*, 9878–9893. [CrossRef]
17. Peng, Y.; Tang, Z.; Zhao, G.; Cao, G.; Wu, C. Motion Blur Removal for Uav-Based Wind Turbine Blade Images Using Synthetic Datasets. *Remote Sens.* **2021**, *14*, 87. [CrossRef]
18. Cao, Z.; Fu, C.; Ye, J.; Li, B.; Li, Y. HiFT: Hierarchical feature transformer for aerial tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 15457–15466.
19. 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.
20. Zhao, B.; Gong, X.; Wang, J.; Zhao, L. Low-Light Image Enhancement Based on Multi-Path Interaction. *Sensors* **2021**, *21*, 4986. [CrossRef]

21. Feng, W.; Quan, Y.; Dauphin, G. Label noise cleaning with an adaptive ensemble method based on noise detection metric. *Sensors* **2020**, *20*, 6718. [CrossRef]
22. Ye, J.; Fu, C.; Cao, Z.; An, S.; Zheng, G.; Li, B. Tracker Meets Night: A Transformer Enhancer for UAV Tracking. *IEEE Robot. Autom. Lett.* **2022**, *7*, 3866–3873. [CrossRef]
23. Ye, J.; Fu, C.; Zheng, G.; Cao, Z.; Li, B. DarkLighter: Light up the darkness for UAV tracking. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; IEEE: Piscatway, NJ, USA, 2021; pp. 3079–3085.
24. Rakhmatulin, I.; Kamilaris, A.; Andreasen, C. Deep neural networks to detect weeds from crops in agricultural environments in real-time: A review. *Remote Sens.* **2021**, *13*, 4486. [CrossRef]
25. Zhu, H.; Wei, H.; Li, B.; Yuan, X.; Kehtarnavaz, N. A Review of Video Object Detection: Datasets, Metrics and Methods. *Appl. Sci.* **2020**, *10*, 7834. [CrossRef]
26. Yang, L.; Liu, S.; Zhao, Y. Deep-Learning Based Algorithm for Detecting Targets in Infrared Images. *Appl. Sci.* **2022**, *12*, 3322. [CrossRef]
27. 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]
28. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
29. Carreira, J.; Sminchisescu, C. CPMC: Automatic object segmentation using constrained parametric min-cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1312–1328. [CrossRef] [PubMed]
30. Van de Sande, K.E.; Uijlings, J.R.; Gevers, T.; Smeulders, A.W. Segmentation as selective search for object recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; IEEE: Piscatway, NJ, USA, 2011; pp. 1879–1886.
31. Pont-Tuset, J.; Arbelaez, P.; Barron, J.T.; Marques, F.; Malik, J. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 128–140. [CrossRef]
32. Wang Lin, L.; Liu, S.; Chen, Y.W. Method and Apparatus of Candidate Generation for Single Sample Mode in Video Coding. US Patent 10,021,418. 10 July 2018 .
33. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
34. Feng, W.; Dauphin, G.; Huang, W.; Quan, Y.; Liao, W. New margin-based subsampling iterative technique in modified random forests for classification. *Knowl.-Based Syst.* **2019**, *182*, 104845. [CrossRef]
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
36. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
37. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99 [CrossRef]
38. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
39. Feng, W.; Quan, Y.; Dauphin, G.; Li, Q.; Gao, L.; Huang, W.; Xia, J.; Zhu, W.; Xing, M. Semi-supervised rotation forest based on ensemble margin theory for the classification of hyperspectral image with limited training data. *Inf. Sci.* **2021**, *575*, 611–638. [CrossRef]
40. Kong, L.; Wang, J.; Zhao, P. YOLO-G: A Lightweight Network Model for Improving the Performance of Military Targets Detection. *IEEE Access* **2022**, *10*, 55546–55564. [CrossRef]
41. Dong, J.; Xia, S.; Zhao, Y.; Cao, Q.; Li, Y.; Liu, L. Indoor target tracking with deep learning-based YOLOv3 model. In Proceedings of the Fourteenth International Conference on Digital Image Processing (ICDIP 2022), Wuhan, China, 20–23 May 2022; SPIE: Bellingham, WA, USA, 2022; Volume 12342, pp. 992–998.
42. Jiang, S.; Xu, B.; Zhao, J.; Shen, F. Faster and simpler siamese network for single object tracking. *arXiv* **2021**, arXiv:2105.03049.
43. Tao, R.; Gavves, E.; Smeulders, A.W.M. Siamese Instance Search for Tracking. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1420–1429.
44. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 850–865.
45. 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, 15–20 June 2019; pp. 4282–4291.
46. 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.

47. 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, New York, NY, USA, 7–8 February 2020; Volume 34, pp. 12549–12556.
48. 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, 20–25 June 2021; pp. 8126–8135.
49. Wang, N.; Zhou, W.; Wang, J.; Li, H. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1571–1580.
50. Liu, Y.; Zhang, S.; Li, Y.; Yang, J. Learning to Adapt via Latent Domains for Adaptive Semantic Segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 1167–1178.
51. Rakshit, S.; Bandyopadhyay, H.; Bharambe, P.; Desetti, S.N.; Banerjee, B.; Chaudhuri, S. Open-Set Domain Adaptation Under Few Source-Domain Labeled Samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4029–4038.
52. Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; Van Gool, L. Domain adaptive faster r-cnn for object detection in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3339–3348.
53. Yu, Q.; Fan, K.; Wang, Y.; Zheng, Y. Faster MDNet for Visual Object Tracking. *Appl. Sci.* **2022**, *12*, 2336. [CrossRef]
54. Moon, J.; Das, D.; Lee, C.G. A Multistage Framework With Mean Subspace Computation and Recursive Feedback for Online Unsupervised Domain Adaptation. *IEEE Trans. Image Process.* **2022**, *31*, 4622–4636. [CrossRef]
55. Acharya, D.; Tennakoon, R.; Muthu, S.; Khoshelham, K.; Hoseinnezhad, R.; Bab-Hadiashar, A. Single-image localisation using 3D models: Combining hierarchical edge maps and semantic segmentation for domain adaptation. *Autom. Constr.* **2022**, *136*, 104152. [CrossRef]
56. He, L.; Liu, C.; Li, J.; Li, Y.; Li, S.; Yu, Z. Hyperspectral image spectral–spatial-range Gabor filtering. *IEEE Trans. Geosci. Remote. Sens.* **2020**, *58*, 4818–4836. [CrossRef]
57. Li, C.; Guo, C.; Loy, C.C. Learning to enhance low-light image via zero-reference deep curve estimation. *arXiv* **2021**, arXiv:2103.00860.
58. 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, 27–30 June 2016; pp. 770–778.
59. Zheng, J.; Ma, C.; Peng, H.; Yang, X. Learning to Track Objects from Unlabeled Videos. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 13526–13535. [CrossRef]
60. Ye, J.; Fu, C.; Zheng, G.; Paudel, D.P.; Chen, G. Unsupervised domain adaptation for nighttime aerial tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8896–8905.
61. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef] [PubMed]
62. 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.
63. Ganin, Y.; Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; PMLR: Moscow Region, Russia 2015; pp. 1180–1189.
64. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
65. Grigorescu, S.E.; Petkov, N.; Kruizinga, P. Comparison of texture features based on Gabor filters. *IEEE Trans. Image Process.* **2002**, *11*, 1160–1167. [CrossRef]
66. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscatway, NJ, USA, 2009; pp. 248–255.
67. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
68. Lukezic, A.; Matas, J.; Kristan, M. D3S-A Discriminative Single Shot Segmentation Tracker. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
69. Zhang, Z.; Peng, H.; Fu, J.; Li, B.; Hu, W. Ocean: Object-Aware Anchor-Free Tracking. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 771–787.
70. Zhang, L.; Gonzalez-Garcia, A.; Weijer, J.V.D.; Danelljan, M.; Khan, F.S. Learning the Model Update for Siamese Trackers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
71. Lukezic, A.; Matas, J.; Kristan, M. D3S—A Discriminative Single Shot Segmentation Tracker. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.

72. Zhang, L.; Gonzalez-Garcia, A.; Joost, V.; Danelljan, M.; Khan, F.S. Learning the Model Update for Siamese Trackers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4010–4019.