symmetry

MDPI

Article

Learning Multifeature Correlation Filter and Saliency Redetection for Long-Term Object Tracking

Liqiang Liu [1,*], Tiantian Feng [2] and Yanfang Fu [1,*]

[1] School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, China
[2] School of Physics and Optoelectronic Engineering, Xidian University, Xi'an 710071, China; ttfeng@stu.xidian.edu.cn
* Correspondence: liuliqiang@xatu.edu.cn (L.L.); fuyanfang@xatu.edu.cn (Y.F.)

Abstract: Recently due to the good balance between performance and tracking speed, the discriminative correlation filter (DCF) has become a popular and excellent tracking method in short-term tracking. Computing the correlation of a response map can be efficiently performed in the Fourier domain by the discrete Fourier transform (DFT) of the input, where the DFT of an image has symmetry in the Fourier domain. However, most of the correlation filter (CF)-based trackers cannot deal with the tracking results and lack the effective mechanism to adjust the tracked errors during the tracking process, thus usually perform poorly in long-term tracking. In this paper, we propose a long-term tracking framework, which includes a tracking-by-detection part and redetection part. The tracking-by-detection part is built on a DCF framework, by integrating with a multifeature fusion model, which can effectively improve the discriminant ability of the correlation filter for some challenging situations, such as occlusion and color change. The redetection part can search the tracked object in a larger region and refine the tracking results after the tracking has failed. Benefited by the proposed redetection strategy, the tracking results are re-evaluated and refined, if it is necessary, in each frame. Moreover, the reliable estimation module in the redetection part can effectively identify whether the tracking results are correct and determine whether the redetector needs to open. The proposed redetection part utilizes a saliency detection algorithm, which is fast and valid for object detection in a limited region. These two parts can be integrated into DCF-based tracking methods to improve the long-term tracking performance and robustness. Extensive experiments on OTB2015 and VOT2016 benchmarks show that our proposed long-term tracking method has a proven effectiveness and high efficiency compared with various tracking methods.

Keywords: visual object tracking; saliency detection; long-term tracking; multifeature fusion; correlation filter tracking

Citation: Liu, L.; Feng, T.; Fu, Y. Learning Multifeature Correlation Filter and Saliency Redetection for Long-Term Object Tracking. *Symmetry* **2022**, *14*, 911. https://doi.org/10.3390/sym14050911

Academic Editors: Dejun Zhang, Whoi-Yul Kim and Moonsoo Ra

Received: 7 March 2022
Accepted: 27 April 2022
Published: 29 April 2022

1. Introduction

Visual object tracking is a fundamental task in computer vision and machine learning, aiming to localize the tracked object in the rest of the image sequences after giving the object localization and scale information in the first frame [1,2]. Although significant achievements have been made to improve tracking performance in the last few decades [3–7], there are still many challenges in object tracking field, especially in long-term tracking, where the tracked object may easily suffer from some challenging situation, such as heavy occlusion, disappearing and reappearing, deformation, and color change. Object tracking also has been widely used in many applications, such as unmanned aerial vehicle (UAV), self-driving car, video monitoring system, telemedicine, autonomous landing, military combat system, and so on.

Tracking-by-detection methods [8,9] have gained much popularity and have had a huge success for visual object tracking in recent years. These tracking methods usually identify objects through a detector and update the detector online to keep up with the

changes of the tracked objects in both appearance and scale. However, existing tracking-by-detection methods also may be misled by corrupted detection results due to deformation or occlusion situation.

Due to the high efficiency and good performance, DCF-based trackers [10–12] have attracted much attention. As one of the state-of-the-art tracking-by-detection methods, CF-based tracking methods have had great success. In addition, computing the correlation of a response map can be performed efficiently in the Fourier domain by the discrete Fourier transform (DFT) of input, where the DFT of an image has symmetry in the Fourier domain. Recently, many trackers have been proposed based on the DCF framework to improve the tracking performance in long-term tracking, which is more difficult compared to short-term tracking due to the occlusion and appearance changes occurring continually in a longer video sequence. Therefore, the reliability of the detection results is crucial for updating the detection model, which suffers the risk of drifting to the background. The search region of standard DCF-based trackers is usually cropped with twice the size of the tracked object at the center position, but sometimes the object moves out of the search region where the detector cannot detect the object. To solve this problem, some methods [13,14] search the object in a larger region which may contain the tracked object with a higher probability, and they can better handle the occlusion and motion. However, these methods are not reliable for facing full occlusion and serious deformation in long-term tracking, and a small wrong estimation could lead to tracking failure due to the accumulation of previous frames. Hence, the critical process is how to identify the reliability of the detection result and refine it during the tracking process. In long-term tracking, an effective redetection module is necessary. Some methods [15,16] simply use the redetected results to replace the original detection results, which may corrupt the detection model further. Some long-term tracking methods [17,18] incorporate with multiple trackers to improve the ability of the detector, which leads to a heavy computation and obviously influence the tracking speed. The main challenge for these long-term trackers is how to meet the real-time requirement while also improving tracking performance.

In this paper, we propose a long-term tracker with multiple features and saliency redetection (MSLT), which consists of tracking-by-detection and redetection parts. Inspired by the excellent Staple [19] tracker, which combines the histogram of oriented gradient (HOG) and color features, and we take it as a baseline tracker. The main contributions of this paper are listed as follows:

- As a crucial part of our tracker, the redetector part contains a reliable estimation module and redetector module. The proposed tracking-by-detection part integrates with multiple features in the correlation filter, which is equipped with color and HOG features for tracking.
- The estimation module determines whether it is necessary to replace the previous tracking result and whether to start the redetection process. Considering the tracking speed and performance, we employ a saliency detector for redetecting the tracked object, which is fast and valid for object detection in a limited region. This re-detection module is more effective and can locate the object after it reappears in the image.
- Our MSLT method is evaluated by extensive experiments, which compare results with several state-of-the-art methods on two benchmarks, including OTB2015 [20] and VOT2016 [21]. Both qualitative and quantitative experiments demonstrate the favorable effectiveness of our tracker.

2. Related Work

In recent years, many trackers have been proposed and achieved great success in visual object tracking field. Here we review three categories of trackers that are relevant to our work.

2.1. Correlation-Filter-Based Tracking

In the visual tracking field, DCF has gained a lot of popularity and achieved impressive performance. In the DCF framework, correlation filters are trained by minimizing a least-squares loss for all circular-shifted samples and transforming the objective function into the Fourier domain to reduce the heavy computation. The first correlation filter framework was proposed by Bolme et al. [22] who used gray features to train a minimum output sum of squared error filter (MOSSE) with a high speed. Henriques et al. [10] exploited the circulant structure of training patches and proposed a kernel correlation filter (KCF) tracker by combining multidimensional features with kernels. Some trackers were proposed to adapt to the change of object scale by using multiscale correlation filters, such as DSST [23] and SAMF [24]. The SRDCF [25] tracker addressed the boundary effects problem by introducing a spatial regularization term to penalize the correlation filter coefficients that enable the correlation filter to be learned on larger image regions, and lead to a more discriminative appearance model. Similarly, the BACF tracker [26] exploited real background patches together with the target patch and used an online adaptation strategy to update the tracker model to alleviate the boundary effects. Recently, with the development of convolution neural networks (CNN) in object detection and classification, some trackers have also used the CNN features pretrained on a large object recognition to replace or combine handcrafted features, such as C-COT [27], HCF [28], ECO [29], and so on. Finally, the CFNet [30] tracker proposed an end-to-end framework, which interpreted the correlation filter learner as a differentiable layer in a deep neural network.

2.2. Tracking-by-Detection

Tracking-by-detection methods regard the tracking problem as a classification problem by learning a discriminative model, such as with a support vector machine (SVM) and partial filter-based tracking. The TLD method [31] consisted of three tasks, including tracking, learning, and detection, in which the tracking and detection tasks ran simultaneously. Inspired by the TLD method, many related trackers have been proposed. LMCF [32] employed a structured output SVM into a CF framework and combined two kinds of algorithms. The MEEM [33] method collected snapshots and picked the best prediction result from the SVM framework. Struct learned a structured output to update the detector. Lu et al. [34] proposed a robust object tracking algorithm by using a collaborative model which exploited both holistic templates and local representations. Zhang et al. [35] proposed a novel circulant sparse tracker (CST), which exploited circulant target templates. These above tracking-by-detection methods focus on short-term tracking and perform poorly when facing some challenging situations.

2.3. Long-Term Tracking

Long-term tracking focuses on solving challenging situations, such as object disappearing and reappearing, partial occlusion, and full occlusion. MUSTer [36] maintained a short-term memory for detection and a long-term mechanism for searching the object via key-point matching. However, the MUSTer method needs to evaluate the integrated trackers in every frame. LCT [37] learned discriminative correlation filters for estimating the translation and scale variation, and the authors also developed a robust online detector using random ferns to redetect objects in case of tracking failure. The PTAV method [38] proposed a framework that contains three parts, including a base tracker T, verifier V and their coordination mechanism. The base tracker T used a DCF-based tracker, while the verifier V used a Siamese network to verify the similarity between two objects. Wang et al. [39] and Tang et al. [40] utilized a reliable redetection mechanism with a DCF-based tracker for long-term tracking.

3. Method

3.1. Framework

The overall framework of our MLST method is shown in Figure 1. It contains three modules, including tracking-by-detection, reliability estimation, and redetection. Firstly, for the tracking-by-detection module, we process it with twice the size of the region of interest patch in the input frame and employ a DCF model based on HOG features and a color histogram model based on color features to obtain the related response maps. Then, we estimate the responses by the peak-to-sidelobe ratio (PSR) and color ratio of the related region, respectively. The reliability estimation can decide whether to recall the redetection module. If the tracking result passes through the reliability estimation, then we keep the original tracking result as the final result, if not, we go through the redetection process, where a saliency detector is recalled with a larger search region. We introduce the related modules in the following sections.

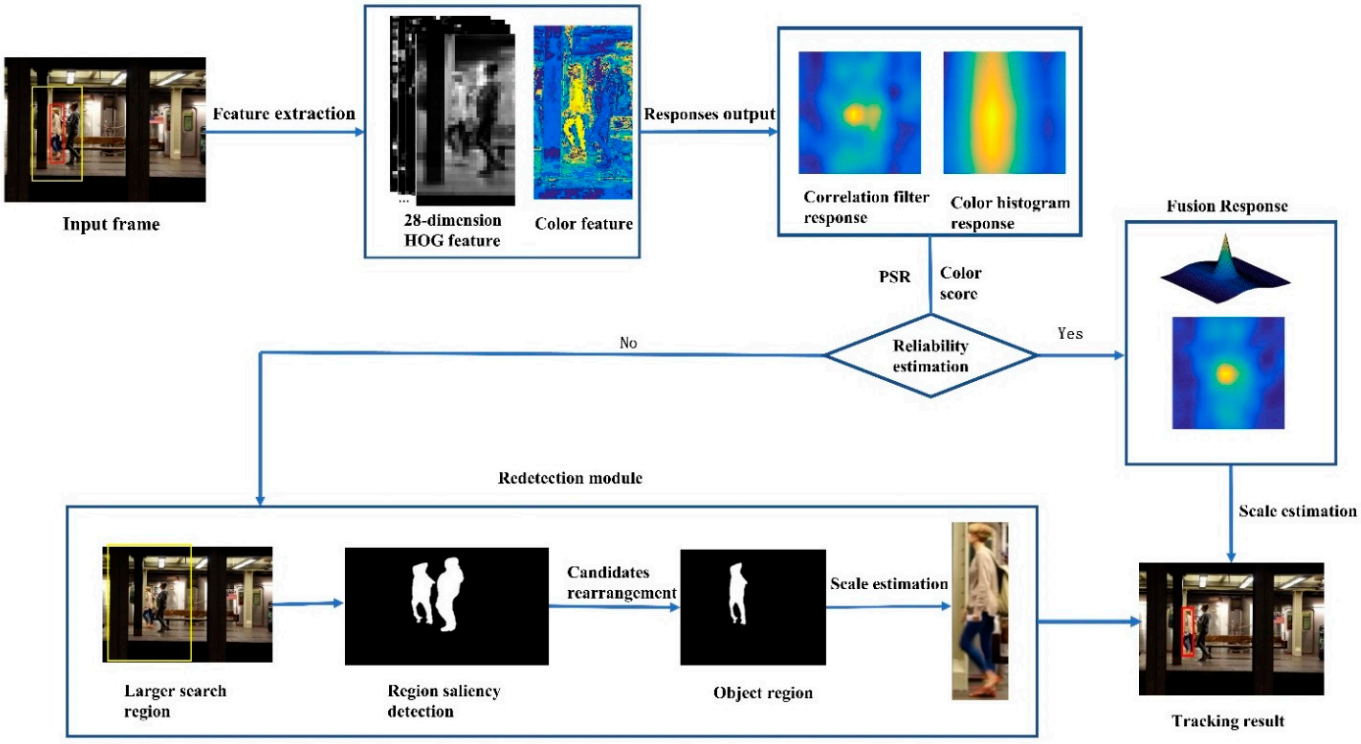

Figure 1. Overall framework of the proposed MSLT. It contains three modules, including tracking-by-detection, reliability estimation, and redetection.

3.2. Multifeature Fusion

- Correlation Filter Response with HOG Features

The standard DCF model learned a discriminative correlation filter on image patch x with d channels, and the size x was 2.5 times that of the tracked object. All training samples were generated from the circular shift operation on the tracked object and extracted 28-dimension HOG features as an appearance description. Thus, the objective function of the DCF used a Tikhanov regularization which can be formulated as,

$$\varepsilon(f) = \left|\left| Xf - y \right|\right|^2 + \lambda_1 \sum_{l=1}^{d} ||f_l||^2 \quad (1)$$

where y is a desired response, which use the Gaussian-shaped ground truth generally, and λ_1 is a regularization factor. To reduce the computation cost, Equation (1) can be

transformed into the Fourier domain through Parseval's Theorem. The objective function has a closed-form solution and the solution for the lth channel can be expressed as,

$$\hat{f}_l = \frac{\hat{x}_l \odot \hat{y}^*}{\sum_{l=1}^{d} \hat{x}_l^* \odot \hat{x}_l + \lambda_1} \quad l = 1, 2, \ldots d, \tag{2}$$

where $\odot$ denotes the element-wise product, the symbol $\hat{}$ stands for the discrete Fourier transform (DFT) of a vector, and $\hat{x}_l^*$ is the complex-conjugate of $\hat{x}_l$. For efficient updates, we used a linear update strategy to update the numerator $\hat{A}_l^t$ and denominator $\hat{B}_l^t$ of Equation (2),

$$\hat{A}_l^t = (1 - \eta_h)\hat{A}_l^{t-1} + \eta_h \hat{x}_l^t \odot \hat{y}^* \tag{3}$$

$$\hat{B}_l^t = (1 - \eta_h)\hat{B}_l^{t-1} + \eta_h \sum_{l=1}^{d} \hat{x}_l^{*t} \odot \hat{x}_l^t \tag{4}$$

where η_h denotes the learning rate of the correlation filter. To reduce the boundary effects during the learning process, we employed Hann windows on the samples [41].

During the tracking stage, an image patch z^t with the same size of training sample x^{t-1} was cropped at the last location, and generated a response map R_h^t by correlating with the learned filter of the last frame,

$$R_h^t = \mathcal{F}^{-1}\left\{ \frac{\sum_{l=1}^{d} \hat{A}_l^{t-1} \odot \hat{z}_l^t}{\hat{B}^{t-1} + \lambda_1} \right\} \tag{5}$$

where $\hat{A}_l^{t-1}$ and $\hat{B}_l^{t-1}$ are the numerator and denominator of the filter in the previous frame, respectively, and $\mathcal{F}^{-1}$ denotes the inverse DFT. The correlation filter response with HOG features is shown in Figure 2, where we selected 28-dimension HOG features.

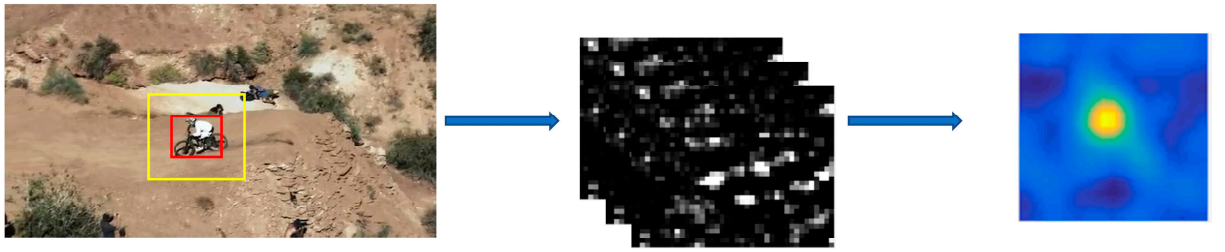

Figure 2. The process of correlation filter response with HOG features.

- Color Histogram Response

For long-term tracking, a color histogram model was adopted for some challenging situations, such as color change and blur motion. Figure 3 shows the generation of a color histogram model. Inspired by [19], the histogram weight vector **m** was obtained via minimizing the regression error; the formula is as follows,

$$\min_{\mathbf{m}} \sum_x \left|\left| \sum_{v \in \Re} \mathbf{m}^T \varphi_x(v) - y \right|\right|^2 + \lambda_2 \left|\left| \mathbf{m} \right|\right|^2 \tag{6}$$

where $\varphi_x(v)$ denotes the feature pixels of patch x in the finite region $\Re$, y are corresponding labels, and λ_2 is a regularization factor of the color histogram model. Following [41], the

equation can be transformed into a linear regression method for every single pixel on target region $\mathcal{O}$ and background region $\mathcal{B}$, and the formula after simplification is as follows,

$$\min_{\mathbf{m}} \frac{1}{|\mathcal{O}|}\sum_{v\in\mathcal{O}}\left|\mathbf{m}^T\varphi_x(v)-1\right|^2+\frac{1}{|\mathcal{B}|}\sum_{v\in\mathcal{B}}\left|\mathbf{m}^T\varphi_x(v)\right|^2 \tag{7}$$

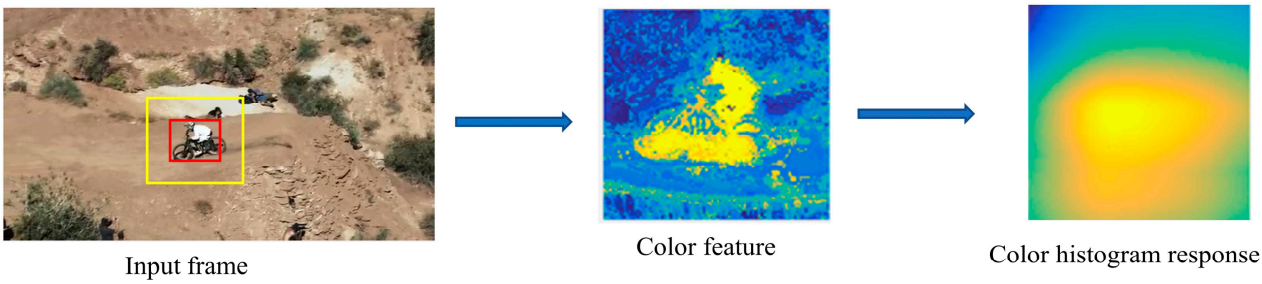

Figure 3. The process of color histogram response.

Solved by method of ridge regression, the solution can be found:

$$m^j=\frac{p^j(\mathcal{O})}{p^j(\mathcal{O})+p^j(\mathcal{B})+\lambda_2} \tag{8}$$

For each dimension $j=1,2,\ldots M$, $p^j(\mathcal{E})$ denotes the jth pixel of vector $\mathbf{p}$. Similarly, the update strategy can be expressed as follows:

$$\mathbf{p}_{\mathrm{t}}(\mathcal{O})=(1-\eta_c)\mathbf{p}_{\mathrm{t}-1}(\mathcal{O})+\eta_c\mathbf{p}'_t(\mathcal{O}) \tag{9}$$

$$\mathbf{p}_{\mathrm{t}}(\mathcal{B})=(1-\eta_c)\mathbf{p}_{\mathrm{t}-1}(\mathcal{B})+\eta_c\mathbf{p}'_t(\mathcal{B}) \tag{10}$$

where η_c is a learning rate of the color histogram model. Similar to the correlation filter model, after obtaining the histogram weight vector, we computed the color histogram response R_c for the given image patch z.

$$R_c=\mathbf{m}^T\varphi_{v\in z}(v) \tag{11}$$

- **Response Fusion**

The above correlation filter response with HOG features and color histogram response can be utilized for object tracking. For more accurate tracking, we combined them with a linear fusion,

$$R_{final}=\gamma\cdot R_h+(1-\gamma)R_c \tag{12}$$

where γ is a fusion weight factor. The position of the tracked object in the current frame was defined as the maximal value of R_{final}, while the scale estimation used the DSST method [23]. Figure 4 shows the fusion between the correlation filter response and the color histogram response. In tracking challenges, HOG features are good for occlusion, while color features are good for deformation, color change, and so on. Therefore, the tracker with response fusion performed well when evaluated on occlusion, color change, and deformation challenges.

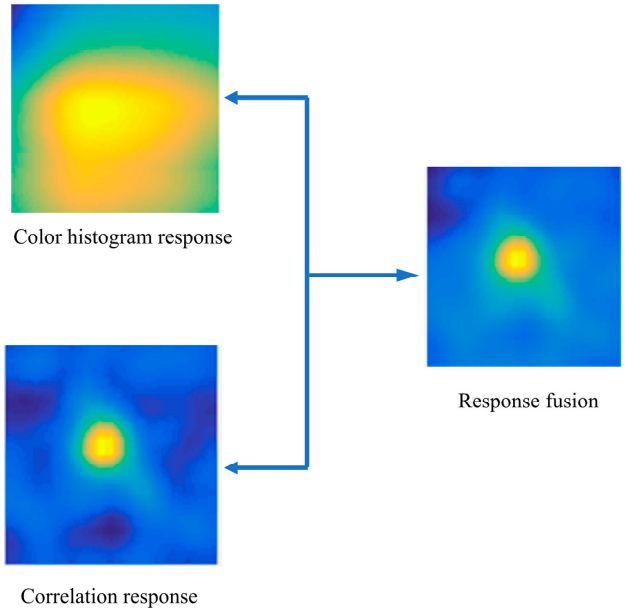

Figure 4. Response fusion with color histogram and correlation response.

3.3. Redetection Module

The redetection module contained two processes, reliability estimation and saliency detection, respectively. When the tracking results arrive, we need to estimate their reliability by the correlation filters and color histogram responses. Then, we introduce how to redetect the object when the tracking result is not reliable.

3.3.1. Reliability Estimation

For the correlation filter response with HOG features, we computed the value of the peak-to-sidelobe ratio (PSR) to quantify the confidence in the results. If the value of PSR was low, then correlation confidence was low. The PSR of the correlation filter response S_h^t can be expressed as follows:

$$S_h^t = \frac{\max\left(R_h^t\right) - \mu_t}{\sigma_t} \tag{13}$$

where μ_t and σ_t denote the mean and standard deviation of the correlation response R_h^t, respectively. The superscript t of the correlation filter response denotes the t-th frame. If a large peak value appears in the target area, while a smooth low value occurs in other areas, it indicates that the tracking result is reliable and matched with the target. On the contrary, when the tracking result is not reliable, the response graph has multiple peak values with low values, and PSR value decreases significantly. Therefore, PSR value can reflect the quality of the tracking results to a certain extent.

Considering the little change between two consecutive frames, we defined the average value of the PSR value in the previous frames as a threshold value. The set of PSR values for previous frames was $C_h = \left\{S_h^2, S_h^3, \ldots S_h^{t-1}\right\}$, and its average value was defined as M_h. The reliability estimation criteria of correlation filter response can be defined as:

$$S_h^t < \tau_1 \cdot M_h \tag{14}$$

where τ_1 is a constant less than 1. The above formula indicates that when the PSR value of the current frame does not satisfy the reliability estimation criteria, the target tracking has failed under the module of the correlation filter with HOG features.

For the color histogram response, a color region was obtained by adding all pixels of target region in the first frame, and then the color score was defined according to the

proportion of the color region in the response to the color histogram obtained in each subsequent frame. The formula of the color score was set as follows:

$$S_c^t = \frac{\sum_v \mathbf{m}^T \varphi_t(v)}{\sum_v \mathbf{m}^T \varphi_1(v)} \tag{15}$$

where S_c^t denotes the color score of the tth frame. Similarly, the reliability estimation criteria of the color histogram response can be defined as:

$$S_c^t < \tau_2 \cdot M_c \tag{16}$$

where τ_2 is a constant less than 1, and M_c denotes the average value of the set $C_h = \left\{1, S_t^2, S_t^3, \ldots S_t^{t-1}\right\}$, which includes the color score of the first frame.

3.3.2. Saliency Detection and Candidates Sort

For the saliency redetector, we used an existing algorithm [42] to obtain the saliency map efficiently. Considering the tracked object may be out of view, we detected the object in a larger region with the saliency detection. If two or more salient objects were detected, we needed to rearrange the candidate targets and choose the best matched. Assuming that N salient candidates $(z_1, z_2, \ldots, z_N)$ were obtained, we computed their correlations with the original correlation filter template H.

$$R(z_n) = \mathcal{F}^{-1}(H \odot Z_n) \tag{17}$$

where Z_n is the feature description of candidate z_n in frequency domain. The symbol $\odot$ denotes the matrix dot product. We sorted the responses of all candidate boxes and set the candidate box with the maximum response as the final detected object. The related formula can be expressed as follows:

$$R_t = \max(R(z_1), R(z_2), \ldots, R(z_N)) \tag{18}$$

Through the threshold process, we can obtain the salient region S_t of an image patch, and its center coordinate (x_s, y_s) can be computed as:

$$x_s = \frac{\sum\limits_{i,j \in [W,H]} S_t(i,j) \cdot i}{\sum\limits_{i,j \in [W,H]} S_t(i,j)} \tag{19}$$

$$y_s = \frac{\sum\limits_{i,j \in [W,H]} S_t(i,j) \cdot j}{\sum\limits_{i,j \in [W,H]} S_t(i,j)} \tag{20}$$

where $S_t(i,j)$ is a saliency value of pixel (i,j) in target region, with the size $[W,H]$. With the calculated coordinate (x_s, y_s) as the position, we used the DSST algorithm [23] to calculate the size of the candidate box.

3.4. Algorithm Description

The formal description of the proposed tracking method is given in Algorithm 1.

Algorithm 1: Long-term tracker with multiple features and saliency redetection (MSLT).

Input: The initial position l_0, tracked object position l_{t-1}, and scale of the $(t-1)$th frame;
Output: The predicted object position l_t and scale of the tth frame;
Repeat:
1. Extract features and compute related HOG features and color features maps in the search region of the tth frame.
2. Compute the correlation filter and color histogram responses, respectively.
3. Compute the reliability estimation PSR value S_h^t and color score S_c^t by using Equations (13) and (15).
4. If $S_h^t < \tau_1 \cdot M_h$ **and** $S_c^t < \tau_2 \cdot M_c$, then
Start the saliency detection in a larger search region, and obtain N salient candidates $(z_1, z_2, \ldots, z_N)$;
If N = 1
Take this object as saliency detection object;
Else
Compute the correlations between salient candidates and original correlation filter template H using (17). Sort the responses of all candidate boxes, and set the maximum response as the final salient object;
End if
Compute the center coordinate (x_s, y_s) of the salient object using Equations (19) and (20), and estimate the related scale;
Else
Fuse the response using Equation (12), and set the maximum value of the fusion response as the central position l_t of the target; estimate the scale of tracked object and obtain the final tracking result.
End if
5. Update the correlation filter model and color histogram by using (3), (4), (9), and (10).

4. Experiments

In this section, we implemented our MSLT method based on the MATLAB 2017a platform and run on a PC machine equipped with an Intel 3.7 GHz and 16 GB RAM. For the parameters, we set the cell size of HOG as 4 and the bin value of color histogram as 32. The learning rates of the correlation filter with HOG features and color histogram were set as 0.01 and 0.04, respectively. The regularization parameters λ_1 and λ_2 were both set as 0.001, and the fusion response parameter was 0.3. The search region size of the saliency detection module was set as four times the size of the tracked object. The related scale parameter setting was according to DSST [23]. We evaluated our MLST on two benchmarks, including OTB2015 [20] and VOT2016 [21]. For a fair comparison, we used the available codes or results provided by the related authors.

4.1. OTB2015 Dataset

The OTB2015 [19] dataset is a popular and classical tracking dataset which contains 100 video sequences; we took part of them for this experiment. It is fully annotated with 11 different attributes, including fast motion (FM), background clutter (BC), deformation (DEF), motion blur (MB), occlusion (OCC), in-plane rotation (IPR), out-of-plane rotation (OPR), out-of-view (OV), low resolution (LR), illumination variation (IV), and scale variation (SV). To be fair, we evaluated all compared trackers based on a one-pass evaluation (OPE) protocol provided in [43], which was employed to evaluate the compared trackers from two metrics, namely distance precision and overlap success. In this experiment, we compared our tracker with seven state-of-the-art trackers, which included SRDCFdecon [44], STAPLE_CA [45], STAPLE [19], SAMF [24], DSST [23], KCF [10], and CSK [46].

Figure 5 shows the plots of the overall precision and success rates of different trackers on the part of OTB2015 dataset [20], where the legend in the plots denotes the average distance precision (DP) score at 20 pixels and the area-under-the-curve (AUC) score [43], respectively. We can see that our MSLT tracker performs better, and the DP score and AUC score are 0.874 and 0.639, respectively. Compared with the baseline STAPLE_CA tracker, our overall DP and AUC scores improve by 4.2% and 2.6%, respectively. Compared with the KCF tracker, our tracker has obvious advantages, where the overall DP score improved by 14% and AUC score improved by 13.3%. The tracking speed of our tracker reaches 31.2 fps, which meets the real-time requirement.

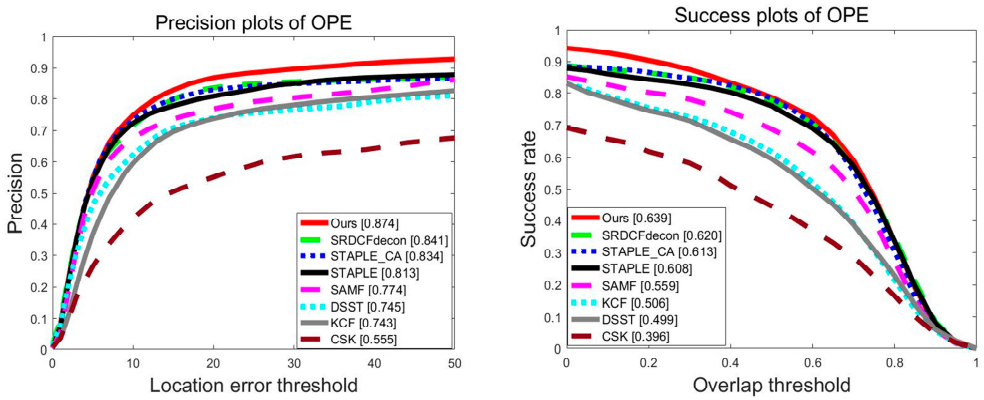

Figure 5. Overall precision plots and success plots of OPE between seven methods and our proposed method.

Figures 6 and 7 show the precision and success plots of OPE among seven methods and our proposed method with different video sequence attributions. It can be seen that our tracker outperforms other trackers in both DP and AUC scores with all contributions except the FM challenge. Take the OCC contribution for example, compared with the SRDCFdecon tracker (ranks second), our tracker improved by 2.1% in the DP and 0.9% in the AUC score.

4.2. VOT2016 Dataset

The visual object challenge (VOT) is a workshop at the IEEE International Conference on Computer Vision (ICCV) and European Conference on Computer Vision (ECCV), started in 2013; the VOT2016 dataset [21] consists of 60 video sequences. In VOT2016, there are ten new difficult sequences replacing ten simple sequences of VOT2015, but no change is made in the evaluation metrics. We compared our MSLT model with six trackers, including KCF [10], SRDCF [25], DSST [23], Staple [19], DAT [47], and ECO [29]. We used three evaluation metrics [48] in this experiment, accuracy, robustness and expected average overlap (EAO).

Tables 1 and 2 present the accuracy and robustness results of seven compared trackers with nine contributions, which include camera motion, empty, illumination change, motion change, occlusion, size change, mean weighted mean, and pooled [48]. A higher accuracy score indicates a better performance, while a lower robustness score indicates a better performance. Table 1 reports that our MSLT achieves the best performance on most attributions, such as occlusion, camera motion, and pooled. Similarly, Table 2 shows that our tracker performs well in robustness results, which illustrate the times of failures in the sequences. For the occlusion contribution, our tracker ranks first with an accuracy of 0.5253 and robustness of 10. Moreover, our MSLT tracker ranks first on both the pooled and weighted mean contributions. The expected average overlap metrics are shown in Figure 8, and the related EAO values are shown in Table 3. It can be seen that the EAO value of our MSLT is 0.3737, which ranks first among all the compared trackers.

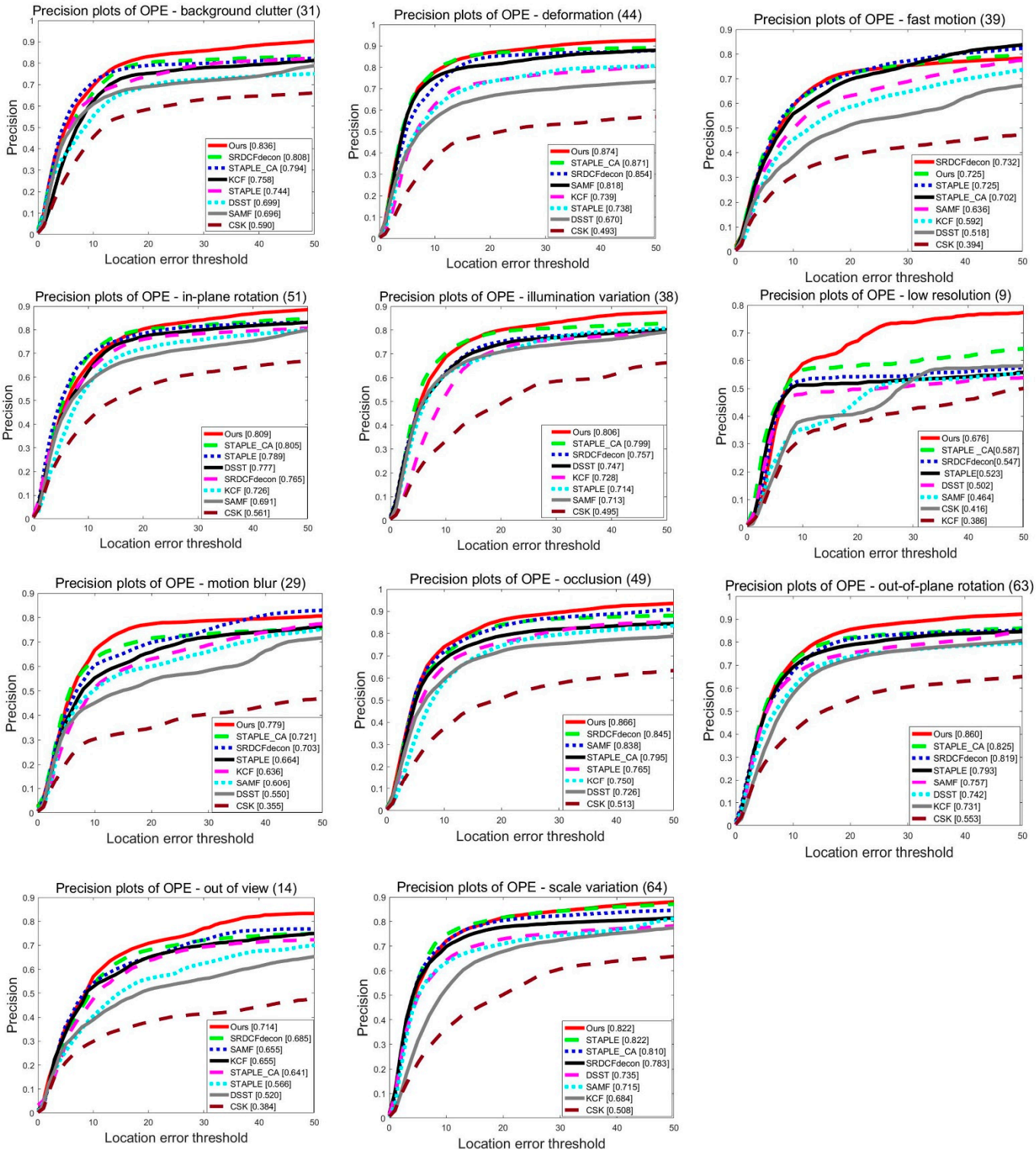

Figure 6. Precision plots of OPE between seven methods and our proposed method with different video sequence attributions.

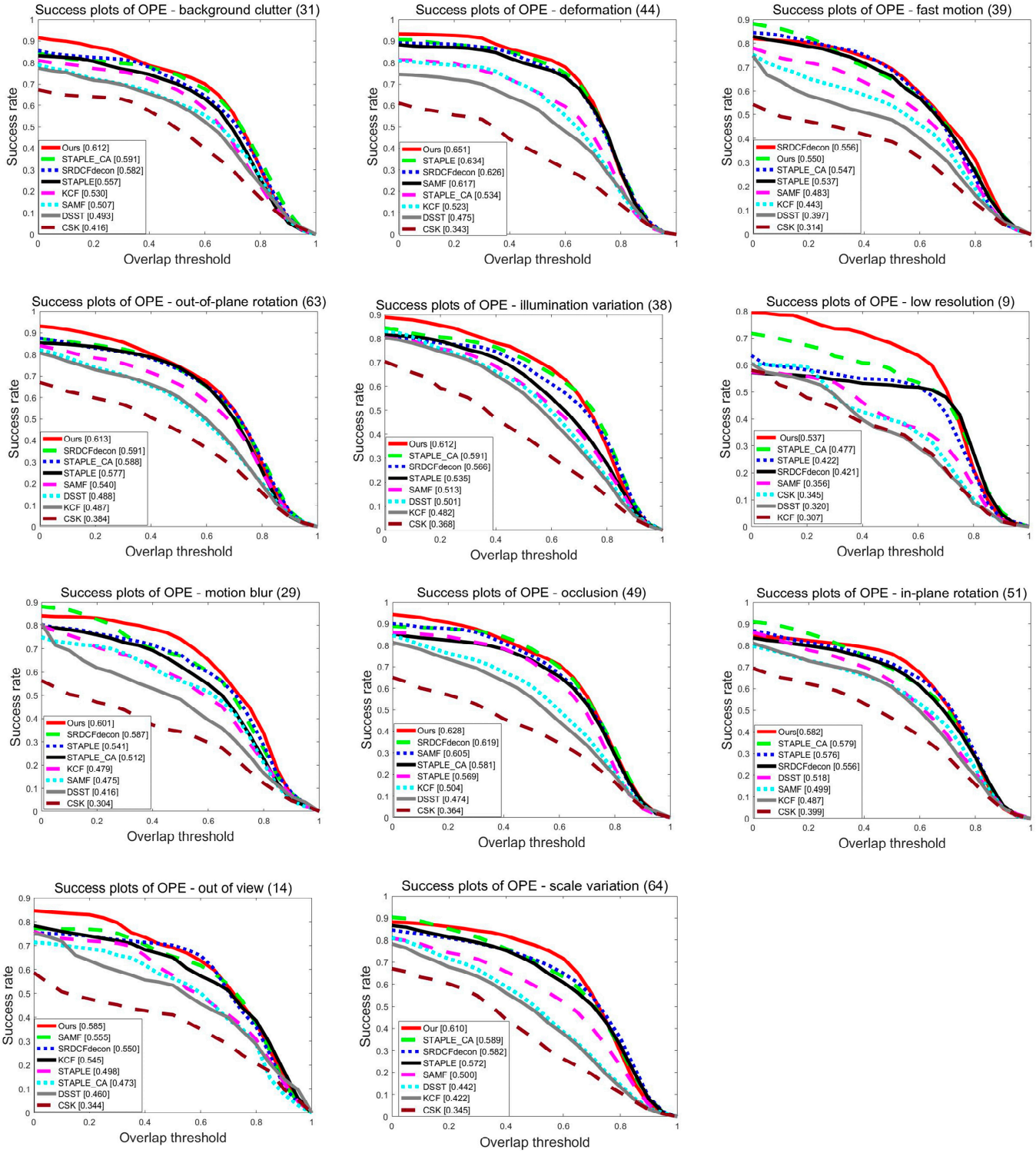

Figure 7. Success rate plots of OPE between seven methods and our proposed method with different video sequence attributions.

Table 1. Accuracy of seven trackers evaluated on VOT2016. The bold fonts mean the best result.

	Camera Motion	Empty	Illum Change	Motion Change	Occlusion	Size Change	Mean	Weighted Mean	Pooled
MSLT	**0.5757**	**0.5943**	0.6975	**0.5078**	**0.5253**	0.4232	0.5338	**0.5484**	**0.5524**
ECo	0.5667	0.5748	**0.7084**	0.4997	0.5019	**0.4631**	0.5423	0.5395	0.5499
Staple	0.5513	0.5833	0.7091	0.5051	0.5110	0.4328	**0.5491**	0.5403	0.5400
SRDCF	0.5517	0.5785	0.6802	0.4846	0.4750	0.4043	0.5290	0.5258	0.5335
DSST	0.5306	0.5794	0.6710	0.4834	0.5036	0.4060	0.5290	0.5245	0.5318
DAT	0.4608	0.4978	0.4350	0.4632	0.3998	0.4507	0.4512	0.4518	0.4687
KCF	0.4937	0.5496	0.6872	0.4291	0.4700	0.4301	0.5058	0.4916	0.4936

Table 2. Robustness of seven trackers evaluated on VOT2016. The bold fonts mean the best result.

	Camera Motion	Empty	Illum Change	Motion Change	Occlusion	Size Change	Mean	Weighted Mean	Pooled
MSLT	**15.000**	**5.000**	**1.000**	**17.000**	**10.000**	17.000	**10.833**	**11.673**	**43.000**
ECo	**15.000**	8.000	2.000	18.000	13.000	**10.000**	11.000	12.582	44.000
Staple	34.000	13.000	7.000	35.000	15.000	24.000	21.333	23.895	81.000
SRDCF	43.000	16.000	8.000	36.000	21.000	22.000	24.333	28.317	90.000
DSST	66.000	31.000	6.000	60.000	33.000	22.000	36.333	44.813	151.00
DAT	36.000	25.000	6.000	30.000	22.000	22.000	25.000	28.353	103.00
KCF	55.000	24.000	8.000	52.000	31.000	20.000	31.667	38.082	122.00

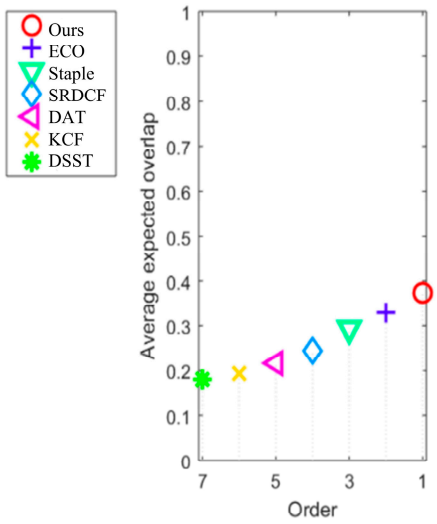

Figure 8. The EAO rank of the compared trackers. Different markers are used for different methods, and a higher EAO means a better performance.

Table 3. The EAO values of the compared trackers. The bold fonts mean the best result.

Method	All
MSLT	**0.3737**
CCOT	0.3293
Staple	0.2941
SRDCF	0.2458
DAT	0.2116
KCF	0.1935
DSST	0.1805

4.3. Qualitative Evaluation

In order to intuitively present the effectiveness of our tracker, we evaluated qualitatively eight trackers with several representative sequences. Figure 9 shows the comparisons of our tracker and seven state-of-the-art trackers: SRDCFdecon [44], STAPLE_CA [45], STAPLE [19], SAMF [24], DSST [23], KCF [10], and CSK [46] on six challenging sequences, which include Skiing, Couple, Bolt, Lemming, freeman1 and Tiger2.

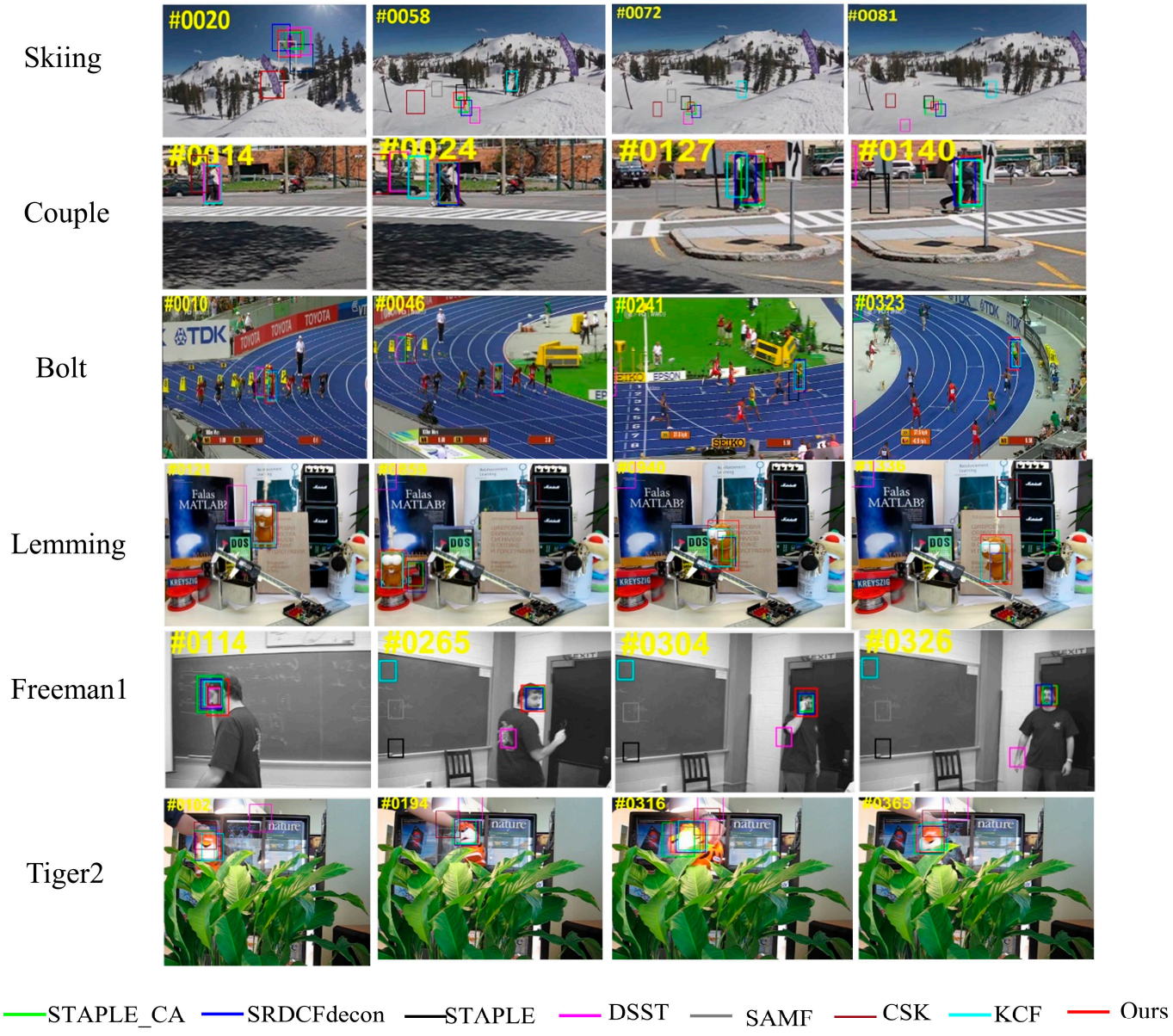

Figure 9. Tracking results of eight trackers on several representative sequences.

We can see that our MSLT tracker achieved good tracking performance compared with other trackers. Taking the Skiing sequence (Contains IV, SV, DEF, IPR and OPR contributions) as an example, most trackers tracked well in the twentieth frame, but only a few trackers could accurately track the object all the time, and our tracker performed well in handling fast motion, due to the redetection module. For the Lemming sequence (containing IV, SV, OCC, OPR, and OV contributions), our tracker performed well even in the 1336th frame, but the bounding boxes of SRDCFdecon, CSK and STAPLE_CA trackers were drifting in the 859th frame.

4.4. Ablation Study

In order to verify the effectiveness of the saliency redetection module, we compared the tracking performance of the proposed algorithm with a redetection module and without a redetection module, while its experimental conditions and parameters remained unchanged. For simplicity, we used the OTB2013 dataset [43] for testing and compared the tracking performance on both the overall and several contributions under the OPE protocol.

Figure 10 shows the overall precision plots and success plots of OPE with and without a redetection module. It can be seen that the tracker with a redetection module achieves better performance with a DP score of 0.898 and an AUC score of 0.678, which significantly improves by 5.9% and 6.4% compared with the performance of the tracker without a redetection module. Figure 11 shows the success rate plots of OPE with different video sequence attributions. We can see that the tracker with a redetection module performs better than the one without a redetection module when evaluated on all contributions.

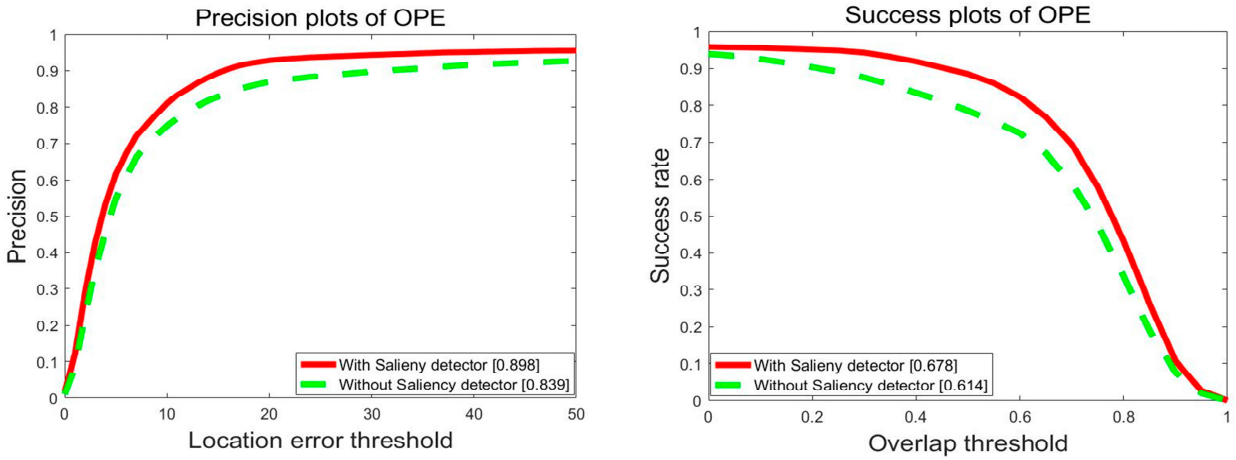

Figure 10. Overall precision plots and success plots of OPE with and without redetection module.

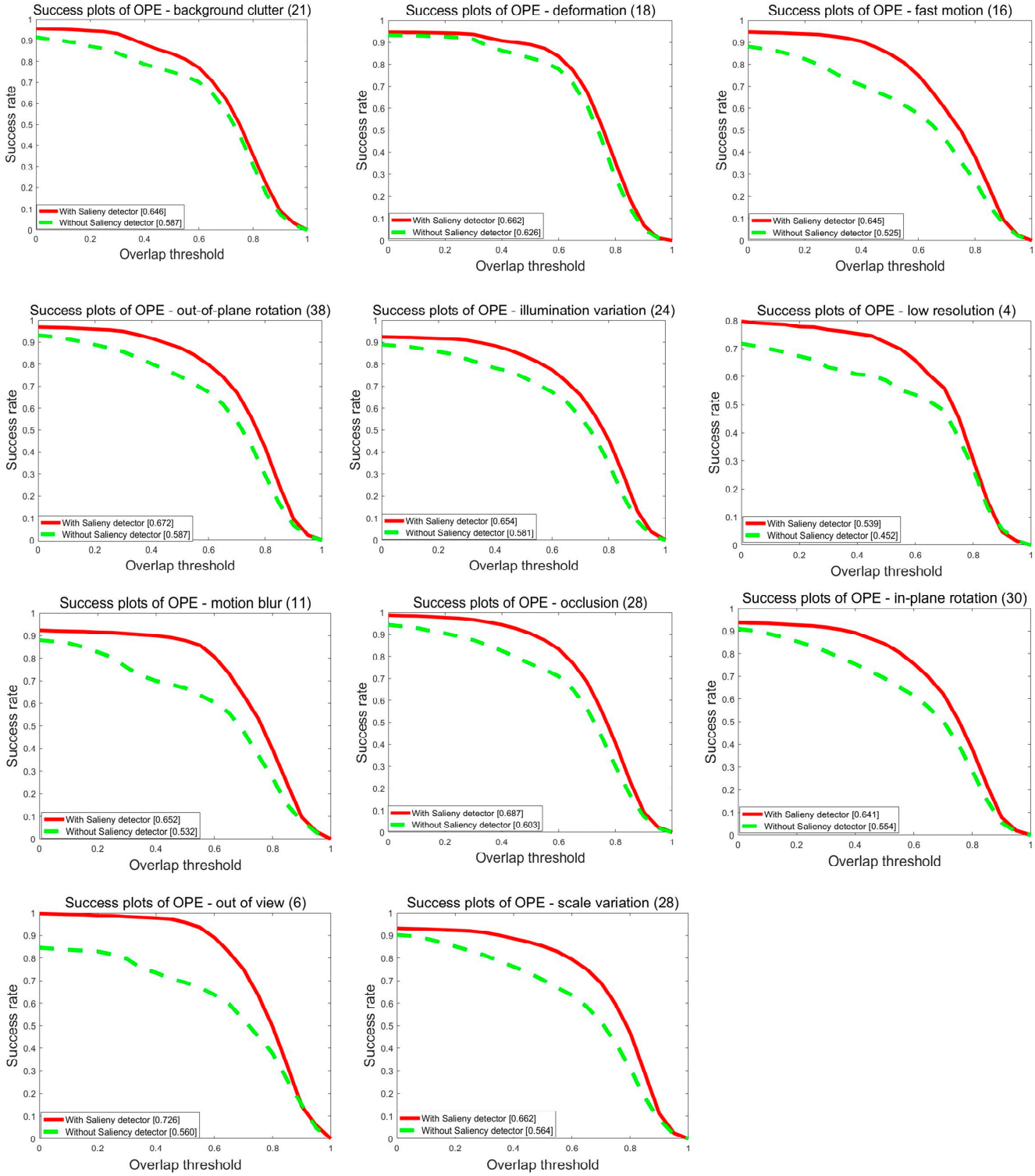

Figure 11. Success rate plots of OPE with different video sequence attributions.

5. Conclusions

This paper proposed a long-term tracker with multiple features and saliency redetection (MSLT). The MSLT tracker consists of two parts, a tracking-by-detection part and a redetection part. The tracking-by-detection part was built on the DCF framework by integrating with HOG and a color histogram fusion model, which was effective for some challenging situations, such as occlusion and color change. Meanwhile, the saliency redetection part could estimate the reliability of tracking result and redetect the tracked object with a saliency detection in a larger region if necessary. Compared with state-of-art trackers on two benchmarks, our MSLT method exhibited obvious advantages on most evaluation metrics. Furthermore, the ablation study indicated that the tracker with our redetection module performed better than without. In the future, we will attempt to introduce a deep learning method in the redetection module and utilize CNN features to improve the discrimination power of the correlation filter model. In addition, we will further explore the applications of object tracking for blurry and dark challenges, and some potential interdisciplinary applications, such as ambient technologies [49], industry 4.0, and so on.

Author Contributions: Conceptualization, T.F. and L.L.; methodology, T.F. and L.L.; software, L.L.; validation, T.F.; formal analysis, L.L.; investigation, L.L..; resources, L.L..; data curation, T.F.; writing—original draft preparation, L.L.; writing—review and editing, T.F., Y.F. and L.L.; visualization, L.L.; supervision, T.F. and Y.F.; project administration, L.L.; funding acquisition, Y.F. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Scientific research Fund of Xi'an Technological University, grant number No. 302020665. This research was supported by International Science and Technology Cooperation Program of Science and Technology Department of Shaanxi Province (No. 2021KW-07).

Data Availability Statement: All datasets evaluated in the paper can be found on official websites, OTB-100: http://cvlab.hanyang.ac.kr/tracker_benchmark/datasets.html accessed on 15 April 2022, VOT-2016: https://www.votchallenge.net/vot2016/ accessed on 15 April 2022.

Acknowledgments: The authors would like to thank the anonymous reviewers for their thoughtful comments and suggestions on the original manuscript.

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

References

1. Smeulders, A.W.; Chu, D.M.; Cucchiara, R.; Calderara, S.; Dehghan, A. Visual tracking: An experimental survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1442–1468. [PubMed]
2. Yilmaz, A.; Javed, O.; Shah, M. Object tracking: A survey. *ACM Comput. Surv.* **2006**, *38*, 13. [CrossRef]
3. Li, P.; Wang, D.; Wang, L.; Lu, H. Deep visual tracking: Review and experimental comparison. *Pattern Recognit.* **2018**, *76*, 323–338. [CrossRef]
4. Zhang, K.; Liu, Q.; Wu, Y.; Yang, M.-H. Robust Visual Tracking via Convolutional Networks without Training. *IEEE Trans. Image Process.* **2016**, *25*, 1779–1792. [CrossRef] [PubMed]
5. Jang, J.; Jiang, H. MeanShift++: Extremely Fast Mode-Seeking With Applications to Segmentation and Object Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 4100–4111. [CrossRef]
6. Sundararaman, R.; De Almeida Braga, C.; Marchand, E.; Pettré, J. Tracking Pedestrian Heads in Dense Crowd. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 3864–3874. [CrossRef]
7. Liu, L.; Cao, J. End-to-end learning interpolation for object tracking in low frame-rate video. *IET Image Process.* **2020**, *14*, 1066–1072. [CrossRef]
8. Chen, F.; Wang, X. Adaptive Spatial-Temporal Regularization for Correlation Filters Based Visual Object Tracking. *Symmetry* **2021**, *13*, 1665. [CrossRef]
9. Fawad; Khan, M.J.; Rahman, M.; Amin, Y.; Tenhunen, H. Low-Rank Multi-Channel Features for Robust Visual Object Tracking. *Symmetry* **2019**, *11*, 1155. [CrossRef]
10. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 583–596. [CrossRef] [PubMed]
11. Zhang, D.; Zhang, Z.; Zou, L.; Xie, Z.; He, F.; Wu, Y.; Tu, Z. Part-based visual tracking with spatially regularized correlation filters. *Vis. Comput.* **2019**, *36*, 509–527. [CrossRef]

12. Gong, L.; Wang, C. Research on Moving Target Tracking Based on FDRIG Optical Flow. *Symmetry* **2019**, *11*, 1122. [CrossRef]
13. Liu, T.; Wang, G.; Yang, Q. Real-time part-based visual tracking via adaptive correlation filters. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4902–4912. [CrossRef]
14. Liu, S.; Zhang, T.; Cao, X.; Xu, C. Structural Correlation Filter for Robust Visual Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4312–4320. [CrossRef]
15. Supancic, J.S.; Ramanan, D. Self-paced learning for long-term tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 2379–2386.
16. Lebeda, K.; Hadfield, S.; Matas, J.; Bowden, R. Long-term tracking through failure cases. In Proceedings of the IEEE International Conference on Computer Vision Workshops (CVPRW), Portland, OR, USA, 23–28 June 2013; pp. 153–160.
17. Lee, H.; Choi, S.; Kim, C. A memory model based on the siamese network for long-term tracking. In Proceedings of the European Conference on Computer Vision Workshops (ECCVW), Glasgow, UK, 8–14 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 100–115. [CrossRef]
18. Zhang, Y.; Wang, D.; Wang, L.; Qi, J.; Lu, H. Learning regression and verification networks for long-term visual tracking. *arXiv* **2018**, arXiv:1809.04320.
19. Bertinetto, L.; Valmadre, J.; Golodetz, S.; Miksik, O.; Torr, P.H.S. Staple: Complementary Learners for Real-Time Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1401–1409. [CrossRef]
20. Wu, Y.; Lim, J.; Yang, M.-H. Object Tracking Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1834–1848. [CrossRef] [PubMed]
21. Matej, K.; Ales, L.; Jiri, M.; Michael, F.; Roman, P.P.; Luka, C.; Tomas, V.; Gustav, H.; Alan, L.; Gustavo, F. The visual object tracking VOT2016 challenge results. In Proceedings of the European Conference on Computer Vision Workshops (ECCVW), Munich, Germany, 8 October 2016.
22. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550. [CrossRef]
23. Danelljan, M.; Häger, G.; Khan, F.S.; Michael, F. Accurate scale estimation for robust tracking. In Proceedings of the 2014 British Machine Vision Conference, Nottingham, UK, 1–5 September 2014; BMVA Press: Nottingham, UK, 2014; pp. 65.1–65.11. [CrossRef]
24. Li, Y.; Zhu, J. A scale adaptive kernel correlation filter tracker with feature integration. In Proceedings of the European Conference on Computer Vision Workshops (ECCVW), Zurich, Switzerland, 6–7 September 2014; pp. 254–265.
25. Danelljan, M.; Hager, G.; Khan, F.S.; Felsberg, M. Learning Spatially Regularized Correlation Filters for Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015; pp. 4310–4318. [CrossRef]
26. Galoogahi, H.K.; Fagg, A.; Lucey, S. Learning Background-Aware Correlation Filters for Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1144–1152. [CrossRef]
27. Danelljan, M.; Hager, G.; Khan, F.S.; Felsberg, M. Convolutional Features for Correlation Filter Based Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW), Santiago, Chile, 7–13 December 2015; pp. 621–629. [CrossRef]
28. Ma, C.; Huang, J.B.; Yang, X.; Yang, M.H. Hierarchical convolutional features for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3074–3082.
29. Danelljan, M.; Robinson, A.; Shahbaz, K.F.; Felsberg, M. ECO: Efficient Convolution Operators for Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6931–6939. [CrossRef]
30. Valmadre, J.; Bertinetto, L.; Henriques, J.; Vedaldi, A.; Torr, P.H.S. End-to-End Representation Learning for Correlation Filter Based Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5000–5008. [CrossRef]
31. Kalal, Z.; Mikolajczyk, K.; Matas, J. Tracking-learning-detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1409–1422. [CrossRef] [PubMed]
32. Wang, M.; Liu, Y.; Huang, Z. Large margin object tracking with circulant feature maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 21–26.
33. Zhang, J.; Ma, S.; Sclaroff, S. Meem: Robust tracking via multiple experts using entropy minimization. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 188–203.
34. Wei, Z.; Lu, H.; Yang, M.H. Robust object tracking via sparsity-based collaborative model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 1838–1845.
35. Zhang, T.; Bibi, A.; Ghanem, B. In defense of sparse tracking: Circulant sparse tracker. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3880–3888.

36. Hong, Z.; Chen, Z.; Wang, C.; Mei, X.; Prokhorov, D.; Tao, D. Multistore tracker (muster): A cognitive psychology inspired approach to object tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 749–758.
37. Ma, C.; Yang, X.; Zhang, C.Y.; Yang, M. Long-term correlation tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5388–5396. [CrossRef]
38. Fan, H.; Ling, H. Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5486–5494.
39. Tang, F.; Ling, Q. Contour-Aware Long-Term Tracking with Reliable Re-Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4739–4754. [CrossRef]
40. Wang, N.; Zhou, W.; Li, H. Reliable Re-Detection for Long-Term Tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 730–743. [CrossRef]
41. Wang, N.; Zhou, W.; Tian, Q.; Hong, R.; Wang, M.; Li, H. Multicue correlation filters for robust visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4844–4853.
42. Liu, L.; Cao, J.; Niu, Y. Visual Saliency Detection Based on Region Contrast and Guided Filter. In Proceedings of the 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, China, 8–11 September 2017; pp. 327–330. [CrossRef]
43. Wu, Y.; Lim, J.; Yang, M. Online Object Tracking: A Benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 2411–2418. [CrossRef]
44. Danelljan, M.; Hager, G.; Khan, F.S.; Felsberg, M. Adaptive Decontamination of the Training Set: A Unified Formulation for Discriminative Visual Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1430–1438. [CrossRef]
45. Mueller, M.; Smith, N.; Ghanem, B. Context-aware correlation filter tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1396–1404.
46. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 702–715. [CrossRef]
47. Possegger, H.; Mauthner, T.; Bischof, H. In Defense of Color-based Model-free Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2113–2120. [CrossRef]
48. Matej, K.; Jiri, M.; Alexs, L.; Tomas, V.; Roman, P.; Gustavo, F.; Georg, N.; Fatih, P.; Luka, C. A Novel Performance Evaluation Methodology for Single-Target Trackers. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2137–2155. [CrossRef]
49. Thakur, N.; Han, C.Y. An Ambient Intelligence-Based Human Behavior Monitoring Framework for Ubiquitous Environments. *Information* **2021**, *12*, 81. [CrossRef]