sensors

Article

Robust Visual Tracking Based on Adaptive Multi-Feature Fusion Using the Tracking Reliability Criterion

Lin Zhou [1], Han Wang [1,*], Yong Jin [1], Zhentao Hu [1], Qian Wei [1], Junwei Li [1] and Jifang Li [2]

[1] School of Computer and Information Engineering, Henan University, Kaifeng 475004, China; zhoulin@henu.edu.cn (L.Z.); jy@henu.edu.cn (Y.J.); hzt@henu.edu.cn (Z.H.); Weiqian@vip.henu.edu.cn (Q.W.); lijunwei@henu.edu.cn (J.L.)

[2] School of Electrical Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450000, China; lijifang@ncwu.edu.cn

* Correspondence: wanghan_henu@163.com

Received: 25 October 2020; Accepted: 12 December 2020; Published: 14 December 2020

Abstract: Multi-resolution feature fusion DCF (Discriminative Correlation Filter) methods have significantly advanced the object tracking performance. However, careless choice and fusion of sample features make the algorithm susceptible to interference, leading to tracking failure. Some trackers embed the re-detection module to remedy tracking failures, yet distinguishing ability and stability of the sample features are scarcely considered when training the detector, resulting in low effectiveness detection. Firstly, this paper proposes a criterion of feature tracking reliability and conduct a novel feature adaptive fusion framework. The feature tracking reliability criterion is proposed to evaluate the robustness and distinguishing ability of the sample features. Secondly, a re-detection module is proposed to further avoid tracking failures and increase the accuracy of target re-detection. The re-detection module consists of multiple SVM detectors trained by different sample features. When the tracking fails, the SVM detector trained by the most reliable sample feature will be activated to recover the target and adjust the target position. Finally, comparison experiments on OTB2015 and UAV123 databases demonstrate the accuracy and robustness of the proposed method.

Keywords: visual object tracking; correlation filter; feature tracking reliability criterion; adaptive feature fusion; multiple online detectors

1. Introduction

Visual single-object tracking is one of the fundamental problems in computer vision, and it involves multiple research fields such as signal processing, image processing and artificial intelligence. The task of visual object tracking is to continuously localize a target in a video sequence with given prior information such as initial location and scale of the target. Classical tracker is able to quickly and accurately localize the target only in the ideal scenario. However, some factors including deformation, occlusion and illumination variation, etc., caused by complex environment make visual object tracking challenging.

1.1. Related Work

In general, methods related to visual object tracking can be divided into two classes: generative methods [1–3] and discriminative methods [4–13]. Generative approaches often need to learn a model of target appearance and locate regions with maximum similarity to target appearance. In order to cope with various disturbances caused by complex environment, generative approaches must establish an effective appearance model, which means high computation complexity. Moreover, generative approaches usually ignore the background information of the target, making it hard to accurately separate the target from the background. To tackle this problem, discriminative adaptive tracking approaches model visual object tracking as a binary classification problem, and construct a classifier to locate target in a certain region of each frame in the video sequence.

In recent years, Discriminative Correlation Filter (DCF) has shown good performance on visual tracking benchmarks [14–16] and attracted extensive attention from researchers. DCF approaches perform a circular sliding window sampling operation to obtain training samples firstly, and then construct a ridge regression loss function on a set of training samples to design a filter. Moreover, by using Fast Fourier Transform (FFT), DCF maps the filter learning from spatial domain to Fourier domain and lightens the computation burden of convolution operation in spatial domain. Minimum Output Sum of Squared Error (MOSSE) [17] tracker proposed by Bolme et al. firstly employed the correlation filter in adaptive tracking and got 600 FPS tracking speed. However, other performance of MOSSE including robustness and accuracy are poor due to it using grayscale feature of images only. Boddeti et al. [18–20] extended the DCF framework to multi-channel feature scene, using high-dimensional features such as Histogram of Oriented Gradient(HOG) [21] and Color Names [22] to enhance accuracy and robustness of tracking. Kernelized correlation filter(KCF) [22] proposed by Henriques et al. used the HOG feature to improve both the accuracy and robustness of the tracker in the scene of motion blur and illumination variation. Danelljan et al. [8] employed the Color Names feature to enhance the robustness of tracker in situations such as deformation and occlusion induced by complex environment.

Since then, many researches on specific problems related to target tracking have been proposed, which has pushed the development of DCF methods. To relieve the negative impact caused by the scale change of target, the Scale Adaptive with Multiple Features(SAMF) [9] tracker proposed by Li et al. sampled the target with different scales and implements a multiple scale searching strategy at each frame. Danelljan et al. proposed a Discriminative Scale Space Tracking(DSST) [10] algorithm, which used a one-dimensional correlation filter independent of target displacement filter to predict the target scale change, further improving the efficiency of the scale estimation algorithm. All methods above employ cosine window technique to lighten boundary effect caused by FFT transform of periodic training samples, which makes it hard for trackers to learn background information with respective to the target, thus reducing the discrimination ability of a tracker. Danelljan et.al proposed the Spatially Regularized Discriminative Correlation Filter (SRDCF) [13] based on spatial regularization constraint, which used a predefined spatial weighting function to allocate more energy at the central region of a filter, thus enhancing the discrimination ability of the tracker.

As is known, some factors including deformation, occlusion, illumination variation and background clutter may make target appearance vary significantly, thus, deteriorating the tracking performance. To enhance the robustness of tracking, several works [23–25] accumulated the confidence scores from different parts of the target with part-based models and then used the parts that have high confidence scores to estimate the target's extent. However, the most widely used strategy in tracking is to establish a robust target representation by employing robust multi-category image features. For example, Staple [26] tracker used complementary features such as HOG and color histograms to train two independent filters respectively, and then combined the estimations of the two filters to predict the target's location. The Efficient Convolution Operators (ECO) [27] tracker efficiently integrated hand-crafted features and multi-resolution deep features, leading to superior target tracking results. Some works also investigated deep neural networks such as convolutional neural network (CNNs) [28] and Faster R-CNN [29] for object tracking. Ning et al. [30] utilized the YOLO [31] detector to generate initial object proposals. Paul et al.

proposed the Siam R-CNN [32] tracker, which is an adaptation of Faster R-CNN with a Siamese architecture. The Siam R-CNN tracker re-detects a template object anywhere in an image by determining if a region proposal is the same object as a template region, and regressing the bounding box for this object. However, the thousands of extracted deep feature channels which include irrelevant and redundant descriptors are not compact, leading to deteriorating of target detection performance.

Intuitively, using multiple features to train filter can improve robustness of target tracking. However, in some special tracking scenarios such as the Unmanned Aerial Vehicle (UAV) navigation [11,12], careless choice of features will make the tracker susceptible to interference. For example, in the scenario of violently changing illumination, using color feature to train the filter may lead to extremely unstable tracking model. Therefore, it is necessary to evaluate the tracking reliability of various features on specific tracking scenario, and then establish a reasonable feature fusion strategy to select the features with high discriminate capability to train the tracking model.

In addition, using multi-category image features to train trackers also increases the computation burden of systems. Some works [13,27,33] refined the model update strategies by reducing the frequency of tracker updating or compressing the space of training samples to maintain tracking efficiency. However, these strategies, which make the tracker lose the continuous information related to the appearance change of target and produce over-fitting to the current state of the target, deteriorate the tracking performance, more seriously, lead to tracking failure. Some trackers [34–36] rectified the target position by conducting an accurate dynamic model of the moving target, however, non-linear characteristic of the dynamic model is hard to describe. Several other methods, such as the LCT [37] tracker, used the re-detection module to retrieve the target from tracking failures. However, there are few of methods consider the distinguishing ability and stability of the sample features when training the detector, resulting in low effectiveness detection. Therefore, making use of the continuous information to reduce the probability of tracking failure and increasing the accuracy of re-detection is also a problem which is worth to be concerned.

1.2. Contributions

To solve the issues mentioned above, we propose a new feature adaptive fusion method in DCF framework. Moreover, we establish a re-detection module that consists of multiple detectors based on feature tracking reliability to re-detect the target in tracking failure scene, enhancing the robustness of tracking. The key innovations of the proposed method are listed as follows,

- We propose a criterion of feature tracking reliability and conduct a novel feature adaptive fusion method in filter learning. Different from traditional multi-resolution feature fusion trackers, the proposed method can adaptively assign greater weights to those features with high reliability and background distinguishing ability and vice versa, leading to a robust and accurate tracking.
- A re-detection module consisting of multiple SVM detectors trained by samples under different image feature maps is embedded to deal with tracking failures. Each detector in the module is labeled with the tracking reliability of its sample feature. The detector with maximum label in re-detector module would be activated to locate the lost target in the scene of tracking failures.

1.3. Paper Organization and Notation

The rest of this paper is organized as follows. Section 2 states the traditional multi-resolution feature fusion DCF framework. Section 3 introduces the proposed method in this paper. In the Section 3.1, we define the concept and evaluation method of feature tracking reliability criterion, and state the feature adaptive fusion scheme in filter learning formulation based on the feature tracking reliability. The re-detection module with multiple online detectors is described in Section 3.2. Section 4 provides an overview of the tracking algorithm. Finally, in Sections 5 and 6, we do experiments on OTB 2015 and

UAV123 datasets to compare the proposed algorithm with competing methods in details and conclude the paper.

In this paper, we use $\mathbb{R}$ to denote the set of real numbers, $\mathbb{C}$ to denote the set of complex numbers and $\mathbb{Z}$ to denote the set of integers. We use capital bold case, e.g., $\mathbf{A}$ to denote matrices. The direct sum and Kronecker product of matrices are denoted by $\oplus$ and $\otimes$, respectively. We use bold lower case, e.g., $\mathbf{x}$ to denote multi-channel signal and the dth channel of signal $\mathbf{x}$ is indicated as upper case, e.g., x^d. We use $\mathbf{x}^{\mathrm{T}}$ and $\mathbf{x}^{\mathrm{H}}$ to denote the transposition and Hermitian transposition of signal $\mathbf{x}$, respectively. The $L^2(T)$ is considered as the Hilbert space equipped with the inner product $\langle g(t), h(t)\rangle = \frac{1}{T}\int_0^T g(t-\tau)\, h(\tau)\, d\tau$. In $L^2(T)$, the circular convolution operation is defined as $g(t) * h(t) = \frac{1}{T}\int_0^T g(t-s)\, h(s)\, \mathrm{d}s$. For function $g \in L^2(t)$, we define its squared $\mathcal{L}^2$-norm as $\|g\|^2_{\mathcal{L}^2} = \frac{1}{T}\int_0^T |g(t)|^2 dt$ and the Fourier coefficients of g is indicated as $\widehat{g}[k] = \frac{1}{T}\int_0^T g(t)\, e^{-i\frac{2\pi}{T}kt} dt, k \in \mathbb{Z}$. The squared ℓ^2-norm of $\widehat{g}$ is defined as $\|\widehat{g}\|^2_{\ell^2} = \sum_{-\infty}^{\infty} |\widehat{g}[k]|^2$.

2. Traditional Multi-Resolution Features Fusion Filter

In filter training, the aim is to obtain a convolution filter $\mathbf{f}$ based on a set of training samples $\left\{\left(\mathbf{x}_j, \mathrm{y}_j\right)\right\}_{j=1}^{M}$ collected from frame 1 to M. Here, y_j is the label of training sample $\mathbf{x}_j$. For the sake of clarity, the filter learning framework in this paper is formulated on data defined in one-dimensional domain, which is displayed in Figure 1. The training sample $\mathbf{x} = (x_1, \ldots, x_L)$ contains L number of feature layers with different resolutions which are extracted from the same image patch. The lth feature layer x_l contains m_l channels, namely $x_l = [x_l^1, x_l^2, \cdots, x_l^{m_l}]$, and thus there are $D = \sum_{l=1}^{L} m_l$ feature channels in sample $\mathbf{x}$. We use N_l to denote the number of spatial sample points in the x_l^d channel of feature layer x_l, which means $x_l^d = [x_l^d(1), x_l^d(2), \cdots, x_l^d(N_l)]$. Thus, the sample space is expressed as $\mathbf{\emptyset} = \mathbb{R}^{m_1 \times N_1} \times \ldots \times \mathbb{R}^{m_L \times N_L}$.

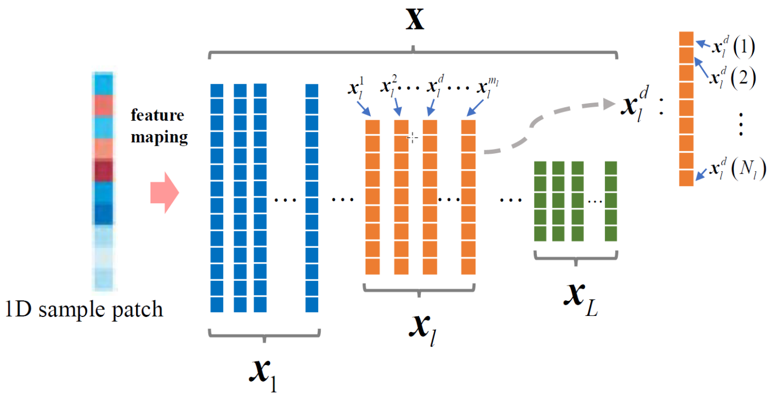

Figure 1. Visualization of the multi-resolution feature mapping in 1D domain.

To achieve sub-pixel level localization accuracy, we convert the learning problem from the discrete spatial domain to continuous spatial domain using an interpolation operator $J_l : \mathbb{R}^{N_l} \to L^2(T), T \in \mathbb{R}$. Specifically, for each feature channel $x_l^d \in \mathbb{R}^{N_l}$, in feature layer x_l, the interpolation operator is described as

$$J_l\left\{x_l^d\right\}(t) = \sum_{n=0}^{N_l-1} x_l^d[n]\, b\left(t - \frac{T}{N_l} n\right), d = 1, \ldots, m_l \tag{1}$$

where $x_l^d \in \mathbb{R}^{N_l},\ d = 1, \ldots, m_l$, and T represents the size of the support region. $J_l\left\{x_l^d\right\}$ is thus a continuous T-periodic real function. The interpolation function b_l is constructed using the standard cubic spline kernel [27],

$$b\left(t\right)=\begin{cases}\left(a+2\right)\left|t\right|^{3}-\left(a+3\right)t^{2}+1 & \left|t\right|\leqslant 1\\ a\left|t\right|^{3}-5at^{2}+8a\left|t\right|-4a & 1\leq\left|t\right|<2\\ 0 & \left|t\right|>2\end{cases} \tag{2}$$

where a is the shape parameter of b. Simultaneously, the label function is also converted into continuous spatial domain, which means $\mathrm{y}_j \in L^2\left(T\right)$.

In the traditional filter learning framework, the objective function of the filter $\mathbf{f}$ is described as,

$$\min E\left(\mathbf{f}\right)=\sum_{j=1}^{M}\alpha_j\left\|\sum_{l=1}^{L}\sum_{d=1}^{m_l}f_l^d * J_l\left\{x_l^d\right\}-\mathrm{y}_j\right\|_{\mathcal{L}^2}^2+\sum_{l=1}^{L}\sum_{d=1}^{m_l}\left\|\mathrm{w}f_l^d\right\|_{\mathcal{L}^2}^2 \tag{3}$$

by solving the Equation (3), we can obtain a multi-channel convolution operator $S_\mathbf{f}$ which is parameterized by a bank of convolution filters $\mathbf{f}=\left(f_1,...,f_L\right), f_l=\left(f_l^1,...,f_l^{m_l}\right)$. The operator $S_\mathbf{f}$ maps a sample $\mathbf{x}$ into a target confidence function that can be described as

$$S_\mathbf{f}\left\{\mathbf{x}\right\}\left(t\right)=\mathbf{f} * J\left\{\mathbf{x}\right\}=\sum_{l=1}^{L}\sum_{d=1}^{m_l}f_l^d * J_l\left\{x_l^d\right\}, \mathbf{x}\in\mathbf{\O} \tag{4}$$

The $S_\mathbf{f}\left\{\mathbf{x}\right\}\left(t\right)$ represents the confidence score of the target at the location $t \in [0, T)$ of the image. Similar to other discriminative methods, the target is localized at the position with the maximum confidence score in the image region.

It can be obviously seen in Equation (3) that there are L kinds of multi-resolution features employed in filter training to enhance the robustness of the tracker. However, the weights of these features are constant during the whole tracking process. As mentioned in Section 1, the rough feature fusion strategy makes the tracker susceptible to the interference in the scene of features changing violently.

3. The Proposed Method

To tackle the problems mentioned above, we propose a novel feature adaptive fusion strategy into tracker training framework in Section 3.1. The feature adaptive fusion strategy is formulated based on our criterion of feature tracking reliability, which is defined to evaluate the distinguishing ability and robustness of the features. In Section 3.2, we propose a re-detection module consisting of multiple SVM detectors trained by different sample features to deal with tracking failures. The re-detection module can further avoid the tracker losing the continuous information of the target appearance and producing over-fitting to the current state of the target.

3.1. Feature Adaptive Fusion Strategy in Filter Training

To tackle the problem mentioned at the end of the Section 2, this paper proposes a novel feature adaptive fusion strategy into filter training objective function as,

$$\min E\left(\mathbf{f}\right)=\sum_{j=1}^{M}\alpha_j\left\|\sum_{l=1}^{L}\sum_{d=1}^{m_l}f_l^d * r_l J_l\left\{x_l^d\right\}-\mathrm{y}_j\right\|_{\mathcal{L}^2}^2+\sum_{l=1}^{L}\sum_{d=1}^{m_l}\left\|\mathrm{w}f_l^d\right\|_{\mathcal{L}^2}^2 \tag{5}$$

In Equation (5) each kind of feature x_1 is weighted by its feature tracking reliability $r_l \in \mathbb{R}$, which is different from the traditional objective function described in Equation (3). The feature weight r_l is calculated by our "feature tracking reliability criterion" which is defined as,

$$r_l = \zeta \max \left(S_{\mathbf{f}}^{l} \{\mathbf{x}\} \right) = \zeta \max \left(\sum_{d=1}^{m_l} \boldsymbol{f}_l^d * J_l \left\{ \boldsymbol{x}_l^d \right\} \right) \tag{6}$$

where the $S_{\mathbf{f}}^{l} \{\mathbf{x}\} = \sum_{d=1}^{m_l} \boldsymbol{f}_l^d * J_l \left\{ \boldsymbol{x}_l^d \right\}$ is the filter output of the feature layer $\boldsymbol{x}_1$. The normalization scalar ζ ensures that $\sum_{l=1}^{L} r_l = 1$.

The feature tracking reliability criterion defined in Equation (6) is formulated based on the fact that the Equation (3) independently solves the least squares problem over all feature layers. Thus, the output of each feature layer $S_{\mathbf{f}}^{l} \{\mathbf{x}\}$ should nearly exactly fit the ideal response y. On the other hand, the response is highly noisy on the feature layers with low discriminative power. The existing noise leads to significant global error reduction of least squares, moreover, reduces the maximal output of the response related to the feature layers [38]. Therefore, robustness and background distinguishing ability of the lth sample feature can be evaluated by the maximum output response of $\boldsymbol{x}_1$. Thus, according to the feature tracking reliability criterion defined in Equation (6), the proposed method can adaptively assign greater weights to those features with high reliability and background distinction. The features in Equation (5) are suitably fused to enhance the robustness of our tracking model.

In practice, the filter in Equation (5) can be efficiently solved by using FFT transform. However, in order to further lighten the computation burden, we dispose the features with dimensionality reduction technique. For feature layer $\boldsymbol{x}_l$, suppose there is a $m_l \times c_l$, $(m_l > c_l)$ matrix $\mathbf{P}_l = \left\{ p_{d,c} \right\}, d = 1, ..., m_l; c = 1, ..., c_l$, which makes $S_{\mathbf{f}}^{l} \{\mathbf{x}\} = S_{\mathbf{Pf}}^{l} \{\mathbf{x}\} = \mathbf{P}_l \boldsymbol{f}_l * J_l \{\boldsymbol{x}_l\} = \sum_{d,c} p_{d,c} \boldsymbol{f}_l^c * J_l \left\{ \boldsymbol{x}_l^d \right\} = \boldsymbol{f}_l * \mathbf{P}_l^{\mathrm{T}} J_l \{\boldsymbol{x}_l\}$, i.e., each filter $\boldsymbol{f}_l^1, ..., \boldsymbol{f}_l^{m_l}$ can be linearly combined by a smaller set of basis filters $\boldsymbol{f}_l^1, ..., \boldsymbol{f}_l^{c_l}$. The filter learning objective function in Fourier domain is finally derived as

$$\min E\left(\mathbf{P}, \mathbf{f}\right) = \sum_{j=1}^{M} \alpha_j \left\| \sum_{l=1}^{L} r_l \widehat{\mathbf{z}}_l^{\mathrm{T}} \mathbf{P}_l \widehat{\boldsymbol{f}}_l - \widehat{\mathrm{y}}_j \right\|_{\ell 2}^{2} + \sum_{l=1}^{L} \sum_{c=1}^{c_l} \left\| \widehat{\mathrm{w}} * \widehat{\boldsymbol{f}}_l^c \right\|_{\ell 2}^{2} + \sum_{l=1}^{L} \lambda \left\| \mathbf{P}_l \right\|_F^2 \tag{7}$$

here, we use $\widehat{\mathbf{z}}_l [k] = X_l^d [k] \, \widehat{b_l} [k]$ to denote the Fourier coefficients related to the feature map $\mathbf{z}_l = J_l \{\boldsymbol{x}_l\}$ in the lth feature layer. The $X_l^d = \sum_{n=0}^{N_l} \boldsymbol{x}_l^d [n] \, e^{-i \frac{2\pi}{N_l} nk}, k \in \mathbb{Z}$ is the Discrete Fourier Transform (DFT) of $\boldsymbol{x}_l^d [n]$. It should be noted that the matrix $\mathbf{P}_l$ is initialized on the first frame by operating the principal component analysis (PCA) on the feature $\boldsymbol{x}_l$. We update the $\mathbf{P}_l$ according the sample features extracted in current frame. To ensure the stability of the target model, the Frobenius norm of matrix $\mathbf{P}_l$ controlled by the weight parameter λ is added as a regular term to limit the updating of $\mathbf{P}_l$.

In conclusion, according to the feature tracking reliability criterion defined in Equation (6), our method can adaptively fuse the sample features used in filter training in Equation (5). With the FFT transform and dimensionality reduction technique, the Equation (5) can be transformed into Equation (7), and the filter can be efficiently learned and updated in Fourier domain by solving Equation (7) using Gauss-Newton and Conjugate Gradient method. In the target detection stage, the output response of the tracker thus can be calculated as

$$S_{\mathbf{f}} \{\mathbf{x}\} = \sum_{l=1}^{L} r_l S_{\mathbf{f}}^{l} \{\mathbf{x}\} = \sum_{l=1}^{L} r_l \boldsymbol{f}_l * \mathbf{P}_l^{\mathrm{T}} J_l \{\boldsymbol{x}_l\}, \mathbf{x} \in \boldsymbol{\emptyset} \tag{8}$$

and the target is localized at the position with the maximum confidence score in the image region.

3.2. Multiple Online Detectors Based on Feature Tracking Reliability

Employing a Gaussian Mixture Model [27], we manage the training samples collected in each frame to construct a compact sample space with N_s training samples. This operation can eliminate the redundant information among the training samples. In addition, we update the filter once every N_uth frame to further improve tracking efficiency. However, as mentioned in Section 1, above strategies may bring about some continuous information loss related to the appearance change of target.

To tackle this problem, we additionally train a long-term filter $S_{\mathrm{L}}\left\{\mathbf{x}\right\}$ with DCF framework to record and track the appearance change of the target firstly. The maximum response of the long-term filter $A_{\mathrm{L}} = \max S_{\mathrm{L}}\left\{\mathbf{x}\right\}$, which is obtained at the estimated position in each frame is to be solved to detect tracking failure. In the case of $A_{\mathrm{L}} = \max S_{\mathrm{L}}\left\{\mathbf{x}\right\} < T_r$, where the T_r is a constant threshold, it is believed that there is a tracking failure, then a detection module will be activated to re-detect the target and revise the tracking result.

Different from traditional detection methods, this paper builds multiple SVM detectors $[\boldsymbol{h}_1, ..., \boldsymbol{h}_L]$ using training samples under different image feature maps independently. Each detector $\boldsymbol{h}_l$ is labeled with the feature tracking reliability r_l related to the feature map of its train samples. When the tracking fails, we use the SVM detector with the maximum label to redetect the target.

Given a training set $\left\{\left(\mathbf{v}_i, b_i\right), i = 1, ..., N, b_i \in \left\{+1, -1\right\}\right\}$ in a frame, The objective function related to multiple SVM detectors is

$$\min_{\mathbf{h}} \frac{1}{2}\|\mathbf{h}\|^2 + \frac{1}{N}\sum_i \mathcal{L}\left(\mathbf{h}; \left(\mathbf{v}_i, b_i\right)\right) \tag{9}$$

where the loss function can be defined as $\mathcal{L}\left(\mathbf{h}; \left(\mathbf{v}, b\right)\right) = \max\left\{0, 1 - b\left\langle\mathbf{h}, \mathbf{v}\right\rangle\right\}$. We can efficiently update $\mathbf{h}$ by

$$\mathbf{h} \leftarrow \mathbf{h} - \frac{\mathcal{L}\left(\mathbf{h}; \left(\mathbf{v}, b\right)\right)}{\|\nabla_{\mathbf{h}}\mathcal{L}\left(\mathbf{h}; \left(\mathbf{v}, b\right)\right)\|^2 + \frac{1}{2\tau}} \nabla_{\mathbf{h}}\mathcal{L}\left(\mathbf{h}; \left(\mathbf{v}, b\right)\right) \tag{10}$$

here, the $\nabla_{\mathbf{h}}\mathcal{L}\left(\mathbf{h}; \left(\mathbf{v}, b\right)\right)$ is the gradient of the loss function in terms of $\mathbf{h}$ and the $\tau \in (0, +\infty)$ is a hyper-parameter that controls the update rate of $\mathbf{h}$. It should be noted that in Equations (9) and (10), the training sample $\mathbf{v} = [v_1, ..., v_L]$ consists of L number of feature layers extracted from the sample patch. The $\mathbf{h} = (\boldsymbol{h}_1, ..., \boldsymbol{h}_L)$ denotes a hyperplane set with respect to L number of SVM detectors. Each SVM detector $\boldsymbol{h}_l$ in $\mathbf{h}$ is labeled by the feature tracking reliability r_l of its sample feature. Only the detector with the maximum label will be activated to redetect the target when tracking failure occurs.

4. Outline of the Proposed Method

This paper presents an overview of the proposed method in Figure 2. It can be seen that the whole tracking process of the proposed method can be divided into five parts, i.e., the Translation estimation part which is described in Equation (9) in Section 3.1, the Tracking failure detection and Target re-detection parts described in Section 3.2, the Feature tracking reliability evaluation part described in Equation (6) and the Scale estimation part. It should be noted that in the Scale estimation part, we design an 1-dimensional scale filter S_s to estimate the scale of target using the same way as DSST [10]. We also design multi-resolution feature maps related to samples for filter S_s and assign weights of these features according to their feature tracking reliability r_l. More details of the tracking process in Figure 2 are discussed in the following Table 1.

In Figure 2 and Table 1, the constant T_r represents the threshold of starting the detection module, and the constant T_a is the threshold of adopting the re-detection result. Meanwhile, the stability threshold T_s in Table 1 is introduced to determine the filter S_{L} and detectors updating. Specifically, S_{L} and multiple SVM detectors are updated only when the maximum response of $S_{\mathrm{L}}\left\{\mathbf{x}\right\}$ is greater than T_s. Some same resolution features including gradient-based features (HOG) and intensity-based features (HOI) are employed to train Long term filter. The feature tracking reliability r_l and filters $S_{\mathbf{f}}$, S_s and S_{L} are updated by a moving average scheme.

Table 1. Outline of the proposed tracking method.

```
Input: The target position $p_1$ and scale $s_1$ in frame 1;
Output: The target position $p_t$ and scale $s_t$ in frame $t$.

Initialization:
  Crop out the image patch centered at $p_1$ and extract multi-category features $\mathbf{x}_1$ in frame 1;
  Initialize the filter $S_\mathbf{f}$ using Equation (6) and set $r_l = 1$ in frame 1;
  Initialize the filter $S_\mathrm{L}$ and $S_s$;
  Initialize the SVM detectors $\mathbf{h}$ using Equation (9);
for $t = 2, t = t + 1, t \leqslant Num$ // Num is the number of frames in the video
  Crop out the image patch centered at $p_{t-1}$ and extract multi-category features $\mathbf{x}$ in frame $t$;
  // Translation estimation
  Estimate the target position $p_t$ in frame $t$ using Equation (8);
  // Tracking failure detection
  Compute the maximum response $A_\mathrm{L} = \max S_\mathrm{L}\{\mathbf{x}\}$ of the filter $S_\mathrm{L}$ at position $p_t$;
  if $A_\mathrm{L} = \max S_\mathrm{L}\{\mathbf{x}\} < T_r$
    //Target re-detection
    Activate detection module and return the candidate position $p_{tmp}$;
    Compute the maximum response $A_{\mathrm{L}\_tmp}$ of the sample extracted at position $p_{tmp}$ on the filter $S_\mathrm{L}$;
    if $A_{\mathrm{L}\_tmp} > T_a$
      Rectify the target position $p_t = p_{tmp}$;
    elseif
      Maintain the target position $p_t$ and discard the candidate position $p_{tmp}$;
    end if
  elseif
  The tracking failure doesn't occur;
  end if
  Output: The target position $p_t$.
  // Scale estimation
  Construct scale pyramid centered at $p_t$ in frame $t$ and estimate $s_t$ using filter $S_s$;
  Output: The targe scale $s_t$ in frame $t$.
  // Model update
  Crop out the image patch centered at $p_t$ in frame $t$ and extract multi-category features $\mathbf{x}$;
  if $t/N_u == 0$
    Update the feature tracking reliability $r_l$ using Equation (6); // Feature tracking reliability evaluation
    Update the filter $S_\mathbf{f}$ using Equation (7);
    Update the scale filter $S_s$;
  end if
  if $A_\mathrm{L} > T_\mathrm{s}$
    Update the long-term filter $S_\mathrm{L}$ and update multiple SVM detectors $\mathbf{h}$ using Equation (10);
  end if
end for
```

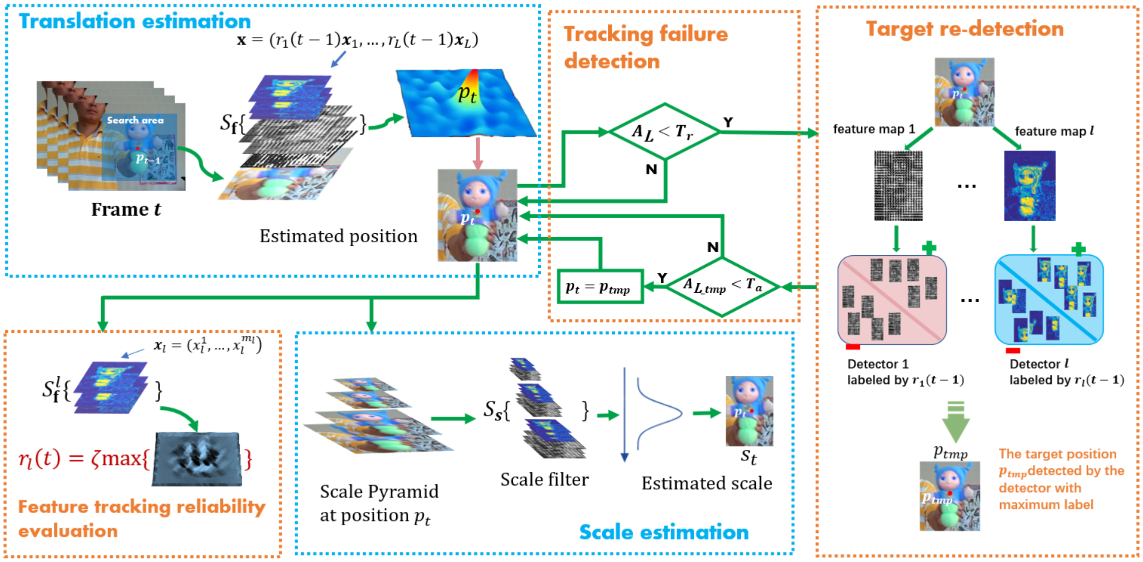
Translation estimation
$\mathbf{x} = (r_1(t-1)\boldsymbol{x}_1, \dots, r_L(t-1)\boldsymbol{x}_L)$
Search area
p_{t-1}
Frame t
$S_{\mathbf{f}}\{$ $\}$
p_t
Estimated position
Tracking failure detection
$A_L < T_r$
Y
N
N
$p_t = p_{tmp}$
Y
$A_{L,tmp} < T_a$
Target re-detection
feature map 1
feature map l
...
Detector 1 labeled by $r_1(t-1)$
Detector l labeled by $r_l(t-1)$
p_{tmp}
The target position p_{tmp} detected by the detector with maximum label
$\boldsymbol{x}_l = (x_l^1, \dots, x_l^{m_l})$
$S_{\mathbf{f}}^l\{$ $\}$
$r_l(t) = \zeta\max\{$ $\}$
Feature tracking reliability evaluation
Scale Pyramid at position p_t
$S_s\{$ $\}$
Scale filter
Estimated scale
S_t
Scale estimation

Figure 2. Overview of the proposed algorithm.

When updating the filter $S_{\mathbf{f}}$ using Equation (7), we independently employ the Gauss-Newton method on each feature layer. For simplicity, we consider learning the filter $\boldsymbol{f}_l$ of the feature layer $\boldsymbol{x}_l$ from a single training sample $\mathbf{x}$. The corresponding loss function in Equation (7) is derived as

$$\min E\left(\mathbf{P}_l \boldsymbol{f}_l\right) = \left\| r_l \widehat{\mathbf{z}}_l^{\mathrm{T}} \mathbf{P}_l \widehat{\boldsymbol{f}}_l - \widehat{y}_j \right\|_{\ell^2}^2 + \sum_{c=1}^{c_l} \left\| \widehat{w} * \widehat{\boldsymbol{f}}_l^c \right\|_{\ell^2}^2 + \lambda \left\|\mathbf{P}_l\right\|_F^2 \tag{11}$$

The Gauss-Newton method is derived by linearizing the residuals in Equation (11) using a first order Taylor series expansion, which approximates the bi-linear term $\widehat{\mathbf{z}}_l^{\mathrm{T}} \mathbf{P}_l \widehat{\boldsymbol{f}}_l$ around the current estimate $\left(\widehat{\boldsymbol{f}}_{l(i)}, \mathbf{P}_{l(i)}\right)$ at iteration i as

$$\begin{aligned} &\widehat{\mathbf{z}}_l^{\mathrm{T}} \left(\mathbf{P}_{l(i)} + \Delta \mathbf{P}_l\right) \left(\widehat{\boldsymbol{f}}_{l(i)} + \Delta \widehat{\boldsymbol{f}}_l\right) \\ &\approx \widehat{\mathbf{z}}_l^{\mathrm{T}} \mathbf{P}_{l(i)} \widehat{\boldsymbol{f}}_{l(i),\Delta} + \widehat{\mathbf{z}}_l^{\mathrm{T}} \Delta \mathbf{P}_l \widehat{\boldsymbol{f}}_{l(i)} = \widehat{\mathbf{z}}_l^{\mathrm{T}} \mathbf{P}_{l(i)} \widehat{\boldsymbol{f}}_{l(i),\Delta} + \left(\widehat{\boldsymbol{f}}_{l(i)} \otimes \widehat{\mathbf{z}}_l\right)^{\mathrm{T}} \operatorname{vec}\left(\Delta \mathbf{P}_l\right) \end{aligned} \tag{12}$$

here, the "vec" represents the vectorization operation of the matrix, we define $\widehat{\boldsymbol{f}}_{l(i),\Delta} = \widehat{\boldsymbol{f}}_{l(i)} + \Delta \widehat{\boldsymbol{f}}_l$, the Gauss-Newton subproblem at iteration is derived by substituting the first-order approximation Equation (12) into Equation (11)

$$\begin{aligned} &\min E\left(\Delta \mathbf{P}_l, \widehat{\boldsymbol{f}}_{l(i),\Delta}\right) \\ &= \left\| r_l \left(\widehat{\mathbf{z}}_l^{\mathrm{T}} \mathbf{P}_{l(i)} \widehat{\boldsymbol{f}}_{l(i),\Delta} + \left(\widehat{\boldsymbol{f}}_{l(i)} \otimes \widehat{\mathbf{z}}_l\right)^{\mathrm{T}} \operatorname{vec}\left(\Delta \mathbf{P}_l\right) \right) - \widehat{y}_j \right\|_{\ell^2}^2 \\ &+ \sum_{c=1}^{c_l} \left\| \widehat{w} * \widehat{\boldsymbol{f}}_{l(i),\Delta}^c \right\|_{\ell^2}^2 + \lambda \left\| \mathbf{P}_{l(i)} + \Delta \mathbf{P}_l \right\|_F^2 \end{aligned} \tag{13}$$

Since the filter $\boldsymbol{f}_l$ is constrained to have finite non-zero Fourier coefficients, Equation (13) is a linear least squares problem. We employ the Conjugate Gradient method to optimize the Gauss-Newton subproblem at each iteration.

5. Experiments and Analysis

Performances related to the proposed method are verified by comparison experiments on OTB2015 (http://cvlab.hanyang.ac.kr/tracker_benchmark/datasets.html) and UAV123 (https://cemse.kaust.edu.sa/ivul/uav123) datasets. As we know, OTB2015 [15] dataset is the most popular tracking benchmark with 100 video sequences which are fully annotated with 11 different attributes. The UAV123 [16] dataset contains a total of 123 video sequences from an aerial viewpoint. To demonstrate the performance of proposed tracker, some state-of-the-art trackers, including ECO [27], CSR_DCF [27], SRDCF [13], Staple [26], DSST [10] and LCT [34] are used to compare with our tracker. It should be noted that only single-object tracking task is considered in the comparison of this paper. For a fair comparison, the performances of the 7 trackers above are compared under the same environment conditions using MATLAB2016b equipped with Windows 10-64bit on Intel(R) Core (TM) i5-9300H CPU and 8GB RAM.

5.1. Experimental Parameters

The proposed method uses two hand-crafted image feature layers, namely Color Names which reflects the color information of the target and HOG layer which reflects the spatial structure and texture characteristics of the target, to demonstrate the performance of the proposed method. The experimental parameters are described in Table 2. The parameters related to the long term filter S_L and the scale filter S_s are selected referring to the LCT [34] and DSST [10] trackers. The parameters in Table 2 are chosen based on the ECO [27] and CSR_DCF [35] trackers and fine-tuned according to the tracking AUC performance.

Table 2. Parameters of Experimental.

Name	Symbol	Value
Shape parameter of interpolation function	a	-0.75
Threshold of starting the detection module	T_r	0.15
Threshold of updating the Long-term filter	T_r	0.38
Threshold of adopting the re-detected result	T_a	1.5 A_L
Number of samples in sample space	N_s	50
Updating Period of tracker $S_\mathbf{f}$	N_u	5
Regularization parameter in Equation (12)	λ	2×10^{-7}
Hyper-parameter of SVM detector	τ	1

5.2. Evaluation Indicators

This paper uses One Pass Evaluation (OPE) criterion including center location error and the bounding box overlap score to evaluate the performance of trackers. We also employ the Success plot, Mean Distance Precision (Mean DP), Mean Overlap Precision (Mean OP), average center location error (CLE) and area-under-curve (AUC) as the expressions of the experiment results. It should be noted that given a estimated bounding box ROI_e and the ground-truth bounding box ROI_g of the target, the bounding box overlap score is defined as

$$IOU = \frac{area\left(ROI_e \cap ROI_g\right)}{area\left(ROI_e \cup ROI_g\right)} \tag{14}$$

Since involving the position and scale related to tracked target simultaneously, IOU is an indicator different with the center location error, and it can widely be used to evaluate the robustness and accuracy of the tracking algorithm.

5.3. Comparisons and Analysis

5.3.1. Impact of the Feature Adaptive Fusion

In the section, we firstly evaluate the effect of feature adaptive fusion based on the feature tracking reliability criterion in our method. The proposed method can adaptively evaluate the tracking reliability of each feature in the target model according to the Equation (6). The tracking reliability of each feature is then used in Equations (5) and (7) as its feature adaptive fusion weight in constructing the target model. Thus, our method can adaptively increase the weights of feature with more stability, robustness and distinction while reduce the weights of the features with poor stability, thereby leading to a robust and accurate tracking. To evaluate the effect of the feature adaptive fusion proposed in this paper, we demonstrate the tracking process on video sequence named "Box" in Figure 3. We present the tracking bounding box comparison of the proposed method with the comparing trackers at the top of the Figure 3, and the change of sample feature tracking reliability calculated using Equation (6) in our method is illustrated at the bottom of the Figure 3.

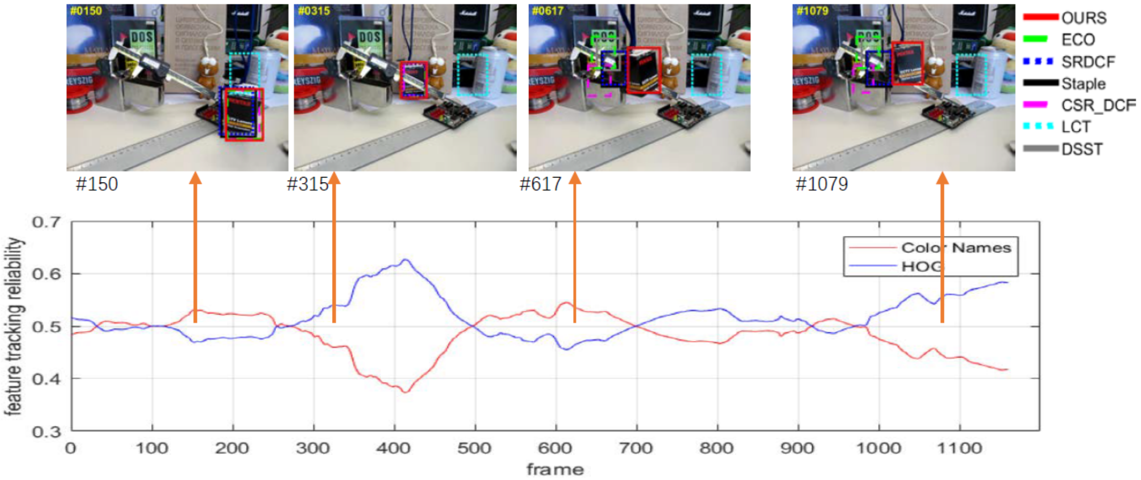

Figure 3. Visualization of the feature weights change in adaptive feature fusion.

As shown in Figure 3, in the 151st frame, due to the non-planar rotation of the target, the HOG feature of the target fluctuates greatly, leading to locating failure of the target, while the color information of the target is still stable. The proposed algorithm can adaptively increase the feature weight of the Color Names and stably track the target; In the 315th frame, the target is partially occluded, the feature is highly noisy in the color information of the target. The algorithm can adaptively increase the weight of the HOG feature and use the structure information to track the target stably; In the 617th frame, when the target appears again from the full occlusion, although the rectangular structure of the target has changed greatly due to the rotation, the target can still be detected by the detector of our method trained with the relatively stable color features; After the 1000th frame, the target's spatial structure tends to be stable, the algorithm thus adaptively increases the weight of the HOG feature to track the target.

5.3.2. Baseline Comparison

In this section, we report the performances of the mean overlap precision (Mean OP), mean distance precision (Mean DP), center location error (CLE) and Success plot related to all the methods on OTB 2015 and UAV123 datasets. It should be noted that the overlap precision score (OP) is defined as the ratio of frames in a video which the IOU is greater than a certain threshold $op(op = 0.5)$. The distance precision score (DP) is defined as the ratio of frames in a video where the Euclidean distance between the tracking output and ground truth is smaller than a threshold $dp(dp = 20(pixel))$.

We present the Mean OP, Mean DP and CLE comparisons of the trackers respectively in the histograms of Figures 4–6, and the best two results are highlighted in red and blue fonts. It can be seen that the proposed method gets the highest Mean OP in Figure 4 and the highest Mean DP in Figure 5 on both OTB2015 and UAV123 datasets. In Figure 6, our method gets the lowest CLE scores. Compared with the second-best method, we can find that our method achieves a gain of 2.3% in Figure 4 of the Mean OP, 1.9% in Figure 5 of the Mean DP, and 2.78 pixels in Figure 6 of the CLE on OTB2015 dataset and a gain of 0.6% in Figure 4 of the Mean OP, 1.1% in Figure 5 of the Mean DP, and 43 pixels in Figure 6 of the CLE on the UAV123 dataset, respectively.

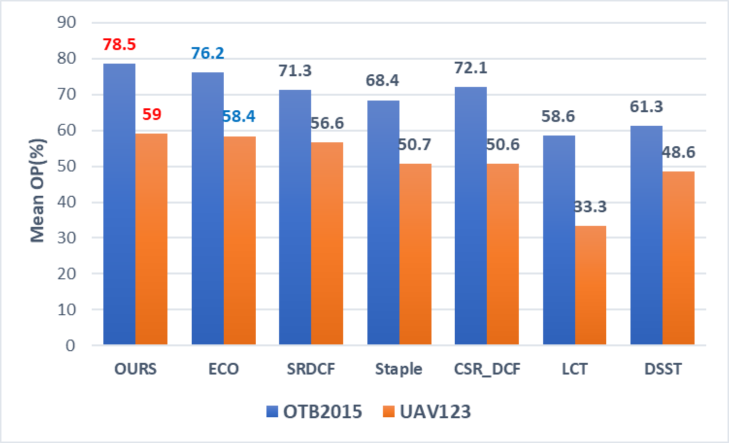

Figure 4. Comparison of Mean overlap precision score (OP) on OTB 2015 and UAV123 datasets.

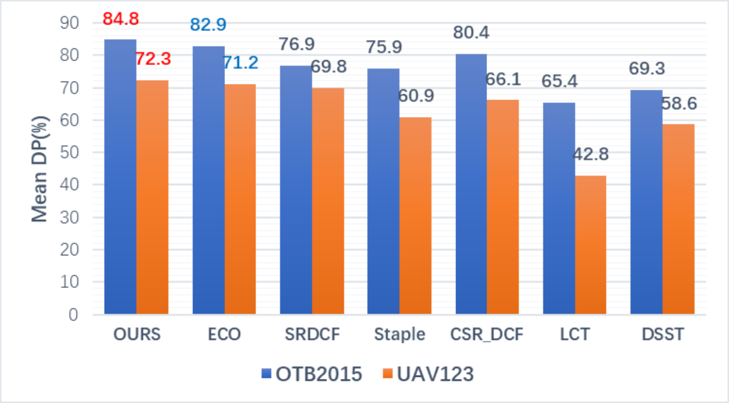

Figure 5. Comparison of Mean distance precision score (DP) on OTB 2015 and UAV123 datasets.

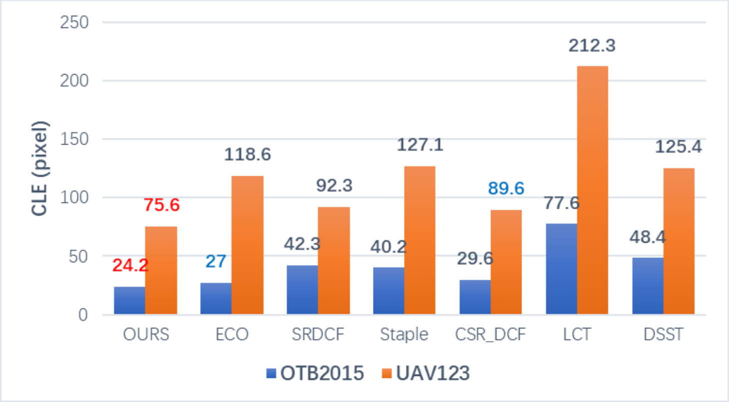

Figure 6. Comparison of center location error (CLE) on OTB 2015 and UAV123 datasets.

The Success plots and the AUC scores of all the trackers on OTB2015 and UAV123 datasets are shown in Figure 7 and 8, respectively. On OTB2015 dataset illustrated in Figure 7, the ECO tracker achieves an AUC score of 62.6%, while our method achieves an AUC score of 63.7%, 1.1% higher than that of the second-best ECO tracker. Meanwhile, on UAV123 dataset illustrated in Figure 8, our method obtains an AUC score of 49.9%, 0.2% higher than that of the second-best method.

Since considering feature tracking reliability in filter learning, our method can adaptively select features with high discrimination for tracking model training. Moreover, a re-detection module is introduced in tracking process to re-detect the target and to align the estimated position in the case of tracking failure. So, our method shows the best performance on the OTB 2015 and UAV123 datasets with the metrics including Mean OP, Mean DP, CLE and Success plot.

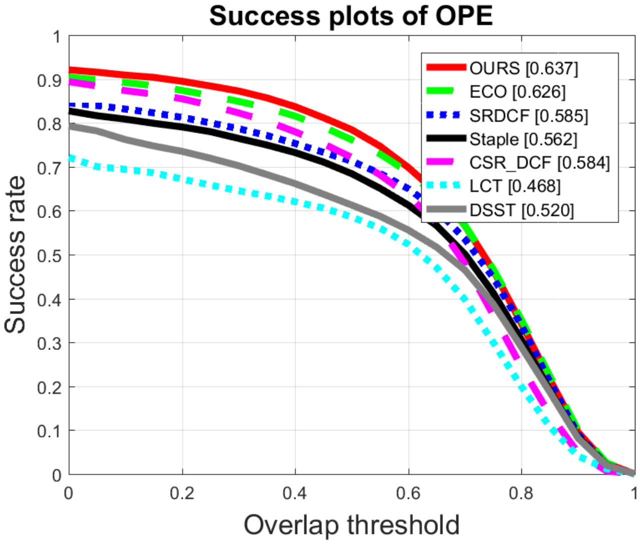

Figure 7. Comparison of Success plots with 7 trackers on OTB2015.

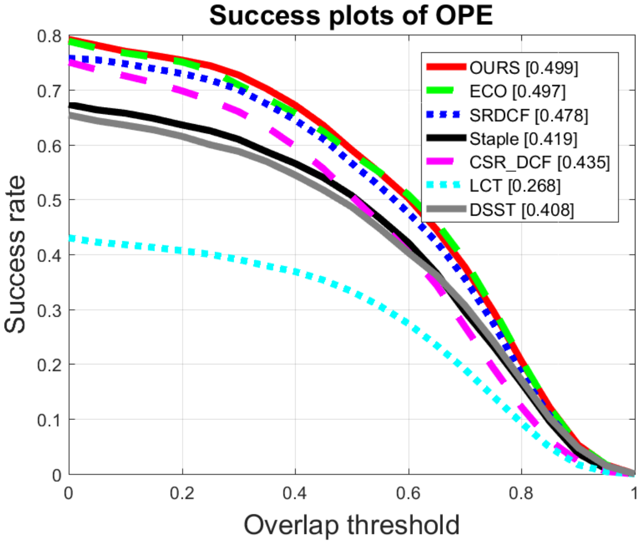

Figure 8. Comparison of Success plots with 7 trackers on UAV123.

5.3.3. Attribute-Based Comparison

We further perform an attribute-based analysis of all the methods respectively on the OTB 2015 and UAV 123 datasets. In OTB2015 dataset, all the sequences are annotated with 11 different attributes. As shown on the 11 attribute axes in Figure 9, the 11 video attributes are: Fast Motion, Background Clutter, Motion Blur, Deformation, Illumination Variation, In-Plane Rotation, Low Resolution, Occlusion, Out-of-Plane Rotation, Out of View and Scale Variation.

Figure 9 shows the AUC score comparisons of the proposed method with other trackers on all 11 attributes in OTB 2015. On each visual attribute axe, the AUC scores of trackers are arranged from the center of the figure to the edge in order from small to large. The AUC scores of the best two trackers are displayed behind the axe labels. It can be seen from the Figure 9 that the proposed method has superior AUC scores on the above 11 attributes and outperforms the other 6 trackers on 10 attributes except the Fast Motion. The advantages of our method are clearly embodied in four attribute axes, namely Background Clutter, Illumination Variation, Motion Blur and Out of View. The Success plots on these four attributes are illustrated in Figure 10.

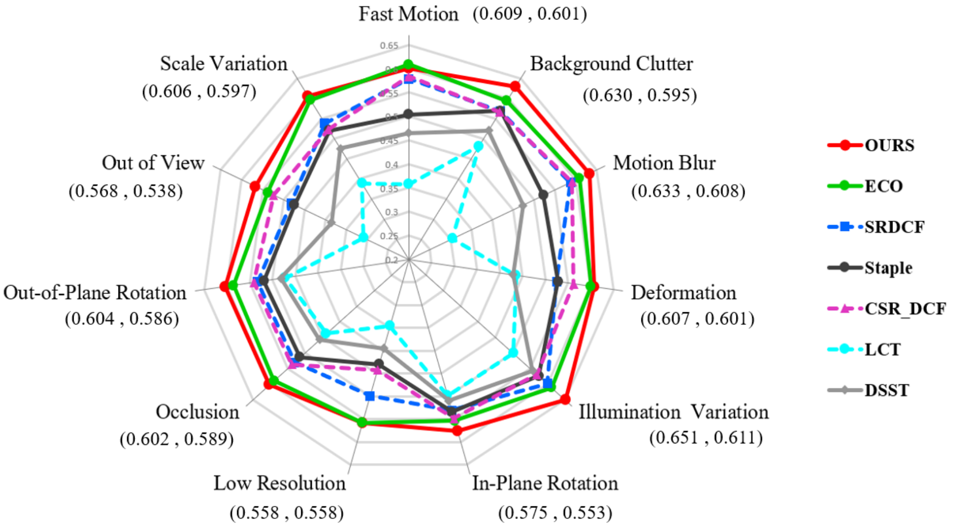

Figure 9. Comparison of area-under-curve (AUC) scores on all visual attributes on OTB2015.

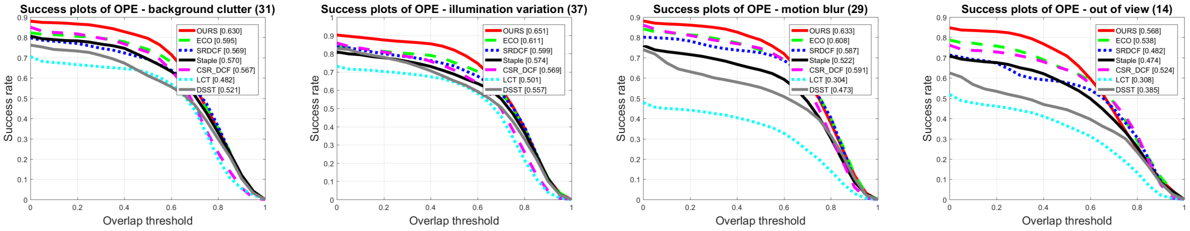

Figure 10. Success plots about Background Clutter Illumination Variation, Motion Blur and Out of View attributes on OTB2015.

In the Background Clutter case of Figure 10, the background nearby the target has similar color or texture to the target, greatly interfering with the object tracking. The proposed method can adaptively assign greater weights to features which have better distinguish capacity according to feature tracking reliability criterion, thus achieving a robust tracking. Our method achieves an AUC score of 63%, 3.5% higher than that of the second-best ECO algorithm. In the Illumination Variation case where the lighting condition changes violently, the features related to color characteristics of the target are extremely unstable and unsuitable to be used for target locating. In this case, our method can adaptively assign small weights to Color Names features that reflect the color information of the target and set greater weights to the HOG features that reflect the spatial structure and texture of the target. Thus, it can reduce damaging effect caused by illumination variation on target tracking. The AUC score of our tracker on Illumination Variation attribute reaches 65.5%, 4% higher than that of second-best ECO tracker. In Motion Blur scenario of Figure 10, the spatial structure and texture of the target tend to be unstable. However, color features can still reflect the target position. Therefore, the proposed method adaptively increases the weight of Color Names and achieves a better tracking performance. It can be seen that our method obtains an AUC score of 63.3%, 2.5% higher than that of the ECO tracker. On the Out of View attribute, the tracking process tends to fail because of temporarily disappearing of the target in the field of view. A re-detection module is introduced to detect the target position and re-initialize the tracker . The re-detection module consists of multiple SVM detectors trained by samples under different image feature maps, and the detector trained by the most reliable sample feature is activated to detect the target when the target returns to the field of view so that the tracker can continue to track the target. The AUC score of our method reaches 56.8%, 3% higher than that of the second-best method.

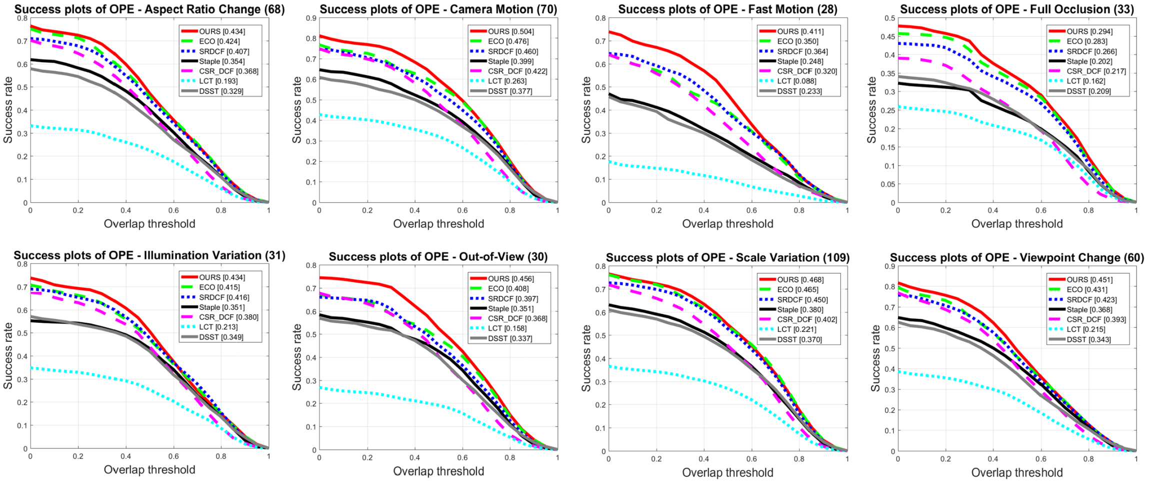

Figure 11. Success plots of trackers with 8 attributes on the UAV123 dataset.

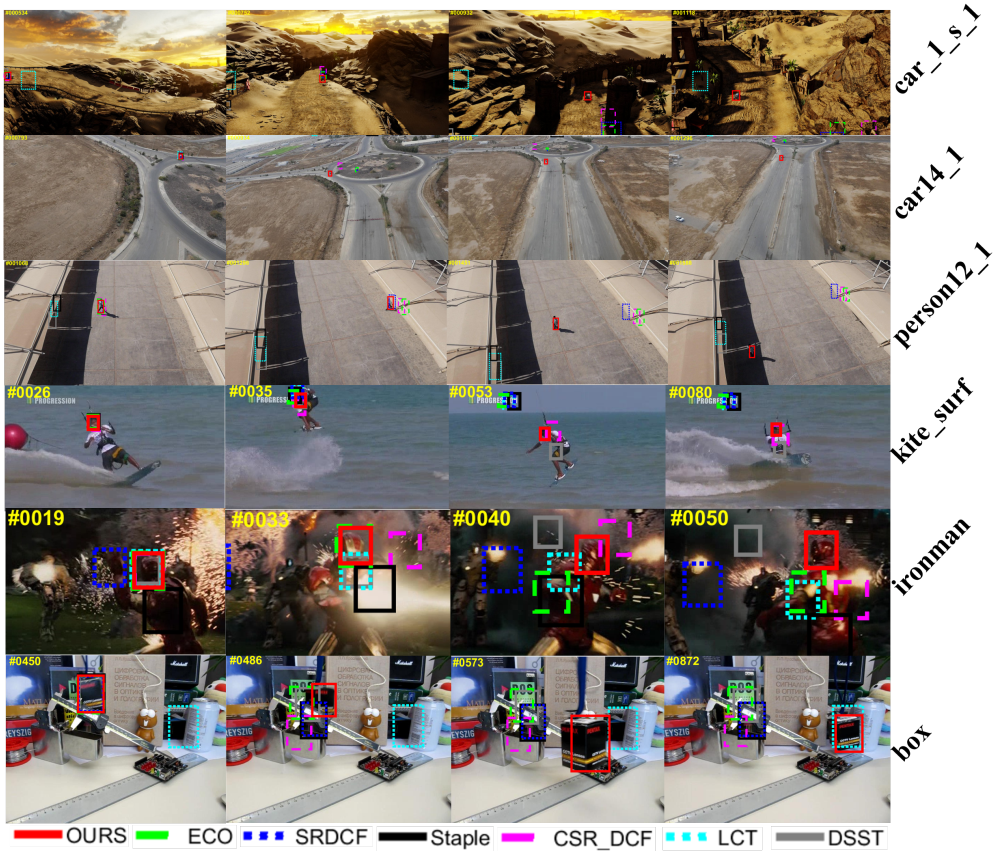

Figure 12. Qualitative comparison of the trackers on several videos.

The UAV123 dataset consists of 12 different attributes, namely Illumination Variation (IV), Scale Variation (SV), Partial Occlusion (POC), Full Occlusion (FOC), Out-of-View (OV), Fast Motion (FM), Camera Motion (CM), Background Clutter (BC), Similar Object (SOB), Aspect Ratio Change (ARC), Viewpoint Change (VC) and Low Resolution (LR). The proposed tracker in this paper performs favorably against other trackers in most attributes defined in the UAV123 dataset. Examples of success plots are demonstrated in Figure 11. It can be clearly seen that the curve corresponding to our tracking method is always higher than that corresponding to other trackers on the 8 Success plots in Figure 11, indicating that the proposed method outperforms the other 6 trackers on these 8 attributes. Especially, due to the adaptively fused robust multi-resolution feature representation and carefully choice of SVM detectors, our method greatly improved the tracking performance in videos with Camera Motion, Fast Motion and Out-of-View attributes. It can be seen that our method gets AUC sco res of 50.4%, 41.1% and 45.6% respectively on Fast Motion and Out-of-View attributes, which are 2.8%, 6.1% and 4.8% higher than the second-best tracker.

In order to compare the tracking performance of the trackers more intuitively, a qualitative comparison of the tracking results on some video sequences from OTB2015 and UAV123 dataset is shown in Figure 12. From the comparison results on "car_1_s_1" and "ironman" sequences, we can find that the proposed tracking method is robust to the illumination variation. The comparison results of "kite_surf" and "box" sequences indicate that our method can efficiently recover the object from occlusion. It can also be proven that the tracker proposed in this paper is competent in dealing with rotation, deformation and other attributes defined in OTB2015 and UAV123 benchmarks as well. Videos demonstrating the results of the propoosed method can also be found from the link: https://drive.google.com/drive/folders/15rcmiSTqQxGFnf9Fm79d4bMX6fobkWbY?usp=sharing (You can also contact us to get the videos for free: wanghan_henu@163.com).

6. Conclusions

This paper firstly proposes a novel feature fusion formulation in filter learning using the criterion of feature tracking reliability. The feature tracking reliability criterion is defined to evaluate the robustness and the background distinguishing ability of the sample features. Then, a re-detection module with multiple SVM detectors labeled with the feature tracking reliability is proposed to reduce the possibility of tracking failure and increase the accuracy of re-detection. Comparative experiments with the state-of-the-art trackers demonstrate that the proposed method yields a robust and accurate tracking in complex tracking scenarios with interference factors including illumination variation, occlusion, out-of-view and background clutter.

Author Contributions: Conceptualization, L.Z.; Methodology, H.W.; Software, J.L. (Jifang Li); Validation, Q.W.; Investigation, Z.H.; Resources, J.L. (Junwei Li); Writing—Original Draft Preparation, H.W.; Writing—Review & Editing, L.Z. and Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Foundation Council of China (61771006, 61976080, U1804149), Key research projects of university in Henan province of china (21A413002,19A413006, 20B510001), First-class Discipline Training Foundation of Henan University (2018YLTD04), the Programs for Science and Technology Development of Henan Province (192102210254).

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

References

1. He, S.; Yang, Q.; Lau, R.W.; Wang, J.; Yang, M.-H. Visual Tracking via Locality Sensitive Histograms. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2427–2434.

2. Jia, X.; Lu, H.; Yang, M.-H. Visual tracking via adaptive structural local sparse appearance model. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1822–1829.
3. Sevilla-Lara, L.; Learnedmiller, E.G. Distribution fields for tracking. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1910–1917.
4. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R. Siamese Box Adaptive Network for Visual Tracking. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6667–6676.
5. Zhou, Q.; Zhong, B.; Lan, X.; Sun, G.; Zhang, Y.; Zhang, B.; Ji, R. Fine-grained spatial alignment model for person re-identification with focal triplet loss. *IEEE Trans. Image Process.* **2020**, *29*, 7578–7589. [CrossRef]
6. Zhong, B.; Bai, B.; Li, J.; Zhang, Y.; Fu, Y. Hierarchical Tracking by Reinforcement Learning-Based Searching and Coarse-to-Fine Verifying. *IEEE Trans. Image Process.* **2019**, *28*, 2331–2341. [CrossRef] [PubMed]
7. 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]
8. Danelljan, M.; Khan, F.S.; Felsberg, M.; Van De Weijer, J. Adaptive Color Attributes for Real-Time Visual Tracking. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1090–1097.
9. Li, Y.; Zhu, J. A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration. In *Computer Vision—ECCV 2014 Workshops. ECCV 2014. Lecture Notes in Computer Science*; Springer, Cham, Switzerland, 2015; Volume 8926, pp. 254–265.
10. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Discriminative Scale Space Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1561–1575. [CrossRef] [PubMed]
11. Garcia-Cruz, X.M.; Sergiyenko, O.Y.; Tyrsa, V.; Rivas-Lopez, M.; Hernandez-Balbuena, D.; Rodriguez-Quiñonez, J.C.; Basaca-Preciado, L.C.; Mercorelli, P. Optimization of 3D laser scanning speed by use of combined variable step. *Opt. Lasers Eng.* **2014**, *54*, 141–151. [CrossRef]
12. Lindner, L.; Sergiyenko, O.; Rivas-López, M.; Ivanov, M.; Rodríguez-Quiñonez, J.C.; Hernández-Balbuena, D.; Flores-Fuentes, W.; Tyrsa, V.; Muerrieta-Rico, F.N.; Mercorelli, P. Machine vision system errors for unmanned aerial vehicle navigation. In Proceedings of the 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE), Edinburgh, UK, 19–21 June 2017; pp. 1615–1620.
13. Danelljan, M.; Hager, G.; Khan, F.S.; Felsberg, M. Learning Spatially Regularized Correlation Filters for Visual Tracking. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4310–4318.
14. Wu, Y.; Lim, J.; Yang, M.-H. Online Object Tracking: A Benchmark. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2411–2418.
15. Yang, M.-H.; Lim, J.; Yang, M.-H. Object Tracking Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1834–1848. [CrossRef]
16. Mueller, M.; Smith, N.; Ghanem, B. A Benchmark and Simulator for UAV Tracking. In *Computer Vision—ECCV 2016. Lecture Notes in Computer Science*; Springer: Amsterdam, The Netherlands, 2016; Volume 9905, pp. 445–461.
17. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
18. Boddeti, V.N.; Kanade, T.; Kumar, B.V.K.V. Correlation Filters for Object Alignment. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2291–2298.
19. Henriques, J.F.; Carreira, J.; Caseiro, R.; Batista, J. Beyond Hard Negative Mining: Efficient Detector Learning via Block-Circulant Decomposition. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Portland, OR, USA, 23–28 June 2013; pp. 2760–2767.

20. Galoogahi, H.K.; Sim, T.; Lucey, S. Multi-channel Correlation Filters. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Portland, OR, USA, 23–28 June 2013; pp. 3072–3079.
21. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.
22. Van De Weijer, J.; Schmid, C.; Verbeek, J.; Larlus, D. Learning Color Names for Real-World Applications. *IEEE Trans. Image Process.* **2009**, *18*, 1512–1523. [CrossRef] [PubMed]
23. Cai, Z.; Wen, L.; Lei, Z.; Vasconcelos, N.; Li, S.Z. Robust Deformable and Occluded Object Tracking With Dynamic Graph. *IEEE Trans. Image Process.* **2014**, *23*, 5497–5509. [CrossRef] [PubMed]
24. Sun, C.; Wang, D.; Lu, H.; Yang, M.-H. Correlation Tracking via Joint Discrimination and Reliability Learning. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 489–497.
25. Xiao, J.; Stolkin, R.; Leonardis, A. Single target tracking using adaptive clustered decision trees and dynamic multi-level appearance models. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4978–4987.
26. Bertinetto, L.; Valmadre, J.; Golodetz, S.; Miksik, O.; Torr, P.H.S. Staple: Complementary Learners for Real-Time Tracking. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1401–1409.
27. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ECO: Efficient Convolution Operators for Tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6931–6939.
28. Seo, Y.; Shin, K.-S. Image classification of fine-grained fashion image based on style using pre-trained convolutional neural network. In Proceedings of the 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA), Shanghai, China, 9–12 March 2018; pp. 387–390.
29. Ren, S., He, K., Girshick, R. and Sun, J., 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
30. Ning, G.; Zhang, Z.; Huang, C.; Ren, X.; Wang, H.; Cai, C.; He, Z. Spatially supervised recurrent convolutional neural networks for visual object tracking. In Proceedings of the 2017 IEEE International Symposium on Circuits and Systems (ISCAS), Baltimore, MD, USA, 28–31 May 2017; pp. 1–4.
31. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA , 27–30 December 2016; pp. 779–788. [CrossRef]
32. Voigtlaender, P.; Luiten, J.; Torr, P.H.; Leibe, B. Siam R-CNN: Visual Tracking by Re-Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6577–6587.
33. Zhou, Q.; Zhong, B.; Zhang, Y.; Li, J.; Fu, Y. Deep Alignment Network Based Multi-Person Tracking with Occlusion and Motion Reasoning. *IEEE Trans. Multimed.* **2019**, *21*, 1183–1194. [CrossRef]
34. Zhao, Z.; Wang, X.; Yao, P.; Bai, Y. A health performance evaluation method of multirotors under wind turbulence. *Nonlinear Dyn.* **2020**, *102*, 1701–1715. [CrossRef]
35. Jin, X.-B.; Lian, X.-F.; Su, T.-L.; Shi, Y.; Miao, B.-B. Closed-Loop Estimation for Randomly Sampled Measurements in Target Tracking System. *Math. Probl. Eng.* **2014**, *2014*, 315908. [CrossRef]
36. Bai, Y.; Wang, X.; Jin, X.-B.; Zhao, Z.; Zhang, B. A Neuron-Based Kalman Filter with Nonlinear Autoregressive Model. *Sensors* **2020**, *20*, 299. [CrossRef] [PubMed]
37. Ma, C.; Yang, X.; Zhang, C.; Yang, M.-H. Long-term correlation tracking. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5388–5396.

38. Lukežic, A.; Vojír, T.; Zajc, L. C.; Matas, J.; Kristan, M. Discriminative Correlation Filter with Channel and Spatial Reliability. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4847–4856.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affliations.