applied sciences

MDPI

Article

Learning Spatial–Temporal Background-Aware Based Tracking

Peiting Gu [1], Peizhong Liu [2,3,*], Jianhua Deng [4] and Zhi Chen [2]

[1] Fujian Provincial Key Laboratory of Data-Intensive Computing, Key Laboratory of Intelligent Computing and Information Processing, School of Mathmatics and Computer Science, Quanzhou Normal University, No. 398, Donghai Street, Quanzhou 362000, China; ptinggu@qztc.edu.cn
[2] College of Engineering, Huaqiao University, No. 269, Chenghuabei Road, Quanzhou 362021, China; 1611422001@stu.hqu.edu.cn
[3] School of Medicine, Quanzhou Medical College, Quanzhou 362011, China
[4] Chengdu Aeronautic Polytechnic, No. 699, East 7th Checheng Road, Chengdu 610100, China; dengjianhua@cuit.edu.cn
* Correspondence: pzliu@hqu.edu.cn; Tel.: +86-595-22339012

Abstract: Discriminative correlation filter (DCF) based tracking algorithms have obtained prominent speed and accuracy strengths, which have attracted extensive attention and research. However, some unavoidable deficiencies still exist. For example, the circulant shifted sampling process is likely to cause repeated periodic assumptions and cause boundary effects, which degrades the tracker's discriminative performance, and the target is not easy to locate in complex appearance changes. In this paper, a spatial–temporal regularization module based on BACF (background-aware correlation filter) framework is proposed, which is performed by introducing a temporal regularization to deal effectively with the boundary effects issue. At the same time, the accuracy of target recognition is improved. This model can be effectively optimized by employing the alternating direction multiplier (ADMM) method, and each sub-problem has a corresponding closed solution. In addition, in terms of feature representation, we combine traditional hand-crafted features with deep convolution features linearly enhance the discriminative performance of the filter. Considerable experiments on multiple well-known benchmarks show the proposed algorithm is performs favorably against many state-of-the-art trackers and achieves an AUC score of 64.4% on OTB-100.

Keywords: boundary effects; spatial–temporal regularization; discriminative correlation filter

Citation: Gu, P.; Liu, P.; Deng, J.; Chen, Z. Learning Spatial-Temporal Background-Aware Based Tracking. *Appl. Sci.* **2021**, *11*, 8427. https://doi.org/10.3390/app11188427

Academic Editor: Antonio Fernández-Caballero

Received: 7 June 2021
Accepted: 3 September 2021
Published: 10 September 2021

1. Introduction

Visual object tracking is one of the popular problems in the computer vision community that has gained a wide spectrum of attention, and that is widely applied to practical applications such as intelligent video monitoring application, intelligent driving car, virtual reality, high-speed translation location and UAV tracking in 6-Generation technology. Although numerous efforts have been made in the previous ten years, there still exist various intractable challenges such as occlusion, fast motion and deformation. Recently, the mainstream object tracking methods are generally categorized into two types, one is a tracking method based on discriminative correlation filter framework, and the other is a tracking method based on deep learning networks. This paper mainly studies the target tracking method based on DCF.

The DCF based paradigm has obtained a top-ranked performance and has high speed attributed to its efficiency in the Fourier domain. Numerous methods based on the DCF baseline framework have been successively proposed and improved through a variety of efficient methods such as multi-scale strategies [1,2], multi-feature fusion methods [3–5], spatial–temporal context based methods [6,7], patch-based strategy [8,9], multi-kernel techniques [10], particle filter methods [11,12] and sparse representation methods [13].

In addition, the Siamese network methods [13–17], end-to-end learning methods [18,19] and the continuous convolution operation methods [20,21] have also been further developed to improve the robustness of the tracker. Although the DCF based methods have been extensively refined, there still exist some imperfections. First, the circulant shifted sampling process on the DCF based framework always suffers unwanted boundary effects, which can easily lead to model drift and deteriorate tracking performance. Second, the negative training samples surrounding the target object have not been exploited adequately, and it is difficult to extract target information from similar backgrounds, and thus complicated appearance changes further degrade the robustness of the learned filters. Third, most traditional DCF based trackers usually employ simple shallow-lever handcrafted features to represent the target. These features always find it tough to capture higher-lever semantic information, and thus cannot locate the target accurately. These imperfections will suppress the discriminative performance of the DCF based tracker from different angles.

In recent years, there are a variety of trackers that have been developed to solve the above-mentioned boundary effect problem from different angles. Among them, the SRDCF tracker [22] restrains the filter coefficients through the spatial regularization term, which makes the filter learn in a larger area of space and thus effectively alleviates unwanted boundary effects. However, the main disadvantage of this method is that the regularization term introduces time-consuming computational resources. The BACF tracker [23] fully exploits the negative training sample surrounding background area, and effectively enhances the discriminative performance of the learned filter to complicated appearance variance. The STRCF [24] tracker introduces spatial–temporal regularization based on the SRDCF tracker, which effectively gains the robustness toward appearance changes, and thus deals well with boundary effects. The ARCF [25] tracker suppresses the occurrence of aberrance by suppressing the rate of change of the response map, which can effectively handle the boundary effect problem.

In this paper, in order to solve the above-mentioned problems efficiently, we develop a robust and efficient DCF based tracker, which is mainly dedicated to handing the boundary effect problem. The main contributions of this paper are summarized as follows:

1. A spatial–temporal regularization background-aware DCF based model is proposed, which can deal robustly with boundary effects and complex appearance changes.
2. Our model can effectively be solved by the alternating direction multiplier method (ADMM), and each sub-problem has a corresponding closed solution.
3. The proposed tracker gains promising tracking performance and significantly outperforms than other state-of-the-art DCF based tracker in accurate and overlap success rate.

The remaining of the work is structured as follows: Section 2 introduces related works, Section 3 presents the proposed tracking framework, and Section 4 introduces experimental verification and analysis. Finally, the conclusions are introduced in Section 5.

2. Related Works

DCF based paradigms have been extensively employed in the field of visual tracking for a long time. Bolme et al. [26] first introduced the correlation filter theory to visual tracking and proposed a single-channel minimum error square sum (MOSSE) correlation filter. The method is chosen as a baseline tracker to expand and optimize from different viewpoints. Henriques et al. [27] proposed DCF with circulant structure kernels (CSK) and employed dense sampling to achieve high-efficiency tracking performance based on tracking-by-detection framework. Danelljan [28] introduced a kernel trick into the DCF framework, and effectively extended the single-channel to multi-channel features to achieve real-time and robust tracking performance. In order to solve the difference of variance on the scale, Li et al. [29] proposed a IBCCF model, which can deal well with the

aspect ratio variance in the tracking process, and effectively improve the tracking robustness and accuracy. In terms of feature fusion, [3] built an effective analysis framework of multiple clues based on DCF for visual tracking, and integrated deep convolution features and traditional hand-crafted features to increase the discriminative performance of the classifier. Bhat et al. [30] systematically analyzed the characteristics of deep and shallow-level hand-crafted features and proposed a novel adaptive feature fusion improvement that measures robustness and preciseness of the tracker, which is performed by employing the complementary relationship between the deep and shallow features. To handle with the boundary effect problem, Li et al. [24] introduced spatial–temporal regularization based on SRDCF, which enable robustness against the boundary effect problem. Dai et al. [31] proposed the ASRCF model to address the boundary effect problem, which is achieved by introducing adaptive spatial regularization module, and can learn the reliable filter coefficients make the filter gain robustness to complex appearance changes.

Currently, deep convolution features are widely exploited in the field of visual tasks [32–35], the combination of the DCF based paradigm and deep network model is a significant drift in the development of visual tracking. With the help of deep network features, deep semantic information is captured, improve the tracking accuracy of tracking instrument. Reference [19] proposed the CREST model, which considers DCF as a layer class convolutional neural network, and integrates feature extraction, the corresponding reflection image is formed, and the formed image is updated to the corresponding neural network to form the corresponding training port to achieve promising tracking results. Zhang et al. [36] incorporated geometric transformation a network architecture based on correlation filtering information, and introduced a spatial alignment module, which can deal effectively with a variety of complex appearance changes and geometric transformations. Zhu et al. [37] make the most of the rich optical flow information between consecutive frames to enhance the representation of features, and treated optical flow estimation, feature extraction, and correlation filter and tracking processes as special layers of deep networks to achieve high-efficient tracking performance. In this work, we combine traditional hand-crafted features with deep network features (VGG-Net) on a DCF based framework, which takes full advantage of the performance dominant position of multiple properties to further raise the accuracy and robustness of the tracker.

3. Spatial–Temporal Regularization Background-Aware Correlation Filter

3.1. Background-Aware Correlation Filters Framework

In this work, the proposed algorithm mainly chose the background-aware filter as the baseline tracker, and then briefly reviews the basic principles of BACF.

Denote by $x \in R^D$ a vectorized image with D channels, y is the ideal correlation response map $y \in R^D$, the main objective function of BACF is to reduce the value of the objective function to the minimum, that is, minimize it:

$$E(w) = \frac{1}{2}\left\| y - \sum_{d=1}^{D} Bx_d * w_t \right\|_2^2 + \frac{\mu}{2}\sum_{d=1}^{D}\|w_d\|_2^2 \quad (1)$$

The letter B represents a clipping matrix in the formula, and it is binary. Input D elements through matrix clipping, and w_d is the representative of the D channel in the channel learning filter.

BACF mainly exploits negative training in real sense samples densely pick up from the background area to learn or update the filter, which can deal effectively with appearance changes, but there still existing some imperfections in addressing boundary effect. First, due to the BACF tracker mainly learns and trains the classifier through current frame, it does not take into account the spatial–temporal information of the historical frame, which affects the discriminative ability of the filter so that it cannot distinguish the target from similar backgrounds. Secondly, in terms of feature representation, BACF mainly employs 31-channel HOG features for representing the target. Limited traditional

hand-crafted features are difficult to capture abstract semantic information, which immediate affects the accuracy of tracking equipment. From the above analysis, it is obvious that the BACF tracker has some deficiency in dealing with the appearance change problem, and then degrades its performance in dealing with boundary effects. In this work, to solve the impact of the above problems, a spatial–temporal regularization background-aware DCF based algorithm is proposed; the framework of the proposed method is shown in Figure 1.

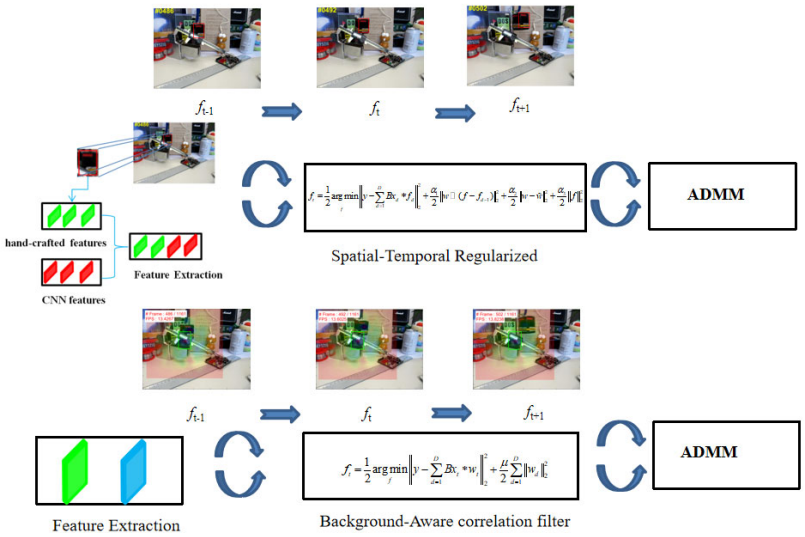

Figure 1. The systematic work-flow of the proposed tracking algorithm.

3.2. The Objective Function of the Proposed Model

Inspired by the above-mentioned analysis, we propose a spatial–temporal regularization background-aware discriminative correlation filter, which is achieved by introducing a spatial–temporal regularization module [24] based on the background-aware filter framework. As shown in Figure 1, the spatial and temporal relation is effectively established between historical frames, and thus gains robustness to appearance changes. Our main objective equation can be expressed as follows:

$$E(f) = \frac{1}{2}\left\| y - \sum_{d=1}^{D} Bx_d * f_d \right\|_2^2 + \frac{\alpha_1}{2}\left\| f - f_{d-1} \right\|_2^2 + \frac{\alpha_2}{2}\left\| f \right\|_2^2 \tag{2}$$

where, f_{d-1} represents a filter that is learned in a channel in a filter, such as a $D-1$ channel. α_1 and α_2 represent regularization parameters. $\left\| f - f_{d-1} \right\|_2^2$ is the introduced spatial–temporal regularization term, make the filter establish a spatial–temporal relationship between the current and historical frame.

3.3. Optimization of the Proposed Model

Correlation filters usually benefit from the frequency domain for high-efficiency computation, and can transform complicated convolution operations into simple element-wise multiplications. Therefore, we convert Equation (2) into the following corresponding frequency domain form:

$$\begin{aligned} E(f, \hat{g}) &= \frac{1}{2}\left\| \hat{y} - \hat{X}\hat{g} \right\|_2^2 + \frac{\alpha_1}{2}\left\| f - f_{d-1} \right\|_2^2 + \frac{\alpha_2}{2}\left\| f \right\|_2^2 \\ \text{s.t.} \quad \hat{g} &= \sqrt{T}(I_D \otimes FB^T) f \end{aligned} \tag{3}$$

The symbol ^ in the equation represents the transformation of the signal in the discrete Fourier transform, for instance $\hat{\alpha} = \sqrt{T} F \alpha$, F is the orthonormal $T \times T$ matrix of com-

plex basis vectors for mapping any T dimensional vectorized signal into the Fourier domain. $\hat{g} = (\hat{g}_1, \hat{g}_2, ..., \hat{g}_D)$ denotes auxiliary variables parameter. $\otimes$ denotes Kronecker product, I_D is $D \times D$ identity matrix, T represents the conjugate transpose operation of a complex vector or matrix. It can be clearly seen that Equation (3) is a convex function, which can be optimized by the ADMM method to obtain a local optimal solution. The specific incremental Lagrangian form can be reformulated as the following equation:

$$\begin{aligned} E(f, \hat{g}, \hat{\mu}) = & \frac{1}{2}\left\|\hat{y} - \hat{X}\hat{g}\right\|_2^2 + \frac{\alpha_1}{2}\left\|f - f_{t-1}\right\|_2^2 + \frac{\alpha_2}{2}\|f\|_2^2 \\ & + \hat{\rho}^T(\hat{g} - \sqrt{T}(I_D \otimes FB^T)f) + \frac{\mu}{2}\left\|\hat{g} - \sqrt{T}(I_D \otimes FB^T)f\right\|_2^2 \end{aligned} \tag{4}$$

where is the introduced penalty parameter of the error term, $\hat{\rho} = [\hat{\rho}_1, \hat{\rho}_2, ..., \hat{\rho}_D]^T$ is the introduced Lagrange auxiliary variable of $DT \times 1$. Employing the ADMM optimization method to solve Equation (4) iteratively, each subproblem $\hat{g}^*$ and f^* has corresponding closed solutions:

Subproblem f^*:

$$\begin{aligned} f^* &= \arg\min_f \left\{ \frac{\alpha_1}{2}\|f - f_{t-1}\|_2^2 + \frac{\alpha_2}{2}\|f\|_2^2 + \hat{\rho}^T(\hat{g} - \sqrt{T}(I_D \otimes FB^T)f) + \frac{\mu}{2}\left\|\hat{g} - \sqrt{T}(I_D \otimes FB^T)f\right\|_2^2 \right\} \\ &= [\alpha_1 + \alpha_2 + \mu T]^{-1}[\alpha_1 f_{t-1} + \hat{\rho}^T\sqrt{T}(I_D \otimes FB^T) + \mu\hat{g}\sqrt{T}(I_D \otimes FB^T)] \\ &= (\frac{\alpha_1 + \alpha_2}{T} + \mu)^{-1}(\frac{\alpha_1 f_{t-1}}{T} + \rho + \mu g) \end{aligned} \tag{5}$$

where g and ρ can be obtained by the following inverse Fourier transform operations:

$$\begin{cases} g = \dfrac{1}{\sqrt{T}}(I_D \otimes BF^T)\hat{g} \\ \rho = \dfrac{1}{\sqrt{T}}(I_D \otimes BF^T)\hat{\rho} \end{cases} \tag{6}$$

Subproblem g^*:

$$g^* = \arg\min_{\hat{g}} \left\{ \frac{1}{2}\left\|\hat{y} - \hat{X}\hat{g}\right\|_2^2 + \hat{\rho}^T(\hat{g} - \sqrt{T}(I_D \otimes FB^T)f + \frac{\mu}{2}\left\|\hat{g} - \sqrt{T}(I_D \otimes FB^T)f\right\|_2^2 \right\} \tag{7}$$

The computational complexity of Equation (7) is $O(T^3D^3)$, we solve the equation at every ADMM iteration will generate heavy computation, which will deteriorate real-time performance. Since the sparse property of X is utilized, each element of $\hat{y}$ $(\hat{y}(1), \hat{y}(2), ..., \hat{y}(m))$ is merely dependent on each $\hat{x}(m) = [\hat{x}_1(m), \hat{x}_2(m), ..., \hat{x}_D(m)]^T$ and $\hat{g}(m) = [conj(\hat{g}_1(m)), conj(\hat{g}_2(m)), ..., conj(\hat{g}_D(m))]^T$, $conj(\cdot)$ represents the complex conjugate operation of a complex vector.

Subproblem $\hat{g}^*$ can be divided into the following M smaller subproblems as follows $\hat{g}(m)^*$ $[m = 1, ..., M]$

$$\hat{g}(m)^* = \arg\min_{\hat{g}(m)} \left\{ \frac{1}{2}\left\|\hat{y}(m) - \hat{x}(m)^T\hat{g}(m)\right\|_2^2 + \hat{\rho}(m)^T(\hat{g}(m) - \hat{f}(m)) + \frac{\mu}{2}\left\|\hat{g}(m) - \hat{f}(m)\right\|_2^2 \right\} \tag{8}$$

where $\hat{f}(m) = [\hat{f}_1(m), \hat{f}_2(m), ..., \hat{f}_D(m)]$ and $\hat{f}_d = \sqrt{D}FB^T f_d$. The solution of each subproblem of $\hat{g}^*$ can be obtained by the following equation:

$$\hat{g}(m)^* = (\hat{x}(m)\hat{x}(m)^T + M\mu I_D)^{-1} \cdot (\hat{y}(m)\hat{x}(m) - M\hat{\rho}(m) + M\mu\hat{f}(m)) \tag{9}$$

The computational complexity of Equation (9) is $O(TD^3)$, with the inverse operation, there will cause huge computational resources, which will affect tracking real-time performance. The Sherman–Morrison formula [38] is applied for further optimize and accelerate the calculation.

$$(uv^T + A)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1}u} \tag{10}$$

where in this situation, $u = v = \hat{x}_t(m)$, $A = \frac{\mu}{1+\gamma} I_D$, Equation (9) can be reformulated as follows:

$$\begin{aligned} \hat{g}(m)^* = & \frac{1}{\mu}(M\hat{y}(m)\hat{x}(m) - \hat{\rho}(m) + \mu\hat{f}(m)) \\ & - \frac{\hat{x}(m)}{\mu l}(M\hat{y}(m)\hat{r}_x(m) - \hat{r}\rho(m) + \mu\hat{r}f(t)) \end{aligned} \tag{11}$$

where, $\hat{r}_x(m) = \hat{x}(m)^T\hat{x}$, $\hat{r}_\rho(m) = \hat{x}(m)^T\hat{\rho}$, $\hat{r}_f(m) = \hat{x}(m)^T\hat{f}$, and $l = \hat{r}_x(m) + T\mu$. After optimization, the computational complexity is smaller than $O(TK)$, and all sub-problems have been solved.

3.4. Lagrangian Parameter Update

In this work, we employ an online adaptive template update strategy to update Lagrangian parameter, the model of the proposed tracker is updated as follows:

$$\hat{x}^m_{model} = (1 - \lambda)\hat{x}^{model}_{m-1} + \lambda\hat{x}_m \tag{12}$$

where m and $m - 1$ represent the $(m)th$ and $(m - 1)th$ frames of the video sequences, respectively, and λ represents an efficiency of learning, namely learning rate parameter of the mode.

4. Experiments

4.1. Implementation Details

The experiment of the proposed method is conducted using MATLAB2017a on a PC with an i7-8700 3.2GHz GPU with 16GB RAM, and NVIDIA GeForce GTX 1070Ti GPU with 11GB RAM. In order to obtain accurate feature information, we employ 31-channel HOG features, color names, and hierarchical convolutional features such as (conv5-4 and conv4-4 layers in VGG-19). The regularization weight factors in Equation (2) are set to 0.01 and 0.08, respectively. In terms of scale estimation, and the number of scales is set to 7 and the scale step is set to 1.02. In terms of the ADMM algorithm, the number of ADMM iterations is set to 5, the penalty weight factor of in Equation (4) is set to 1, and the learning rate λ is set to 0.0192, and then updated by $\mu(i+1) = \min(\mu_{\max}, \beta\mu^{(i)})$, where β=10 and $\mu_{\max} = 10^3$.

4.2. The Overall Tracking Results on OTB Dataset

We compare the proposed method with 10 object tracking methods, including ECO-HC [20], GradNet [39], MCPF [11], DeepSRDCF [40], DeepLMCF [41], CREST [19], UCT [42], ARCF [25], SRDCF [28] and STAPLE_CA [6] on a well-known tracking benchmark datasets [43,44], which contains almost 50 and 100 video annotations video sequences separately with 11 different attributes. We evaluate the performance of 10 trackers by using two metrics provided in [43] on OTB-50 [44] and OTB-100 dataset, the results of tracking are displayed by the report of overlapping success rate and range accuracy. The distance precision (DP) shows that the ratio of frames whose center location error is within a certain threshold. The overlap success plot, which shows the number of thresholds given is less than the overlap number of bounding boxes, and generally, the given threshold of OS is set to 0.5. We report by using some data provided by [43] through the evaluation protocol, the data is applied to the tracking performance of the protocol, and the overlapping success graph and precision graph on these data sets are used. The results are shown in Figures 2 and 3 respectively.

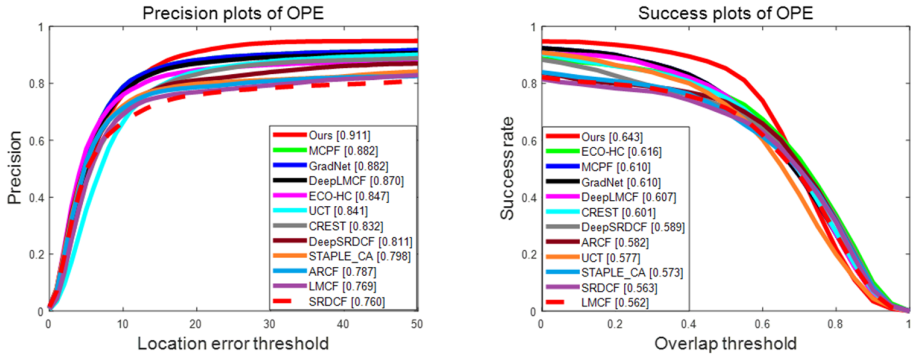

Figure 2. The overall tracking performances of precision and overlap success plots on OTB-50 dataset using OPE evaluation.

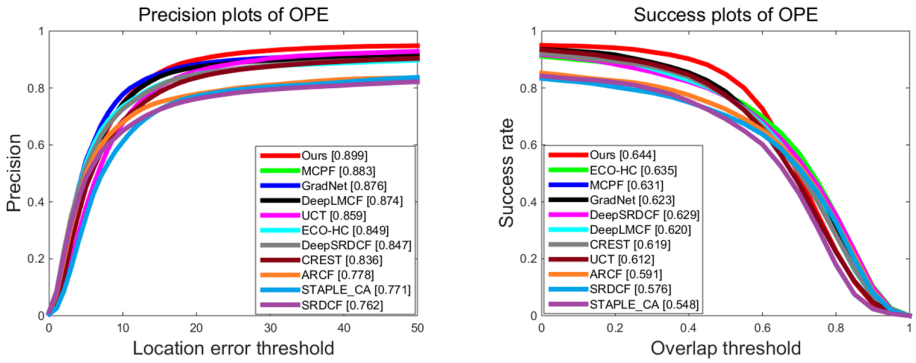

Figure 3. The overall tracking performances of precision and overlap success plots on OTB-100 dataset using OPE evaluation.

We compare 10 kinds of newly discovered OS methods with DP methods, and the proposed tracker performs well with DP of 91.1% and OS of 64.3%. It can be easily seen from Figure 1 that our tracker performs favorably against the other existing methods; in particular, it significantly outperforms the MCPC, GradNet and DeepLMCF trackers that utilize visual deep features.

It can be easily seen from Figure 3 the proposed tracker performs well with DP of 89.9% and OS of 64.4%, where the average DP of 89.9 outperformed the recent state-of-the-art trackers such as MCPF (88.3), GradNet (87.6), DeepLMCF (87.4), UCT (85.9), ECO-HC (84.9), DeepSRDCF (84.7), CREST (80.1), ARCF (77.8), STAPLE_CA (77.1), SRDCF (76.2), and the average OS of 66.4 outperformed ECO-HC (63.5), MCPF (63.1), GradNet (62.3), DeepSRDCF (62.9), DeepLMCF (62), CREST (61.9), UCT (61.2), ARCF (59.1), SRDCF (57.6), STAPLE_CA (54.8). These tracking results demonstrate the validity of introducing the spatial–temporal module to DCF based framework.

Table 1 shows the mean FPS results of the proposed tracker and some mainstream trackers. It can be seen from the table that the proposed tracker combines traditional hand-crafted and deep network features, which significantly increases the amount of calculation and makes the tracker's speed worse than STRCF and BACF, which only use hand-crafted trackers but still outperform some trackers such as SRDCF and MCPF.

Table 1. The mean FPS results of the proposed trackers with several mainstream tracking methods on OTB-2015.

	ARCF	STRCF	BACF	SRDCF	STAPLE_CA	MCPF	Ours
Mean FPS	15.3	24.2	26.7	3.8	35.3	1.8	5

Figure 4 shows the attribute-based evaluation results of six video attributes on OTB50. The results demonstrate proposed tracker outperforms existing competing trackers, this empirically shows how combining spatial–temporal regularization with background-aware module improves the reliability of such trackers against complex boundary effect and appearance variations. Trackers such as MCPF, ECO-HC and DeepLMCF have shown to be less robust to background clutter, out of view, deformation, illumination and variation, respectively.

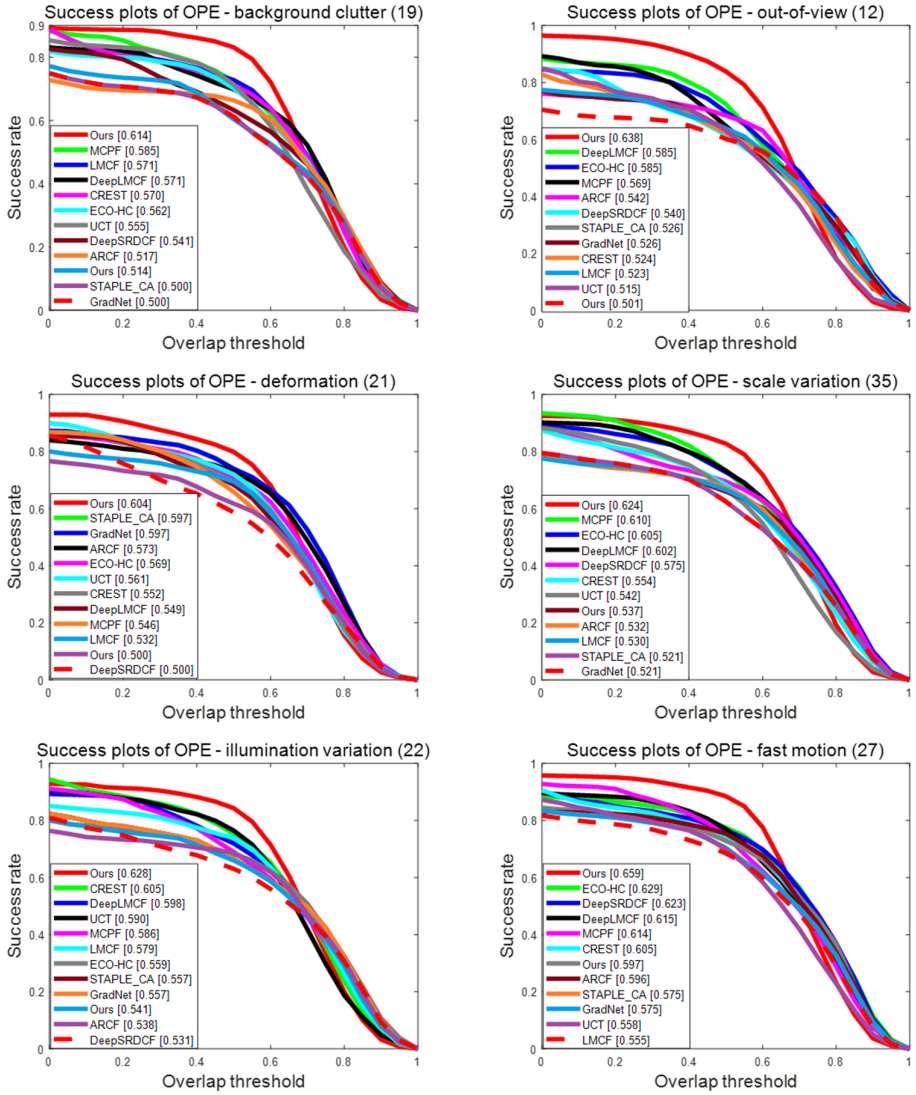

Figure 4. The success plots of the proposed trackers with 6 video attributes on the OTB50 dataset.

4.3. The Overall Tracking Results on Temple-Color 128 Dataset

We evaluated the dataset based on the tracking results generated by the temple color-128 dataset [45], which contains 128 color sequences, and then compares it with the most advanced and highest level method, including ECO [20], ECO-HC [20], STRCF [24], TADT [46], MCPF [11], DeepSRDCF [40], STAPLE [32], BACF [23], PTAV [45], SRDCF [28] and the success rate of overlapping and the accuracy of distance accuracy are taken as the evaluation indexes. Table 2 shows that the proposed tracking instrument achieves good performance in the most advanced trackers with a OS of 56.6% and DP of 78% which is closely followed by ECO (60%), where the average OS of 56.6% outperformed the baseline

trackers such as BACF (51.9%) and homogeneous tracker such as STRCF (55.3%), it demonstrates the effectiveness of the combination of background-aware and spatial–temporal regularization term, which can deal effectively with boundary effect issue. This different discovery further illustrates the combination of deep convolution features, which can further improve the recognition ability of the classifier. Due to different feature combinations, there is a certain degree of complementarity, where the traditional manual feature is generally used to capture some details of the superficial appearance, while the deep feature represents the higher-level information, the semantic information of the target space. It also demonstrates that the introduced spatial–temporal regularization module significantly gains robustness to appearance variance caused by the unwanted boundary effect and improves the discriminative ability of the learned filter to some extent.

Table 2. The overall tracking results on the TC-128 datasets.

Tracker Name	DP	OS
ECO	0.741	0.605
ECO-HC	0.726	0.551
MCPF	0.774	0.545
BACF	0.648	0.519
TADT	0.756	0.562
PTAV	0.742	0.546
DeepSRDCF	0.738	0.536
STAPLE	0.668	0.497
SRDCF	0.675	0.485
STRCF	0.723	0.553
Ours	0.780	0.566

4.4. The Overall Tracking Results on UAV123 Dataset

We evaluate the tracking results on the UAV123 dataset [47], which contains 123 challenging sequences with comparisons to state-of-the-art methods, including ECO-HC [20], DSST [1], SRDCF [28], BACF [23], STAPLE_CA [6], STAPLE [32], KCF [28], SAMF [2] and ARCF [25], Table 3 show that the proposed tracker performs well with OS of 57.2% which is closely followed by ARCF (60%), where the average OS of 57.2% outperformed the baseline trackers such as BACF (51.9%), it further demonstrate that the effectiveness of the combination of spatial–temporal regularization and background-aware module.

Table 3. The overall tracking results on the UAV123 datasets.

Tracker Name	OS
ECO-HC	0.507
DSST	0.448
SRDCF	0.465
BACF	0.519
STAPLE_CA	0.562
STAPLE	0.546
KCF	0.406
SAMF	0.485
ARCF	0.600
Ours	0.572

4.5. The Qualitative Evaluation

Figure 5 shows a qualitative comparison between the proposed tracker and the 10 most advanced trackers (MCPF, GradNet, SRDCF, ECO-HC, UCT, ARCF, DeepSRDCF,

STAPLE_CA, DeepLMCF and CREST) on five representative challenging sequences (MotorRolling, DragonBaby, Skiing, Bolt2 and Matrix) from OTB-2015. From the figure, we can see that the proposed tracker deals well with motion blur, fast motion, deformation, scale variation, out-of-view, out-plane rotation, occlusion and background clutter scenarios challenges.

In the Matrix and MotorRolling sequences, the object mainly experiences significant appearance changes such as scale variation and background clutter challenges. Most of the trackers such as GradNet, ARCF, CREST, UCT, ECO-HC, DeepSRDCF, DeepLMCF and STAPLE_CA lose the target and fail to recover from tracking drift, therefore, compared with the traditional hand-crafted feature or CNN feature, it is more effective to fuse the hand-crafted feature and multiple powerful hierarchical convolution features. However, due to the proposed method, the proposed tracker can deal effectively with challenges of motion blur and occlusion.

As for the skiing, DragonBaby and Bolt2 sequences, the object undergoes a certain degree of appearance change, such as fast motion, deformation and scale variance. The proposed tracker locates the target accurately from #28 to #34 in Skiing sequence, and performs more robust than ECO-HC, ARCF, SRDCF and DeepSRDCF trackers. GradNet, MCPF, ECO-HC, UCT, ARCF, CREST, DeepSRDCF, DeepLMCF and STAPLE_CA trackers fail to track the target object successfully from #46 to #81 in the DragonBaby sequence when fast motion and large-scale scale change occurs. GradNet, ARCF, ECO-HC, UCT, CREST and SRDCF in the sequence with other features, such as bolt2 sequence deformation and background clutter, the performance of the tracker is not so good. Even so, through the experiment, it is found that the tracker can still locate the related targets well, which further verifies the reliability of the proposed method.

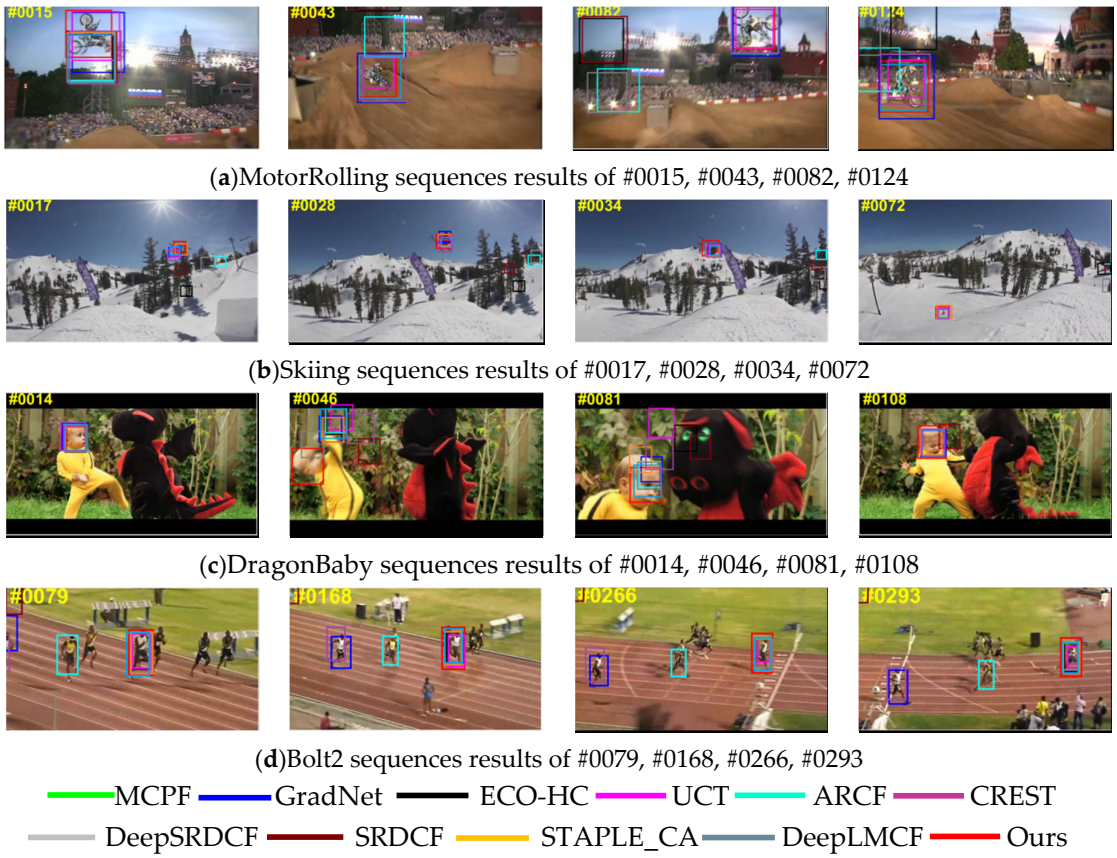

Figure 5. Qualitative evaluation of the proposed tracker with comparison to 10 trackers on 5 challenge sequences (from left to right and top to down are Matrix, MotorRolling, Skiing, DragonBaby and Bolt2).

5. Conclusions

In this work, we integrate a spatial–temporal regularization module into a background-aware correlation filter framework, which is performed by adaptive balancing between active and passive model learning, thus gaining robustness to target appearance variance. The proposed model can be effectively solved by the ADMM optimization algorithm, which accelerates the convergence of the algorithm. In addition, from the perspective of feature representation, the proposed model effectively combines higher-level deep features with shallow hand-crafted features, which makes the filter capture more abstract semantic information, and then improves the discriminative ability of the learned filter. Compared with many new and advanced technologies, the performance of the proposed tracker is still better.

Author Contributions: Conceptualization, P.G. and P.L.; methodology, P.G.; software, J.D.; validation, P.G., P.L. and Z.C.; formal analysis, J.D.; investigation, P.G.; resources, J.D.; data curation, P.G. and Z.C.; writing—original draft preparation, P.G. and P.L.; writing—review and editing, P.G. and P.L.; visualization, P.G. and Z.C.; supervision, J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by Fujian Provincial Big Data Research Institute of Intelligent Manufacturing, and by Fujian Provincial Science and Technology Major Project (No.2020HZ02014), and by the Education and Scientific Research Project for Young and Middle-aged Teachers of Fujian Province 2018 (No. JT180357).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: http://cvlab.hanyang.ac.kr/tracker_benchmark/datasets.html (accessed on 7 June 2021).

Acknowledgments: The authors would like to thank all of the reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Danelljan, M.; Häger, G.; Khan, F.; Felsberg, M. *Accurate Scale Estimation for Robust Visual Tracking*; BMVA Press: London, UK, 2014.
2. Li, Y.; Zhu, J. *A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration*; European Conference on Computer Vision; Springer: Cham, Switzerland, 2014; pp. 254–265.
3. Wang, N.; Zhou, W.; Tian, Q.; Hong, R.; Wang, M.; Li, H. Multi-cue 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.
4. Ma, C.; Huang, J.; Yang, X.; Yang, M. Hierarchical Convolutional Features for Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3074–3082.
5. Qi, Y.; Zhang, S.; Qin, L.; Yao, H.; Huang, Q.; Lim, J.; Yang, M.H. Hedged Deep Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4303–4311.
6. 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.
7. Zhang, K.; Zhang, L.; Yang, M.H.; Zhang, D. Fast Tracking via Spatio-Temporal Context Learning. European Conference on Computer Vision (ECCV), pp. 127–141, Zurich, Switzerland, September, 2014.
8. Liu, S.; Zhang, T.Z.; Cao, X.C.; Xu, C.S. 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.
9. Liu, T.; Wang, G.; Yang, Q.X. 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.
10. Tang, M.; Yu, B.; Zhang, F.; Wang, J.Q. High-Speed Tracking with Multi-kernel Correlation Filters. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4874–4883.

11. Zhang, T.Z.; Xu, C.S.; Yang, M.H. Multi-task Correlation Particle Filter for Robust Object Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4335–4343.
12. Zhang, T.Z.; Liu, S.; Xu, C.S.; Liu, B.; Yang, M.H. Correlation Particle Filter for Visual Tracking. In *IEEE Transactions on Image Processing*; IEEE: Piscataway, NJ, USA, 2018; Volume 27; pp. 2676–2687.
13. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2019; pp. 4282–4291.
14. Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; Torr, P.H.S. Fast Online Object Tracking and Segmentation: A Unifying Approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1328–1338.
15. Zhang, Z.P.; Peng, H.W. Deeper and Wider Siamese Networks for Real-Time Visual Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4591–4600.
16. Zhang, W.; Du, Y.; Chen, Z.; Deng, J.; Liu, P. Robust adapitive learning with Siamese network architecture for visual. *Vis. Comput.* **2021**, *37*, 881–894.
17. Voigtlaender, P.; Luiten, J.; Torr, P.H.S.; Leibe, B. Siam R-CNN:Visual Tracking by Re-Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp.6578–6588.
18. 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 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5000–5008.
19. Song, Y.B.; Ma, C.; Gong, L.J.; Zhang, J.W.; Lau, R.W.H.; Yang, M.Y. CREST: Convolutional Residual Learning for Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2555–2564.
20. Danelljan, M.; Bhat, G.; Khan, F.S.; 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. 6638–6646.
21. Danelljan, M.; Robinson, A.; Khan, F.S.; Felsberg, M. Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016.
22. Danelljan, M.; Häger, 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.
23. Galoogahi, H.K.; Fagg, A.; Lucey, S. Learning Background-Aware Correlation Filters for Visual Tracking. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1144–1152.
24. Li, F.; Tian, C.; Zuo, W.M.; Zhang, L.; Yang, M.H. Learning Spatial-Temporal Regularized Correlation Filters for 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. 4904–4913.
25. Huang, Z.H.; Fu, C.H.; Li, Y.M.; Lin, F.L.; Lu, P. Learning Aberrance Repressed Correlation Filters for Real-Time UAV Tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 2891–2900.
26. 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.
27. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. In Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012.
28. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans. pattern Anal. Mach. Intell.* **2015**, *37*, 583–596.
29. Li, F.; Yao, Y.J.; Li, P.H.; Zhang, D.; Zuo, W.M.; Yang, M.H. Integrating Boundary and Center Correlation Filters for Visual Tracking with Aspect Ratio Variation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2001–2009.
30. Bhat, G.; Johnander, J.; Danelljan, M.; Khan, F.S.; Felsberg, M. Unveiling the Power of Deep Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 483–498.
31. Dai, K.; Wang, D.; Lu, H.C.; Sun, C.; Li, J.H. Visual Tracking via Adaptive Spatially-Regularized Correlation Filters. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4670–4679.
32. 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.
33. You, H.F.; Tian, S.W.; Yu, L.; Lv, Y.L. Pixel-Level Remote Sensing Image Recognition Based on Bidirectional Word Vectors. *IEEE Trans. Geosci. Remote. Sens.* **2020**, *58*, 1281–1293.
34. Cai, W.W.; Wei, Z.G. PiiGAN: Generative Adversarial Networks for Pluralistic Image Inpainting. *IEEE Access* **2020**, *8*, 48451–48463.

35. Convertini, N.; Dentamaro, V.; Impedovo, D.; Pirlo, G.; Sarcinella, L. A Controlled Benchmark of Video Violence Detection Techniques. *Information* **2020**, *11*, 321, doi:10.3390/info11060321.
36. Zhang, M.; Wang, Q.; Xing, J.; Gao, J.; Peng, P.; Hu, W.; Maybank, S. Visual Tracking via Spatially Aligned Correlation Filters Network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 469–485.
37. Zhu, Z.; Wu, W.; Zou, W.; Yan, J. End-to-End Flow Correlation Tracking with Spatial-Temporal Attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 548–557.
38. Sherman, J.; Morrison, W.J. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Stat.* **1950**, *21*, 124–127.
39. Li, P.; Chen, B.Y.; Ouyang, W.L.; Wang, D.; Yang, X.Y.; Lu, H.C. GradNet: Gradient-Guided Network for Visual Object Tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6162–6171.
40. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Convolutional Features for Correlation Filter Based Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Santiago, Chile, 7–13 December 2015; pp. 58–66.
41. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4021–4029.
42. Zhu, Z.; Huang, G.; Zou, W.; Du, D.; Huang, C. UCT: Learning Unified Convolutional Networks for Real-Time Visual Tracking. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 1973–1982.
43. Wu, Y.; Lim, J.; Yang, M. Object Tracking Benchmark. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*; IEEE: Piscataway, NJ, USA, 2015; Volume 37; pp. 1834–1848.
44. Yi, W.; 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.
45. Fan, H.; Ling, H.B. Parallel Tracking and Verifying: A Framework for Real-Time and High Accuracy Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5486–5494.
46. Li, X.; Ma, C.; Wu, B.Y.; He, Z.Y.; Yang, M.H. Target-Aware Deep Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1369–1378.
47. Mueller, M.; Smith, N.; Ghanem, B. A Benchmark and Simulator for UAV Tracking. In Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016.Convertini, N.; Dentamaro, V.; Impedovo, D.; Pirlo, G.; Sarcinella, L. A Controlled Benchmark of Video Violence Detection Techniques. *Information* **2020**, *11*, 321, doi:10.3390/info11060321.