Probabilistic Adaptive Spatial-Temporal Regularized Correlation Filters for UAV Tracking

Rui Li LXR1010@139.COM
Southwest University

Xiao Li IVY@SWU.EDU.CN
Southwest University

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

Most existing trackers based on spatial-temporal regularized correlation filters exploit response map variation to adapt regularization terms to object appearance changes automatically. However, these trackers ignore the high uncertainty of the response map when the object is occluded or similar objects around, making them unable to learn reliable filters accurately. Furthermore, most correlation filters use linear interpolation directly to update the filter model at each frame, which may cause model degradation once the tracking result is inaccurate or missing. In this work, we propose a novel probabilistic adaptive spatial-temporal regularized correlation filters (PASTRCF) to solve the two issues mentioned above. A probabilistic model constructing the reliability of the response map is introduced to accurately utilize the information in the response map to learn regularization coefficients adaptively. The adaptive threshold mechanism provides an appropriate strategy to update the filter model to alleviate model degradation. Extensive experiments on UAV benchmarks have proven the favorable performance of our method compared to the state-of-art trackers, with robust tracking while ensuring real-time performance.

Keywords: UAV Tracking, Spatio-temporal regularization, Correlation filters.

1. Introduction

Visual tracking is a fundamental and challenging computer vision problem (Yilmaz et al., 2006), and it has been widely applied in numerous fields, especially in unmanned aerial vehicle (UAV) applications, where it has been used for aerial cinematography, people following, traffic patrolling, and wildlife rescue. However, robust and accurate tracking has remained a challenging task (Fu et al., 2022) due to fast camera motion, small objects, background clutter, large occlusion, and similar objects around et al.

There are currently two main research focuses in the object tracking field: correlation filter-based methods and deep learning-based algorithms. The power capacity and computational resources of UAVs are limited. Deep learning-based tracking methods (Li et al., 2018a; Danelljan et al., 2019; Dai et al., 2020) require a high computational load due to complex deep features, which reduces the tracking speed and makes it hard to meet the real-time demands of UAV tracking. Correlation filters (Bolme et al., 2010; Henriques et al., 2012) use handcrafted features and convert the complicated correlation operations to element-wise multiplications in the frequency domain, which reduces the computational

complexity and significantly improves the tracking speed. Therefore, correlation filters can meet the real-time demands of UAV tracking.

Although the early correlation filters are effective, two major imperfections exist boundary effect and filter degradation. Danelljan et al. (2015), Galoogahi et al. (2017) and Li et al. (2018b) exploited predefined spatial regularization constraints to alleviate the boundary effect, but these constraints are fixed and cannot adapt to various changes. To solve this issue, Huang et al. (2019), Pu et al. (2020) and Li et al. (2020) utilized response maps to adapt regularization coefficients. However, the high uncertainty of response maps is ignored, which cannot accurately describe the similarity between the object and appearance model when encountering occlusion and similar objects around (see Fig. 1). Temporal regularization (Li et al., 2018b; Dai et al., 2019; Li et al., 2020; Zhang et al., 2021) is applied to mitigate model degradation, but these trackers use a linear interpolation update strategy, which still leads to model degradation due to occlusion or inaccurate tracking results.

In this work, we develop a novel tracking approach that resolves the abovementioned problems,i.e., the PASTRCF tracker. A probabilistic model and an adaptive threshold mechanism are introduced respectively for measuring the reliability of response maps to adapt regularization terms and for providing appropriate update strategies to mitigate model degradation. PASTRCF performs favorably compared with the state-of-the-art trackers while meeting the real-time demands of UAV tracking. The contributions of this work can be summarized as follows:

- A new probabilistic model describing the reliability of response maps is proposed, which could accurately exploit the information in response maps to efficiently learn adaptive regularization terms and obtain more reliable filter coefficients.
- An adaptive threshold mechanism is introduced to assist model update, which alleviates the model degradation caused by inappropriate model update strategy in the case of occlusion or inaccurate tracking results.
- Our tracker is evaluated on the UAV benchmarks UAVDT and DTB70. Extensive experiments show that PASTRCF, which achieves very remarkable performance with a real-time speed, can effectively solve the challenging problems of fast camera motion, small objects, similar objects around, and large occlusion in UAV scenes.

2. Related Work

Correlation filters (CF) have been widely used in UAV tracking recently. MOSSE (Bolme et al., 2010) is the earliest CF-based tracker, which used only grayscale samples to train the filter. Henriques et al. (2012) exploited cyclic shift samples and introduced kernel functions to optimize the filter coefficients in the frequency domain efficiently. Henriques et al. (2015) extended CSK to multi-channel HOG features to enhance the feature representation ability. Similarly, Danelljan et al. (2014) introduced the color naming features to achieve robust tracking. To cope with scale variation, Danelljan et al. (2017b) and Li and Zhu (2015) utilized multi-scale searching strategies to adapt scale estimation.

The boundary effect caused by cyclic shift samples reduces the recognition ability of filters. To address this issue, Danelljan et al. (2015) used spatial regularization weights, and

Galoogahi et al. (2017) extracted more realistic training samples from the background to alleviate the boundary effect effectively. Both spatial constraints have been extensively studied (Lukežic et al., 2017; Galoogahi et al., 2017; Mueller et al., 2017). Li et al. (2018b) introduced a temporal regularization term to alleviate model degradation. However, these trackers used constant regularization terms, which cannot adapt to object appearance changes.

Since 2019, most researchers have focused on adaptive spatial-temporal regularization coefficients to adapt object appearance automatically. Dai et al. (2019) learned reliable spatial weights for specific object appearance variations. Li et al. (2020) fully exploited response map variation to learn automatic spatial-temporal regularization terms. Chu et al. (2021) proposed a re-detection strategy and anti-occlusion method based on AutoTrack to achieve better performance. However, the high uncertainty of response maps is ignored, and they all use linear interpolation strategies to update the filter model. Unlike the above studies, we propose a probabilistic model, which measures the reliability of response maps, to learn more reliable spatial-temporal regularization terms and develop an adaptive threshold mechanism to provide an appropriate model update strategy, improving tracking robustness.

3. Proposed Method

In this section, the probabilistic model describing the reliability of response maps is introduced in section 3.1, which fully exploits the foreground and background information of response maps to adapt regularization coefficients accurately. A dynamic threshold mechanism is proposed as an update strategy to assist model update in section 3.2.

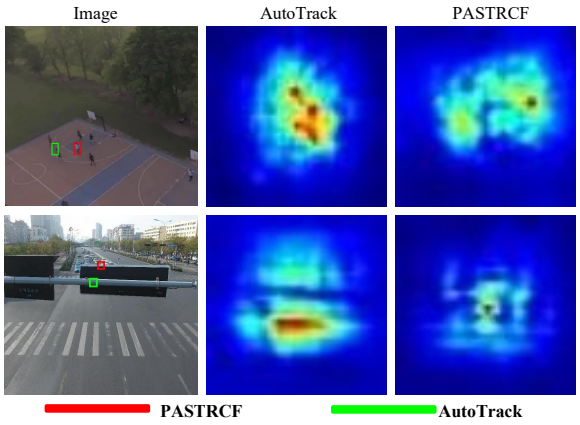


Figure 1: The Visualization of the high uncertainty of response maps.

3.1. Probabilistic Model

Response maps contain information regarding the resemblance of the current object and the appearance model. However, the response map will become highly uncertain when the object is occluded or similar objects around. As shown in Fig. 1, the first row depicts a challenging sequence with multiple distractors. Due to their similarity and proximity, the

response map generated by AutoTrack has many high response locations, which causes it briefly fail by jumping to a distractor object. In the second row, due to occlusion, the response map of AutoTrack is highly uncertain, providing an ambiguous and inaccurate choice to localize the object. Unlike AutoTrack, we proposed a probabilistic model to accurately capture the uncertainty in ambiguous cases and provide a strong secondary mode in the object center with a high response in the actual object position.

We learn the reliability of response maps using a Bayesian classifier from color distributions of the object and background region of the response map. Let $\mathrm{R_G}$ denote the object region of the response map, and $\mathrm{R_B}$ denotes the background region of the response map. Given a search region O, the probability of pixel θ in the object region of the response map is computed as follows:

$$\mathrm{P}\left(\theta \in \mathrm{R_G} \mid \mathrm{i}\right) = \frac{\mathrm{P}\left(\mathrm{i} \mid \theta \in \mathrm{R_G}\right) \mathrm{P}\left(\theta \in \mathrm{R_G}\right)}{\sum_{\mathrm{O} \in (\mathrm{R_G}, \mathrm{R_B})} \mathrm{P}\left(\mathrm{i} \mid \theta \in \mathrm{O}\right) \mathrm{P}\left(\theta \in \mathrm{O}\right)} \tag{1}$$

where i denotes the bin of color histograms, $\mathrm{H_i}\left(\mathrm{X}\right)$ denotes the bin i in the color histogram $\mathrm{H}\left(\mathrm{X}\right)$ of response map, $\mid \mathrm{C} \mid$ is the cardinality of region C. To improve computational efficiency, we approximate the likelihood terms based on the distribution of color histogram (Possegger et al., 2015):

$$\begin{aligned} &\mathrm{P}(\theta \in \mathrm{R_G}) \approx |\mathrm{R_G}|/(|\mathrm{R_G}| + |\mathrm{R_B}|) \\ &\mathrm{P}\left(\mathrm{i} \mid \theta \in \mathrm{R_G}\right) \approx \mathrm{H_i}\left(\mathrm{R_G}\right)/|\mathrm{R_G}| \\ &\mathrm{P}\left(\mathrm{i} \mid \theta \in \mathrm{R_B}\right) \approx \mathrm{H_i}\left(\mathrm{R_B}\right)/|\mathrm{R_B}| \end{aligned} \tag{2}$$

Therefore, the reliability of pixel θ in the response map can be simplified as:

$$\mathrm{P}\left(\theta \in \mathrm{R_G} \mid \mathrm{i}\right) = \mathrm{H_i}\left(\mathrm{R_G}\right)/\left(\mathrm{H_i}\left(\mathrm{R_G}\right) + \mathrm{H_i}\left(\mathrm{R_B}\right)\right) \tag{3}$$

Finally, by solving (3), we can get a probabilistic matrix that describes the reliability of each pixel in the response map, which can be used to learn effective regularization terms and obtain more robust filter coefficients in the training phase.

3.2. Adaptive Threshold Mechanism

Most existing trackers update filter models at each frame without considering whether tracking results are accurate or not, which causes model degradation once severe occlusion or inaccurate tracking results exist. Wang et al. (2017) introduced Average Peak-to Correlation Energy (APCE) to solve this problem, which indicates the fluctuated degree of response maps and the confidence level of detected objects. It is defined as follows:

$$\mathrm{APCE} = \frac{\left|\mathrm{F_{max}} - \mathrm{F_{min}}\right|^2}{\mathrm{mean}\left(\sum_{\mathrm{w,h}} \left(\mathrm{F_{w,h}} - \mathrm{F_{min}}\right)^2\right)} \tag{4}$$

Where $\mathrm{F_{max}}$, $\mathrm{F_{min}}$ and $\mathrm{F_{w,h}}$ denote the maximum, minimum and the w-th row h-th column elements of the response map. The historical average response values are generally chosen as the APCE threshold. The filter will be updated when APCE is greater than the threshold, which indicates the tracking result is accurate and the object is not occluded,

otherwise, the update is stopped. Although this method is effective, the variation tendency between adjacent frames is neglected and the single threshold cannot update the filter model adaptively. For this, the APCE of the last three frames is used to design a dynamic threshold mechanism in our work for updating model adaptively. The dynamic threshold of the i-th frame is designed as follows:

$$\mathrm{F}_{\mathrm{threshold}} = \left(\sqrt{(\alpha + \mathrm{e}) \times \mathrm{M}}\right)^{-\alpha} \times 10^5, \left(\mathrm{M} = \frac{1}{3}\sum_{\mathrm{i=t-2}}^{\mathrm{t}} \mathrm{APCE(i)}\right) \tag{5}$$

When objects are occluded continuously, or inaccurate tracking results exist, M will decrease and the threshold will increase to avoid updating model, which prevents the filter model from being polluted and alleviates model degradation.

4. Tracking Approach

In this section, we introduce the overall tracking process of the proposed method. The main structure can be seen in Fig. 2. We will introduce the objective function used to train the filter model in section 4.1 and its optimization step is shown in section 4.2. Then, the object localization and model update will be shown in section 4.3 and section 4.4 respectively.

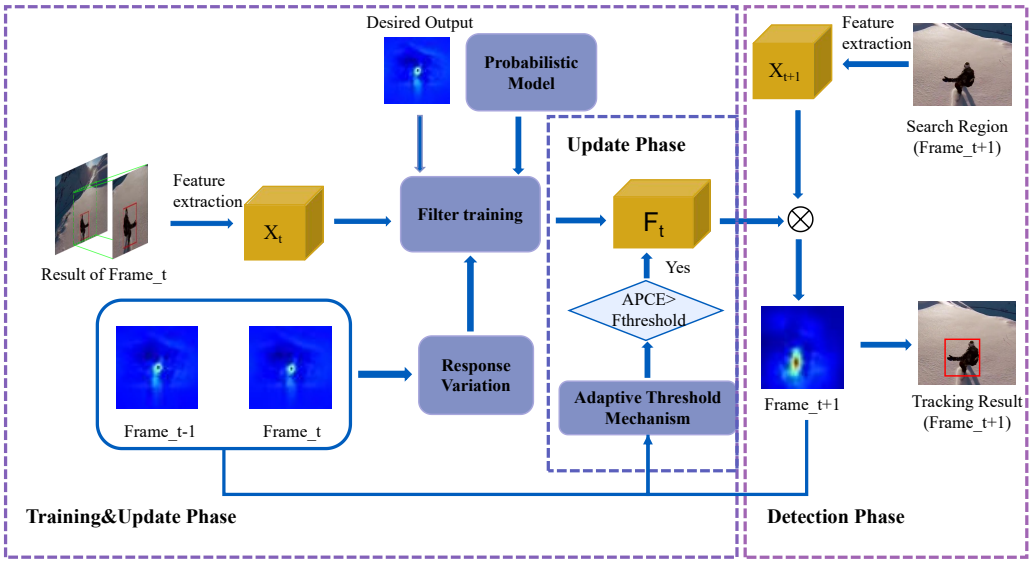


Figure 2: The flowchart of our proposed PASTRCF tracker. Our proposed probabilistic model combined with response map variations is used in the filter training phase to learn a robust filter. The trained filter is exploited in the detection phase to obtain the tracking response map and localize the object. The response maps of the last three frames are exploited to design the adaptive threshold mechanism, which is used in the model update phase to update the filter effectively.

4.1. Objective Function of Our PASTRCF Model

Our baseline AutoTrack: Li et al. (2020) fully exploits the hidden local or global information in response maps to adapt the spatio-temporal regularization term and learns the correlation filter by minimizing the following objective function:

$$\mathrm{E}\left(\mathrm{H_t}, \mu_\mathrm{t}\right) = \frac{1}{2}\left\|\mathrm{y} - \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{x_t^k} * \mathrm{h_t^k}\right\|_2^2 + \frac{1}{2}\sum_{\mathrm{k}=1}^{\mathrm{K}}\left\|\tilde{\omega} \circ \mathrm{h_t^k}\right\|_2^2 + \frac{\mu_\mathrm{t}}{2}\sum_{\mathrm{k}=1}^{\mathrm{K}}\left\|\mathrm{h_t^k} - \mathrm{h_{t-1}^k}\right\|_2^2 + \frac{1}{2}\left\|\mu_\mathrm{t} - \tilde{\mu}\right\|_2^2 \quad (6)$$

Where $\mathrm{y} \in \mathrm{R}^{\mathrm{T}\times 1}$ is the desired response output, $\mathrm{x_t^k} \in \mathrm{R}^{\mathrm{T}\times 1}(\mathrm{k} = 1, 2, 3, \ldots \mathrm{K})$ is the extracted feature with length T in frame t, and K denotes number of channel. $\mathrm{h_t^k}, \mathrm{h_{t-1}^k} \in \mathrm{R}^{\mathrm{T}\times 1}$ denote the filter of the k-th channel trained in frame t and frame t-1 respectively, $\mathrm{H_t} = \left[\mathrm{h_t^1},\ \mathrm{h_t^2},\ \mathrm{h_t^3}, \ldots, \mathrm{h_t^k}\right]$, μ_t is temporal regularization coefficient, $\tilde{\omega}$, $\tilde{\mu}$ denote automatic spatial regularization parameter and reference temporal regularization parameter respectively, $*$ represents correlation operation, $\circ$ is element-wise multiplications operation.

Although AutoTrack has achieved advanced performance on UAV benchmarks, there are still two limitations: 1) the high uncertainty of response maps is not taken into account when occlusion or similar objects around, and 2) there exists model degradation due to updating the model at each frame without considering whether the tracking result is accurate.

Our Objective Function: Motivated by AutoTrack and combined with the probabilistic model that we introduced in section 3.1, our objective function is defined by:

$$\begin{aligned}\mathrm{E}\left(\mathrm{H_t}, \mu_\mathrm{t}\right) = &\frac{1}{2}\left\|\mathrm{y} - \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{x_t^k} * \left(\mathrm{P^T h_t^k}\right)\right\|_2^2 + \frac{1}{2}\sum_{\mathrm{k}=1}^{\mathrm{K}}\left\|\tilde{\omega} \circ \mathrm{h_t^k}\right\|_2^2 \\ &+ \frac{\mu_\mathrm{t}}{2}\sum_{\mathrm{k}=1}^{\mathrm{K}}\left\|\mathrm{h_t^k} - \mathrm{h_{t-1}^k}\right\|_2^2 + \frac{1}{2}\left\|\mu_\mathrm{t} - \tilde{\mu}\right\|_2^2\end{aligned} \quad (7)$$

Where P is the probability matrix obtained by our probability model, which describes the reliability of each pixel in the response map. By exploiting the response probability of the previous frame, we can obtain a more efficient filter model at the current frame.

4.2. Model Optimization with ADMM

For effective optimization, we introduce an auxiliary variable $\hat{\mathrm{g}}_\mathrm{t}^\mathrm{k}$ by ordering $\hat{\mathrm{g}}_\mathrm{t}^\mathrm{k} = \sqrt{\mathrm{T}}\mathrm{F}\mathrm{P}^\mathrm{T}\mathrm{h}_\mathrm{t}^\mathrm{k}$ $\left(\hat{\mathrm{G}} = \left[\hat{\mathrm{g}}_\mathrm{t}^1, \hat{\mathrm{g}}_\mathrm{t}^2, \hat{\mathrm{g}}_\mathrm{t}^3, \cdots, \hat{\mathrm{g}}_\mathrm{t}^\mathrm{k}\right]\right)$, where $\mathrm{F} \in \mathrm{C}^{\mathrm{T}\times\mathrm{T}}$ represents the orthogonal matrix to convert any T-dimensional vectorized signal into the Fourier domain, $\hat{\mathrm{X}}$ represents the discrete Fourier transform form of X. Then we can express the objective function (7) in frequency domain:

$$\begin{aligned}\mathrm{E}\left(\mathrm{H_t}, \mu_\mathrm{t}, \hat{\mathrm{G}}_\mathrm{t}\right) = &\frac{1}{2}\left\|\mathrm{y} - \sum_{\mathrm{k}=1}^{\mathrm{K}} \hat{\mathrm{x}}_\mathrm{t}^\mathrm{k} \circ \hat{\mathrm{g}}_\mathrm{t}^\mathrm{k}\right\|_2^2 + \frac{1}{2}\sum_{\mathrm{k}=1}^{\mathrm{K}}\left\|\tilde{\omega} \circ \mathrm{h_t^k}\right\|_2^2 \\ &+ \frac{\mu_\mathrm{t}}{2}\left\|(\sqrt{\mathrm{T}}\mathrm{F}\mathrm{P}^\mathrm{T})^{-1}\left(\hat{\mathrm{g}}_\mathrm{t}^\mathrm{k} - \hat{\mathrm{g}}_\mathrm{t-1}^\mathrm{k}\right)\right\|_2^2 + \frac{1}{2}\left\|\mu_\mathrm{t} - \tilde{\mu}\right\|_2^2\end{aligned} \quad (8)$$

(8) can be solved by alternating direction method of multipliers (ADMM). The Augmented Lagrangian form of (8) can be formulated as:

$$\begin{aligned} \mathrm{L_t}\left(\mathrm{H_t},\mu_t,\hat{\mathrm{G}}_t,\hat{\mathrm{M}}_t\right) &= \mathrm{E}\left(\mathrm{H_t},\mu_t,\hat{\mathrm{G}}_t\right)+\frac{\gamma}{2}\sum_{k=1}^{K}\left\|\hat{g}_t^k-\sqrt{T}FP^Th_t^k\right\|_2^2 \\ &+\sum_{k=1}^{K}\left(\hat{g}_t^k-\sqrt{T}FP^Th_t^k\right)^T\hat{m}_t^k \end{aligned} \tag{9}$$

Where $\hat{\mathrm{M}}_t=[\hat{m}_1,\hat{m}_2,\ldots,\hat{m}_k]\in R^{T\times K}$ is Fourier transform of Lagrange multiplier, γ is the step size of regularization parameter. By introducing $v_t^k=m_t^k/\gamma\left(V_t^k=\left[v_t^1,v_t^2,\ldots,v_t^K\right]\right)$, the optimization of (9) is equivalent to solving (10).

$$\mathrm{L_t}\left(\mathrm{H_t},\mu_t,\hat{\mathrm{G}}_t,\hat{\mathrm{V}}_t\right)=\mathrm{E}\left(\mathrm{H_t},\mu_t,\hat{\mathrm{G}}_t\right)+\frac{\gamma}{2}\sum_{k=1}^{K}\left\|\hat{g}_t^k-\sqrt{T}FP^Th_t^k+\hat{v}_t^k\right\|_2^2 \tag{10}$$

Then, the ADMM algorithm is adopted by alternately solving the following subproblems.

Subproblem $\hat{\mathrm{G}}$: if $\mathrm{H_t}$, μ_t and $\hat{\mathrm{V}}_t$ are given, the optimal $\hat{\mathrm{G}}^*$ can be obtained as:

$$\hat{\mathrm{G}}^*=\arg\min_{\hat{\mathrm{G}}}\left\{\frac{1}{2}\left\|\hat{y}-\sum_{k=1}^{K}\hat{x}_t^k\circ\hat{g}_t^k\right\|_2^2+\frac{\mu_t}{2}\sum_{k=1}^{K}\left\|\hat{g}_t^k-\hat{g}_{t-1}^k\right\|_2^2+\frac{\gamma}{2}\sum_{k=1}^{K}\left\|\hat{g}_t^k-\sqrt{T}FP^Th_t^T+\hat{v}_t^k\right\|_2^2\right\} \tag{11}$$

It is difficult to solve (11) directly due to its high complexity. So we consider processing on all channels of each pixel and reformulate our formulation as

$$\begin{aligned} \Pi_j^*\left(\hat{\mathrm{G}}_t\right)=\underset{\Pi_j(\hat{\mathrm{G}}_i)}{\operatorname{argmin}}\Bigg\{&\left\|\hat{y}_j-\Pi_j\left(\hat{\mathrm{X}}_t\right)^T\Pi_j\left(\hat{\mathrm{G}}_t\right)\right\|_2^2+\mu_t\left\|(\sqrt{T}FP^T)^{-1}\left(\Pi_j\left(\hat{\mathrm{G}}_t\right)-\Pi_j\left(\hat{\mathrm{G}}_{t-1}\right)\right)\right\|_2^2 \\ &+\gamma\left\|\Pi_j\left(\hat{\mathrm{G}}_t\right)+\Pi_j\left(\hat{\mathrm{V}}_t\right)-\Pi_j\left(\sqrt{T}FP^TH_t\right)\right\|_2^2\Bigg\} \end{aligned} \tag{12}$$

Where $\Pi_j(\hat{X})\in C^{K\times 1}$ denotes the vector of all K channels values on pixel j (j=1,2,3,... ,T). The derivation of (11) can use Sherman Morrison formula to obtain analytical solution:

$$\Pi_j^*\left(\hat{\mathrm{G}}_t\right)=\frac{1}{\gamma+B}\left(I-\frac{\Pi_j\left(\hat{\mathrm{X}}_t\right)\Pi_j\left(\hat{\mathrm{X}}_t\right)^T}{\Pi_j\left(\hat{\mathrm{X}}_t\right)^T\Pi_j\left(\hat{\mathrm{X}}_t\right)+\gamma+B}\right)\rho \tag{13}$$

Where $\rho=\Pi_j\left(\hat{\mathrm{X}}_t\right)\hat{y}_j+\gamma\Pi_j\left(\sqrt{T}FP^TH_t\right)-\gamma\Pi_j\left(\hat{\mathrm{V}}_t\right)+B\Pi_j\left(\hat{\mathrm{G}}_{t-1}\right)$. Note that (13) only contains vector multiply-add operation and thus can be computed efficiently.

Subproblem H: if $\mu_t,\hat{\mathrm{G}}_t,\hat{\mathrm{V}}_t$ are fixed, the optimal solution of h^k can be efficiently calculated and solution is presented below:

$$h^{k^*}=\arg\min_{h^k}\left\{\frac{1}{2}\left\|\tilde{\omega}\circ h_t^k\right\|_2^2+\frac{\gamma}{2}\left\|\hat{g}_t^k-\sqrt{T}FP^Th_t^k+\hat{v}_t^k\right\|_2^2\right\} \tag{14}$$

We introduce a diagonal matrix $W = \mathrm{diag}(\tilde{\mu}) \in R^{T\times T}$, and then the closed-form solution of h^k can be computed by:

$$h^{k^*} = \left[W^T W + \gamma TP\right]^{-1} \gamma TP \left(g_t^k + v_t^k\right) = \frac{\gamma TP\left(g_t^k + v_t^k\right)}{\widetilde{\omega}^T \circ \widetilde{\omega} + \gamma TP} \tag{15}$$

Subproblem μ_t: the solution to subproblem μ_t can be easily obtained as follows:

$$\begin{aligned} \mu_t^* &= \arg\min_{\mu_t} \left\{ \frac{\mu_t}{2} \left\| \left(\sqrt{T} F P^T\right)^{-1} \left(\hat{g}_t^k - \hat{g}_{t-1}^k\right) \right\|_2^2 + \frac{1}{2} \left\|\mu_t - \widetilde{\mu}\right\|_2^2 \right\} \\ &= \widetilde{\mu} - \frac{\sum_{k=1}^K \left\| TP\left(g_t^k - g_{t-1}^k\right)\right\|_2^2}{2} \end{aligned} \tag{16}$$

Lagrangian Multiplier Update: We update Lagrangian multipliers as:

$$\hat{V}^{i+1} = \hat{V}^i + \gamma^i \left(\hat{G}^{i+1} - \hat{H}^{i+1}\right), \quad \left(\gamma^{i+1} = \min\left(\gamma_{\max}, \beta\gamma^i\right)\right) \tag{17}$$

Where i and i + 1 denote the iteration index, it is worth mentioning that our objective function has a better convergence speed and only needs three iterations to get the optimal solution. The regularization constant γ represents step size, $\hat{G}^{i+1}$ and $\hat{H}^{i+1}$ are solutions of two subproblems above at iteration i+1. By iteratively solving the four subproblems above, our objective function can be solved efficiently. The optimized filter $\hat{G}_t$ and spatial-temporal regularization terms can be obtained in frame t, which will be used in frame t+1.

4.3. Object Localization

The tracking object is localized by searching for the maximum value of the response map, which is determined by:

$$R_t = \Gamma^{-1} \sum_{k=1}^{K} \left(\hat{x}_t^k \circ \hat{g}_{t-1}^k\right) \tag{18}$$

Where R_t is the response map at frame t, Γ^{-1} represents the inverse Fourier transform operator and $\hat{x}_t^k$ denotes the Fourier transform of the feature vector extracted from the frame t. Through (18), the optimal object location can be obtained.

4.4. Model Update

Different from other CF-based trackers, our dynamic threshold mechanism can update the tracking model adaptively. Our tracker is not updated when the object is occluded or the tracking result is inaccurate (APCE < Fthreshold). Otherwise, the update method of the filter model is as follows:

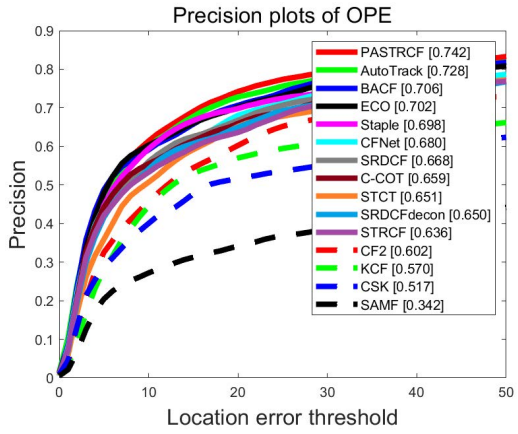
$$h_t = (1 - \lambda) h_{t-1} + \lambda h_t \tag{19}$$

Where λ represents the online learning rate, h_t and h_{t-1} denote the filter in frame t and frame t-1 respectively. The model update strategy we proposed does not require updating every frame, which not only reduces the effect of occlusion but also improves tracking speed.

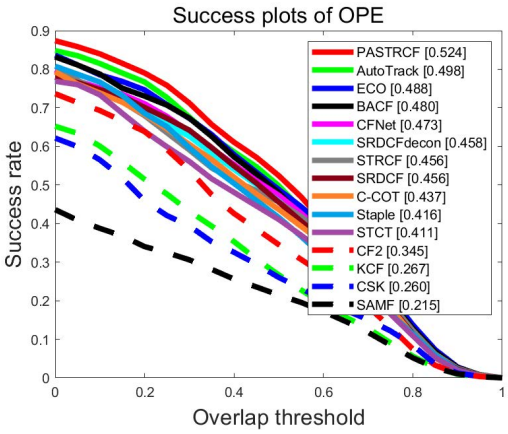
5. Experiments

In this section, we firstly evaluate our tracker with 15 current state-of-the-art trackers including AutoTrack (Li et al., 2020), BACF (Galoogahi et al., 2017), C−COT (Danelljan et al., 2016b), CF2 (Ma et al., 2015), CFNet (Shen et al., 2021), CSK (Henriques et al., 2012), ECO (Danelljan et al., 2017a), KCF (Henriques et al., 2015), CN (Danelljan et al., 2014), SRDCF (Danelljan et al., 2015), SRDCFdecon (Danelljan et al., 2016a), Staple (Bertinetto et al., 2016), STRCF (Li et al., 2018b), STCT (Wang et al., 2016) and SAMF (Li and Zhu, 2015) on UAV and DTB70 benchmarks. Then, we conduct the ablation study to verify the effectiveness of each module. We use the same evaluation criteria in two benchmarks and all details can be found in Dawei et al. (2018) and Li and Yeung (2017).

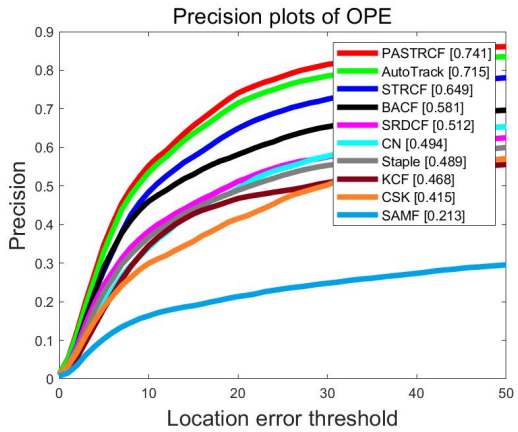
Implementation details: ADMM iteration and the index of dynamic threshold α are both set to 3. The other hyperparameters are the same as our baseline AutoTrack. All experiments are conducted using MATLAB R2019b on a PC with AMD Ryzen 7 5800H CPU, 32GB RAM, and NVIDIA RTX 3060Ti (6GB RAM) GPU.



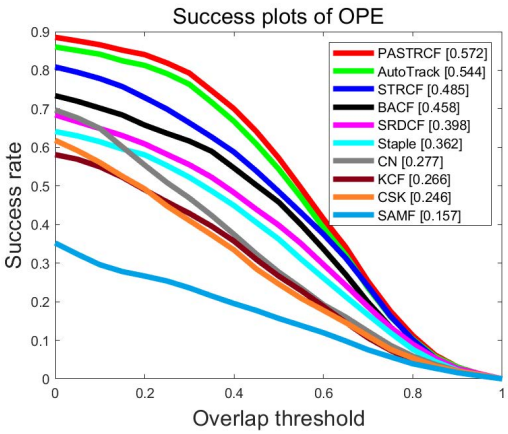
(a) Precision plots on UAVDT.



(b) Success plots on UAVDT.



(c) Precision plots on DTB70



(d) Success plots on DTB70

Figure 3: Overall performance on UAVDT and DTB70.

5.1. Quantitative Evaluation

Overall performance evaluation: The overall performance evaluation of PASTRCF with other state-of-the-art trackers on UAVDT and DTB70 benchmarks is illustrated in Fig. 3. The proposed PASTRCF tracker has outperformed all other trackers on both datasets. More specially, on UAVDT benchmark, PASTRCF has an advantage of 1.4 % and 2.6 % in precision and success rate over the second tracker (AutoTrack) respectively. On DTB70 benchmark, PASTRCF also achieves the best performance and obtains 2.6 % and 2.8 % gain in precision and success rate respectively than AutoTrack. Overall, the proposed PASTRCF tracker achieves superiority and effectiveness against other state-of-the-art trackers.

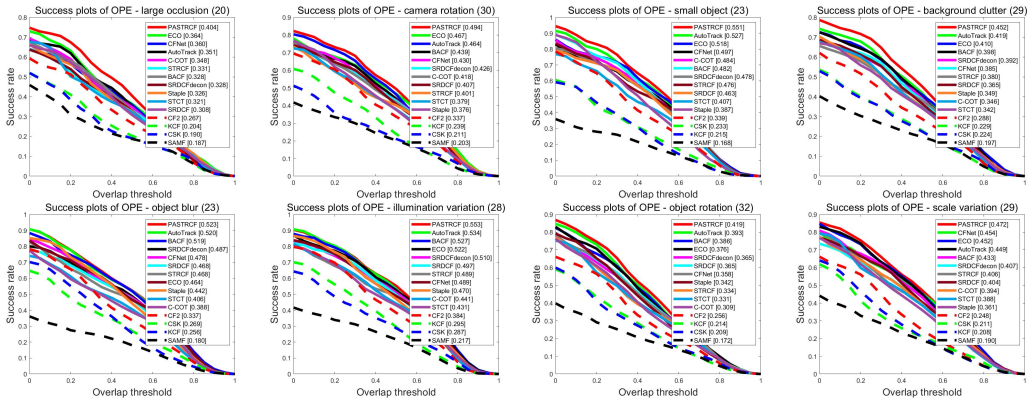


Figure 4: Attribute-based comparison on UAVDT.

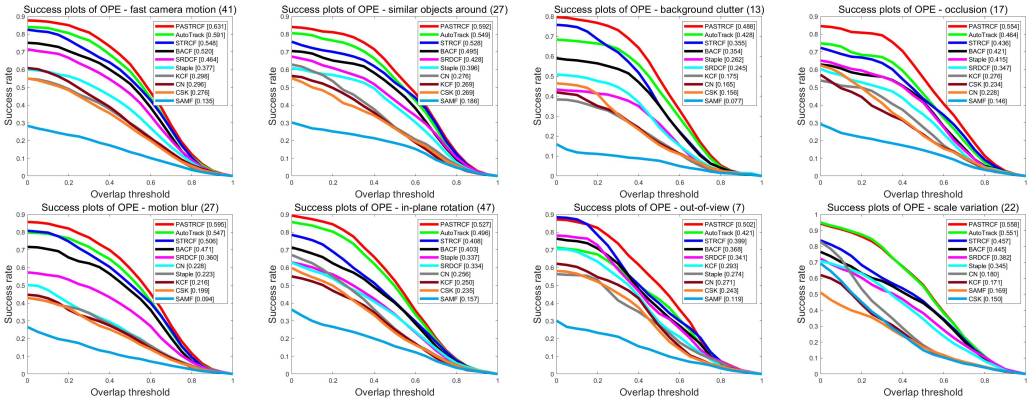


Figure 5: Attribute-based comparison on DTB70.

There may be two reasons why PASTRCF outperforms on both benchmarks: 1) Our probabilistic model can successfully describe the reliability of response maps to alleviate the ambiguous cases of response maps caused by occlusion or similar objects around, further improving the tracking performance. 2) Our adaptive threshold mechanism can provide an

appropriate model update strategy to prevent the filter from being polluted by inaccurate tracking results and mitigate model drift, thereby increasing tracking robustness.

Attribute-based evaluation: The success plots of different attributes on UAVDT benchmark and DTB70 benchmark are demonstrated in Fig. 4 and Fig. 5 respectively, and the precision results on UAVDT benchmark are demonstrated in Table 1. We can see that PASTRCF achieves almost the best performance in these attributes. It performs well under large occlusion (LO), camera rotation (CR), small object (SO), and background clutter (BC) on UAVDT, which obtains 5.3 %, 3.0 %, 2.4 %, 3.3 % gain respectively in success rate and gains advancement of 5.9 %, 1.6 %, 2.9 %, 3.9 % respectively in precision than AutoTrack. On DTB70 benchmark, the success rate of PASTRCF in fast camera motion (FCM) and similar objects around (SOA) attributes obtain 4.0 % and 4.3 % gain respectively than AutoTrack. In other attributes, PASTRCF has also shown great improvement from AutoTrack and achieved performance with a high ranking.

Table 1: Precision comparision results of 8 attributes on UAVDT benchmark. Red, green and blue respectively mean the first, second and third place.

Attribute	**ALL**	**BC**	**CR**	**IV**	**LO**	**OB**	**OR**	**SV**	**SO**
PASTRCF	0.742	0.668	0.695	0.801	0.553	0.779	0.682	0.658	0.857
AutoTrack	0.728	0.629	0.679	0.805	0.494	0.778	0.663	0.637	0.828
BACF	0.706	0.633	0.648	0.775	0.487	0.743	0.635	0.604	0.815
C−COT	0.659	0.557	0.623	0.720	0.460	0.662	0.561	0.559	0.792
CF2	0.602	0.486	0.569	0.679	0.381	0.651	0.482	0.453	0.695
CFNet	0.680	0.567	0.643	0.727	0.447	0.717	0.599	0.611	0.775
CSK	0.517	0.452	0.427	0.566	0.344	0.541	0.437	0.439	0.587
ECO	0.702	0.611	0.644	0.769	0.508	0.710	0.627	0.632	0.791
KCF	0.570	0.458	0.534	0.657	0.344	0.652	0.454	0.490	0.581
SRDCF	0.668	0.583	0.634	0.728	0.451	0.692	0.589	0.588	0.747
SRDCFdecon	0.650	0.574	0.610	0.723	0.425	0.695	0.578	0.549	0.738
Staple	0.698	0.607	0.673	0.787	0.474	0.752	0.625	0.589	0.842
STRCF	0.636	0.530	0.565	0.672	0.420	0.657	0.528	0.543	0.764
STCT	0.651	0.560	0.613	0.699	0.466	0.633	0.575	0.599	0.710
SAMF	0.342	0.312	0.338	0.367	0.265	0.302	0.259	0.287	0.330

Our tracker achieves excellent performance in UAV scenarios, where CR causes object appearance variation, FCM and BC reflect boundary effect, and LO, SOA, and SO are also observable challenges. We mainly attribute this to our proposed tracking approach. The probabilistic model exploits the reliability of response maps to resolve an ambiguous case of response maps caused by SOA and LO, which learns effective spatial-temporal regularization terms to alleviate the boundary effect caused by FCM and BC, and obtains a more robust filter to adapt the appearance variation caused by CR. Then, the adaptive threshold mechanism will set a dynamic threshold to avoid updating the filter when the object is severely occluded and prevent the filter from being polluted, achieving a more stable filter model and significantly improving the tracking robustness.

5.2. Qualitative Evaluation

Some qualitative tracking results of PASTRCF and other advanced trackers are shown in Fig. 6. The videos (from top to bottom) are S0310, S0601, and S0801 on UAVDT benchmark, SnowBoarding4, Basketball, and RcCar4 on DTB70 benchmark. The qualitative results can prove that PASTRCF achieves robust tracking and it is competent in dealing with different challenging attributes, especially in large occlusion, background clutter, camera motion, similar objects around, and small object.

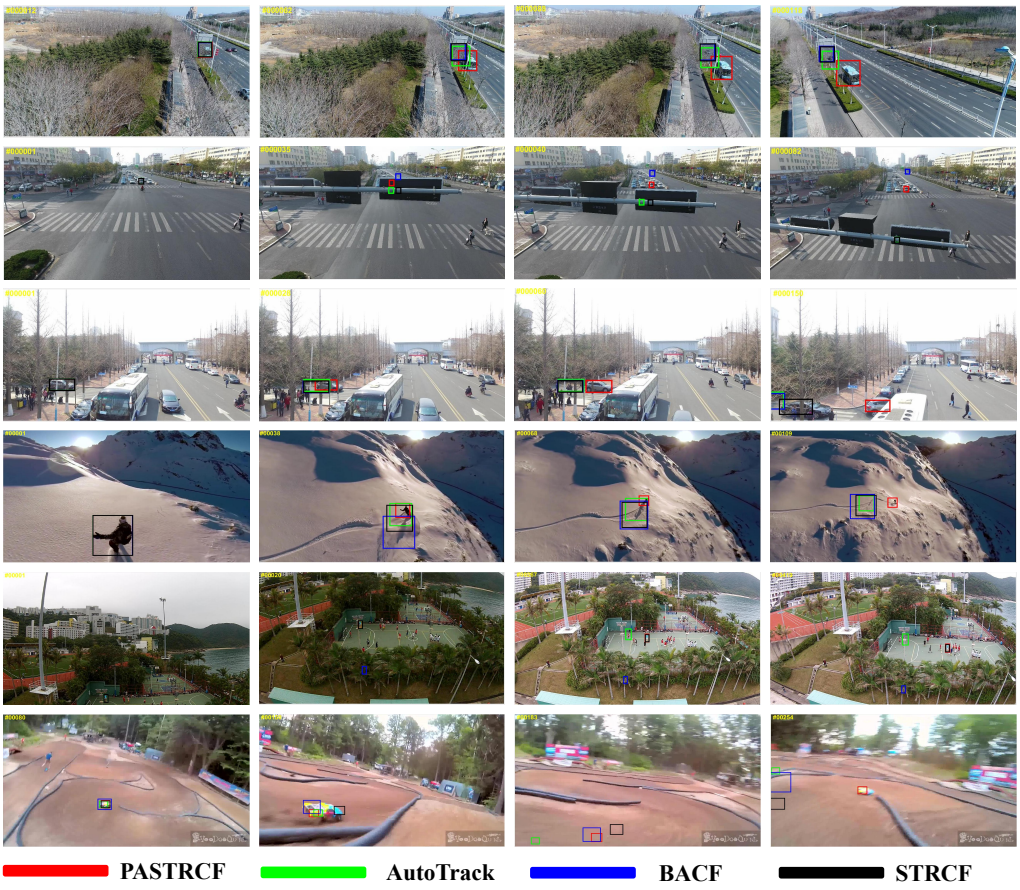


Figure 6: Qualitative evaluation of our proposed method with comparison to three trackers (AutoTrack, BACF and STRCF).

5.3. Ablation Study

To verify the effectiveness of different components in our tracker, an ablation study of PASTRCF is performed in this section. The overall evaluation is shown in Table 2. 1) AutoTrack is the baseline tracker without our probabilistic model as well as the adaptive threshold update strategy, 2) AutoTrack_PM is the baseline with our probabilistic model,

3) AutoTrack_MU is the baseline with the adaptive threshold mechanism, 4) PASTRCF is our proposed tracker with probability model as well as the adaptive threshold mechanism.

Table 2: Ablation study of PASTRCF on UAVDT.

Tracker	**AutoTrack**	**AutoTrack_PM**	**AutoTrack_MU**	**PASTRCF**
Precision	0.728	0.740	0.738	0.742
Success	0.498	0.521	0.516	0.524
Speed	38.272	37.580	40.963	40.824

It can be seen from Table 2 that our probabilistic model and adaptive threshold mechanism have contributed to the substantial improvement over our baseline method. Compared with our probabilistic model, our adaptive threshold mechanism increases the tracking speed because we only update the filter model when the tracking result is accurate and the object is not occluded, which decreases the update time and improves the execution speed.

6. Conclusions

In this work, a novel probabilistic adaptive spatial-temporal regularization correlation filters (PASTRCF) is proposed for robust UAV tracking. Compared with other trackers that use response variation to adapt regularization terms, we first introduce a probabilistic model to solve the uncertainty of response maps when objects encounter occlusion or similar objects around, and accurately exploit the reliable information in response maps for adaptively learning regularization terms. Then we design an adaptive model update mechanism to effectively mitigate model degradation and it also improves the tracking speed. In addition, our PASTRCF tracker is effectively optimized using the ADMM algorithm. Comprehensive experiments have validated that PASTRCF outperforms some state-of-the-art trackers on both UAVDT benchmark and DTB70 benchmark, not only achieving advanced performance but also meeting the real-time demands in UAV scenarios.

References

Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip H. S. Torr. Staple: Complementary learners for real-time tracking. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1401–1409, 2016. doi: 10.1109/CVPR.2016.156.

David S. Bolme, J. Ross Beveridge, Bruce A. Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2544–2550, 2010. doi: 10.1109/CVPR.2010.5539960.

Hongyu Chu, Kuisheng Liao, Yanhua Shao, Xiaoqiang Zhang, Yanying Mei, and Yadong Wu. Ao-autotrack: Anti-occlusion real-time uav tracking based on spatio-temporal context. In Huimin Ma, Liang Wang, Changshui Zhang, Fei Wu, Tieniu Tan, Yaonan Wang,

Jianhuang Lai, and Yao Zhao, editors, *Pattern Recognition and Computer Vision*, pages 256–267, Cham, 2021. Springer International Publishing.

Kenan Dai, Dong Wang, Huchuan Lu, Chong Sun, and Jianhua Li. Visual tracking via adaptive spatially-regularized correlation filters. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4665–4674, 2019. doi: 10.1109/CVPR.2019.00480.

Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6297–6306, 2020. doi: 10.1109/CVPR42600.2020.00633.

Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost Van De Weijer. Adaptive color attributes for real-time visual tracking. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1090–1097, 2014. doi: 10.1109/CVPR.2014.143.

Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg. Learning spatially regularized correlation filters for visual tracking. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4310–4318, 2015. doi: 10.1109/ICCV.2015.490.

Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg. Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1430–1438, 2016a. doi: 10.1109/CVPR.2016.159.

Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 472–488, Cham, 2016b. Springer International Publishing. ISBN 978-3-319-46454-1.

Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6931–6939, 2017a. doi: 10.1109/CVPR.2017.733.

Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg. Discriminative scale space tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1561–1575, 2017b. doi: 10.1109/TPAMI.2016.2609928.

Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4655–4664, 2019. doi: 10.1109/CVPR.2019.00479.

Dawei, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair

Weiss, editors, *Computer Vision – ECCV 2018*, pages 375–391, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01249-6.

Changhong Fu, Bowen Li, Fangqiang Ding, Fuling Lin, and Geng Lu. Correlation filters for unmanned aerial vehicle-based aerial tracking: A review and experimental evaluation. *IEEE Geoscience and Remote Sensing Magazine*, 10(1):125–160, 2022. doi: 10.1109/MGRS.2021.3072992.

Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey. Learning background-aware correlation filters for visual tracking. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1144–1152, 2017. doi: 10.1109/ICCV.2017.129.

João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 702–715, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33765-9.

João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2015. doi: 10.1109/TPAMI.2014.2345390.

Ziyuan Huang, Changhong Fu, Yiming Li, Fuling Lin, and Peng Lu. Learning aberrance repressed correlation filters for real-time uav tracking. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2891–2900, 2019. doi: 10.1109/ICCV.2019.00298.

Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018a. doi: 10.1109/CVPR.2018.00935.

Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, and Ming-Hsuan Yang. Learning spatial-temporal regularized correlation filters for visual tracking. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4904–4913, 2018b. doi: 10.1109/CVPR.2018.00515.

Siyi Li and Dit-Yan Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4140–4146. AAAI Press, 2017.

Yang Li and Jianke Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Computer Vision - ECCV 2014 Workshops*, pages 254–265, Cham, 2015. Springer International Publishing. ISBN 978-3-319-16181-5.

Yiming Li, Changhong Fu, Fangqiang Ding, Ziyuan Huang, and Geng Lu. Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11920–11929, 2020. doi: 10.1109/CVPR42600.2020.01194.

Alan Lukežic, Tomáš Vojír, Luka Cehovin Zajc, Jirí Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4847–4856, 2017. doi: 10.1109/CVPR.2017.515.

Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. Hierarchical convolutional features for visual tracking. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3074–3082, 2015. doi: 10.1109/ICCV.2015.352.

Matthias Mueller, Neil Smith, and Bernard Ghanem. Context-aware correlation filter tracking. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1387–1395, 2017. doi: 10.1109/CVPR.2017.152.

Horst Possegger, Thomas Mauthner, and Horst Bischof. In defense of color-based model-free tracking. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2113–2120, 2015. doi: 10.1109/CVPR.2015.7298823.

Lei Pu, Xinxi Feng, and Zhiqiang Hou. Spatial adaptive regularized correlation filter for robust visual tracking. *IEEE Access*, 8:11342–11351, 2020. doi: 10.1109/ACCESS.2020.2964716.

Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13901–13910, 2021. doi: 10.1109/CVPR46437.2021.01369.

Lijun Wang, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu. Stct: Sequentially training convolutional networks for visual tracking. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1373–1381, 2016. doi: 10.1109/CVPR.2016.153.

Mengmeng Wang, Yong Liu, and Zeyi Huang. Large margin object tracking with circulant feature maps. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4800–4808, 2017. doi: 10.1109/CVPR.2017.510.

Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38:13, 2006.

Yi Zhang, Guixi Liu, Haoyang Zhang, and Hanlin Huang. Robust visual tracker combining temporal consistent constraint and adaptive spatial regularization. *Neural Comput. Appl.*, 33(14):8355–8374, jul 2021. ISSN 0941-0643. doi: 10.1007/s00521-020-05589-w.