



Deep Convolutional Correlation Iterative Particle Filter for Visual Tracking

Reza Jalil Mozhdehi^{a,**}, Henry Medeiros^{b,2}

^aDepartment of Electrical and Computer Engineering, Marquette University, WI, USA

^bDepartment of Agricultural and Biological Engineering, University of Florida, FL, USA

ABSTRACT

This work proposes a novel framework for visual tracking based on the integration of an iterative particle filter, a deep convolutional neural network, and a correlation filter. The iterative particle filter enables the particles to correct themselves and converge to the correct target position. We employ a novel strategy to assess the likelihood of the particles after the iterations by applying K-means clustering. Our approach ensures a consistent support for the posterior distribution. Thus, we do not need to perform resampling at every video frame, improving the utilization of prior distribution information. Experimental results on two different benchmark datasets show that our tracker performs favorably against state-of-the-art methods.

© 2022 Elsevier Ltd. All rights reserved.

1. Introduction

Visual tracking is a challenging computer vision problem in which the size and location of a specific target are provided in the first video frame, and the target is then followed in subsequent frames by estimating its size and position. What makes the problem particularly challenging is the fact that the appearance of the target may change significantly in scenarios such as those involving partial occlusion or deformation. The introduction of deep convolutional neural networks (CNNs) (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015) to extract target features for visual tracking was a turning point in the design of tracking algorithms (Wang and Yeung, 2013; Hong et al., 2015). These convolutional features, in conjunction with correlation filters such as those proposed in (Dai et al., 2019b; Zhang et al., 2018a; Qi et al., 2016; Ma et al., 2015a; Mozhdehi and Medeiros, 2017; Fu et al., 2020; Raju et al., 2021), significantly improve tracking performance in comparison with traditional correlation filters based on hand-crafted features (Zhang et al., 2014; Zhong et al., 2014; Hare et al., 2011; Kalal et al., 2012). The methods described in (Mozhdehi and Medeiros, 2017) and (Zhang et al., 2017) were the first ap-

proaches to demonstrate the effectiveness of employing particle filters in conjunction with correlation-convolution trackers. In these methods, particle filters are used to sample several image patches, which are then processed by a CNN. The weight of each sample is calculated by applying a correlation filter to the convolutional features.

In this paper, we propose a deep convolutional correlation iterative particle filter (D2CIP) tracker. D2CIP is an extension of our previous visual trackers (Mozhdehi and Medeiros, 2017; Mozhdehi et al., 2018b,a; Mozhdehi and Medeiros, 2020), which represent a new class of tracking algorithms that integrate Sequential Monte Carlo strategies with correlation-convolution techniques. Our proposed tracker uses multiple particles as the inputs to a CNN (Simonyan and Zisserman, 2015) and then applies the correlation filter used in the ECO tracker (Danelljan et al., 2017a) to generate the correlation map of each particle. As Fig. 1 illustrates, trackers based on correlation filters attempt to determine the new position of the target by analyzing the displacement between the center of the correlation map, which corresponds to the previous target position, and the new peak in the map. More specifically, the correlation between the target model generated at previous frames and convolutional features extracted from the current frame are used to determine the new target position. Larger displacements between the previous target position and the peak of the correlation map lead to a degradation in the quality of the correlation map, since the corresponding convolutional features are less similar to the target model. Our proposed method differs

^{**}Corresponding author: Tel.: +001-414-458-1579;

e-mail: reza.jalilmozhdehi@marquette.edu (Reza Jalil Mozhdehi), hmedeiros@ufl.edu (Henry Medeiros)

²Part of this work was performed while the author was with the Department of Electrical and Computer Engineering, Marquette University.

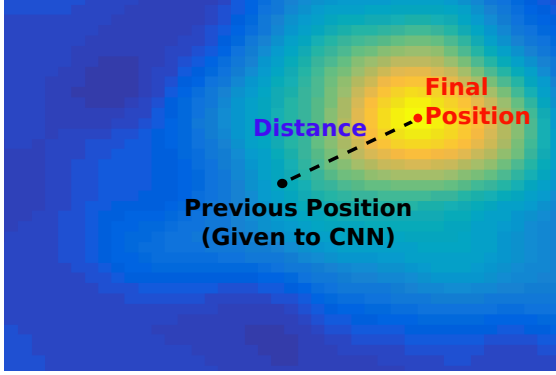


Fig. 1. Example of a correlation map of a given frame with respect to the previous target position.

from previous approaches in that it decreases this distance for each particle through an iterative procedure. At each iteration, the particles approach the target location and an improved correlation map is computed. To our knowledge, iterative particle filters have not been used in conjunction with CNNs and correlation filters before.

The second major contribution of our work is a novel target state estimation strategy. In our previous particle-correlation trackers (Mozhdehi and Medeiros, 2017; Zhang et al., 2017; Mozhdehi et al., 2018b,a; Mozhdehi and Medeiros, 2020) and similar state-of-the-art methods (Zhang et al., 2018b; Yuan et al., 2019), assessing the likelihood of the particles is challenging because many particles may be in close proximity to one another. In the framework proposed in this paper, the particles converge to a few locations after a series of iterations. Thus, we propose a novel method based on particle clustering and convergence consistency to evaluate the final particle locations. This novel method enables our proposed tracker to overcome challenges associated with multi-modal likelihood distributions. Additionally, existing particle-correlation trackers must perform resampling at every frame because shifting the particles to the peak of the correlation maps changes the support of the posterior distribution. Our iterative particle filter overcomes this issue and hence does not disregard information from prior samples. We tested our tracker on the LaSOT dataset (Fan et al., 2019), the TREK-150 dataset (Dunnhofer et al., 2021), and the Visual Tracker Benchmark v1.1 beta (OTB100) (Wu et al., 2015). The results show that our tracker outperforms several state-of-the-art methods.

2. Related work

The successful application of deep convolutional neural networks to object detection tasks (Liu et al., 2020; Lin et al., 2017; Girshick et al., 2016; Ren et al., 2015; Bashiri et al., 2018) has led to an increased interest in the utilization of such networks for visual tracking applications. Most CNN-based tracking algorithms use a CNN to examine image patches and determine the likelihood that a particular patch corresponds to the target. Li et al. (Li et al., 2016b) presented a tracker which samples image patches from the region surrounding the previous target position and uses multiple image cues as inputs to a CNN. The au-

thors later employed Bagging (Shinde et al., 2014) to improve the robustness of their online network weight update process (Li et al., 2016a). The multi-domain network (MDNet) tracker (Nam and Han, 2016) samples image patches at multiple positions and scales to account for target size variations. MDNet uses three convolutional layers to extract common target features and several domain- (or target-) specific fully connected layers to differentiate between a certain target category and the background. SANet (Fan and Ling, 2017) extends the MDNet architecture by employing a recurrent neural network (RNN) to predict the target position. However, much of the performance of MDNet and SANet is due to the fact that they are trained by utilizing the benchmark datasets on which they are evaluated.

One effective mechanism to determine the similarity between an image patch and the target are correlation filters (Choi et al., 2016; Tang and Feng, 2015; Danelljan et al., 2015). Trackers based on correlation filters measure the correlation between the target model and an image patch in the frequency domain and are agnostic to the features used to represent the targets. By employing correlation filters on the convolutional features generated by multiple layers of a deep CNN, HCFT (Ma et al., 2015a) shows substantial performance improvement in comparison with other visual trackers. Later, Qi et al. (Qi et al., 2016) introduced HDT, which augments HCFT with a hedging algorithm to weigh the outputs of the convolutional layers, better leveraging semantic information and improving tracking results (Li et al., 2018b). Instead of considering convolutional layers independently, Danelljan et al. proposed C-COT (Danelljan et al., 2016b), which employs a continuous fusion method among multiple convolutional layers and uses a joint learning framework to leverage different spatial resolutions. The authors later addressed C-COT’s problems of computational complexity and model over-fitting in the ECO algorithm (Danelljan et al., 2017a). Their factorized convolution operator and their novel model update method decrease the number of parameters in the model and improve tracking speed and robustness.

Recently, several methods have been proposed to improve the performance of correlation-based trackers. One strategy entails combining different types of features and constructing multiple correlation-based experts (Wang et al., 2018). Spatial-temporal information can also be used to address unwanted boundary effects in correlation trackers by spatially penalizing the filter coefficients (Danelljan et al., 2015). To address the additional computational costs associated with such strategies, in contrast to methods that train the model using samples from multiple frames, Li et al. (2018a) update the correlation model using samples from the current frame and the previously learned correlation filter. Zhu et al. (2018) further extend such strategies through a spatial-temporal attention mechanism that uses optical flow information in consecutive frames. Finally, Sun et al. (2018) use an approach based on reliability information (Du et al., 2020), which performs real-time tracking by estimating the importance of sub-regions within the correlation filter.

Particle filters provide an effective and general framework for improving the performance of correlation-based trackers. However, using particle filters in conjunction with correlation filters also involves some challenges. As shown in (Mozhdehi

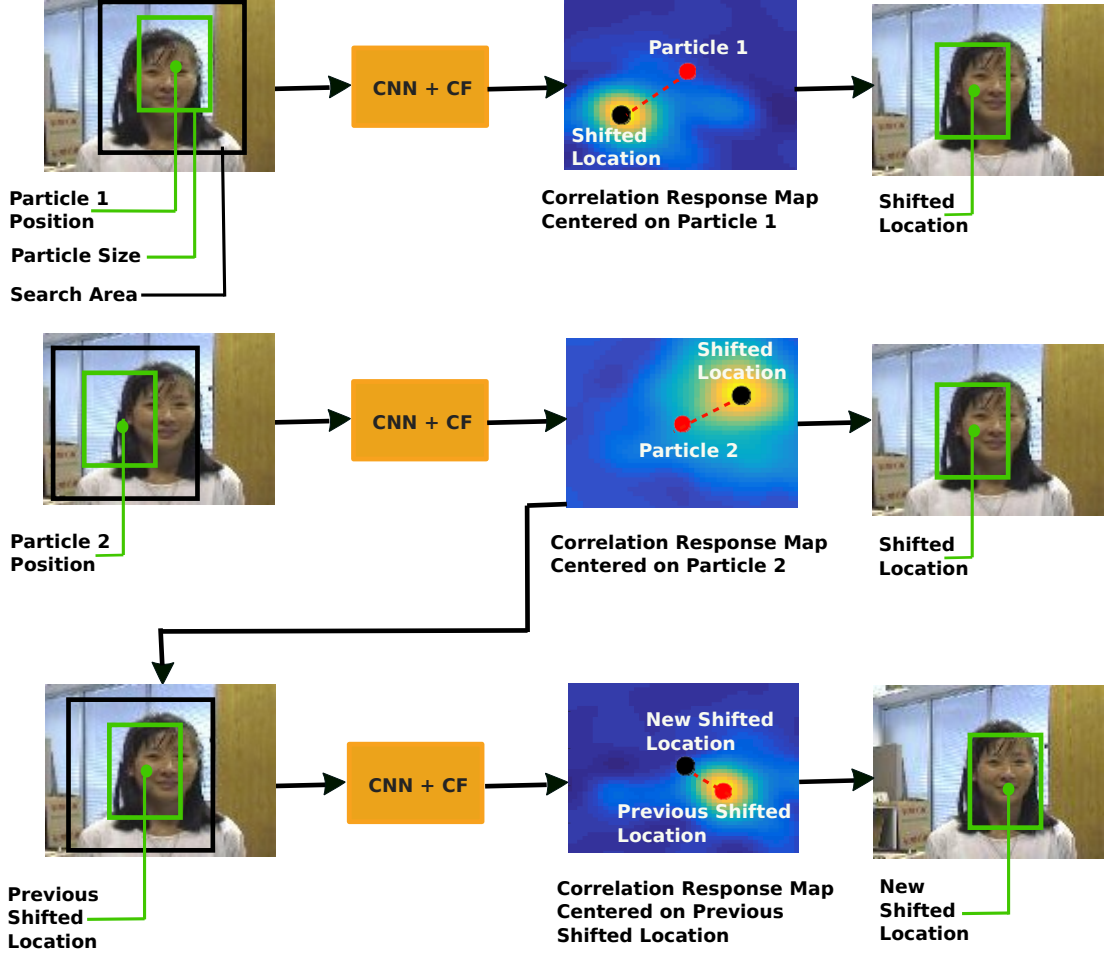


Fig. 2. The top and middle rows illustrate how two distinct particles with different weights converge to the same shifted location. Two different patches centered in particles 1 and 2 are given to the CNN and correlation filter to generate their correlation response maps. Each of these particles is then shifted to the peak of its corresponding correlation response map. As shown, their shifted locations are identical. Thus, this location is associated with two different weights in the posterior distribution because the correlation maps corresponding to the two particles are different. The middle and bottom rows show how the particle and its shifted location may generate different correlation maps. In the bottom row, a patch centered in the shifted location of particle 2 is generated. This patch results in a different correlation map at a different shifted location in comparison with the patch corresponding to particle 2. Thus, if the shifted location of particle 2 is used in the computation of the posterior distribution, its weight should be calculated based on the correlation map on the bottom row instead of the one on the middle row.

and Medeiros, 2017; Zhang et al., 2018b), particle-correlation trackers use the sum of the elements of the correlation maps as the weights of the particles. However, in challenging situations, such as in the presence of occlusions or target deformations, the correlation maps are not reliable and generate weights that do not reflect the similarity between the target model and the image patch under consideration. Additionally, particle-correlation trackers generally estimate the target state based on the particle with the maximum weight (Mozhdehi and Medeiros, 2017; Zhang et al., 2018b, 2017), which is not always an accurate method because many particles may have similar weights. Furthermore, the aforementioned correlation-particle trackers perform resampling at every frame and consequently disregard previous particle information. As shown in (Fan et al., 2015), iterative particle filters can improve sampling and lead to more distinctive particle likelihood models. However, such methods have been not used in conjunction with correlation-convolution trackers so far.

3. The change of support problem in convolution-correlation particle filters

Recursive Bayesian estimation algorithms attempt to determine the distribution of the target state x_t given a set of observations $y_{0:t} = \{y_0, y_1, \dots, y_t\}$ using Bayes rule

$$p(x_t|y_{0:t}) = \frac{p(y_t|x_t, y_{0:t-1})p(x_t|y_{0:t-1})}{p(y_t|y_{0:t-1})}. \quad (1)$$

Since analytical solutions to Eq. 1 are only available for very specific classes of problems (e.g., when all the distributions are normal), particle filters employ a Monte Carlo strategy to approximate the distribution of the target state. That is, a set of particles $\{x_t^{(i)}\}_{i=0}^N$ is sampled from a proposal distribution $q(x_t^{(i)}|x_{t-1}^{(i)}, y_t)$ and the contribution of each sample is weighed according to a likelihood function $p(y_t|x_t^{(i)})$ and a transition distribution $p(x_t^{(i)}|x_{t-1}^{(i)})$. In a particle filter, the weight of each par-

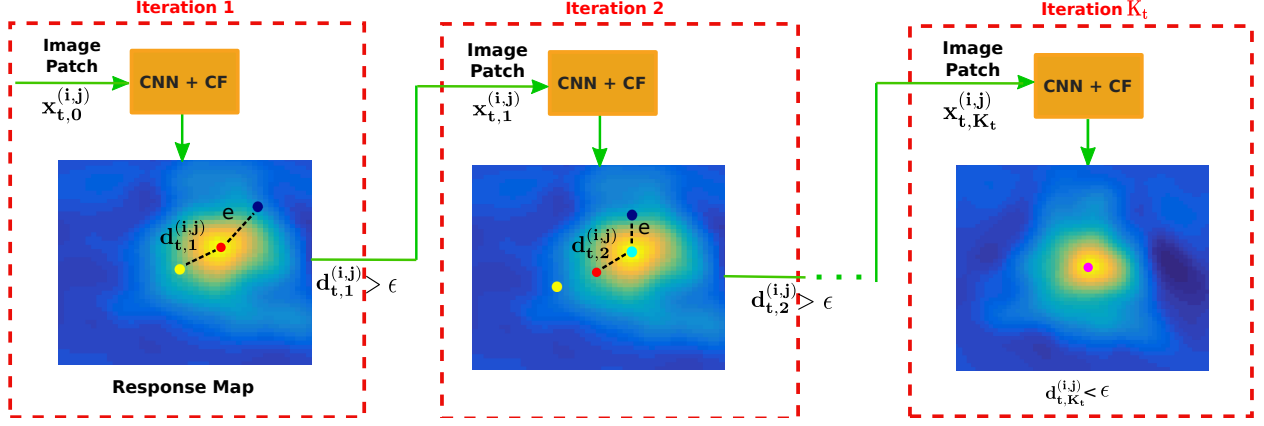


Fig. 3. Illustration of the proposed iterative particle position refinement. In the first iteration, particle $x_{t,0}^{(i,j)}$ shown by the yellow point is given to the CNN and the correlation filter to calculate its response map. Because the displacement $d_{t,1}^{(i,j)}$ between the estimated position and the particle is higher than ϵ , the particle needs to be refined. In this scenario, the estimated position is not accurate and there is an error e between the peak of the correlation map and the ground truth position (black point). The red point is then considered the new particle $x_{t,1}^{(i,j)}$ for the second iteration. The cyan point shows the estimated position in the second iteration, which needs further refinement despite the reduction in the error e . The purple point represents the position at the K_i -th iteration, which does not need to be refined because $d_{t,K_i}^{(i,j)} < \epsilon$.

ticle is calculated by

$$\omega_t^{(i)} \propto \omega_{t-1}^{(i)} \frac{p(y_t|x_t^{(i)})p(x_t^{(i)}|x_{t-1}^{(i)})}{q(x_t^{(i)}|x_{t-1}^{(i)}, y_t)}, \quad (2)$$

and posterior target state distribution is approximated by

$$p(x_t|y_t) \approx \sum_{i=1}^N \tilde{\omega}_t^{(i)} \delta(x_t - x_t^{(i)}), \quad (3)$$

where $\tilde{\omega}_t^{(i)}$ are the normalized particle weights. For additional details, we refer the reader to (Arulampalam et al., 2002).

Particle filters used in correlation trackers such as (Mozhdehi and Medeiros, 2017; Zhang et al., 2017; Mozhdehi et al., 2018b,a; Mozhdehi and Medeiros, 2020; Zhang et al., 2018b; Yuan et al., 2019) sample particles from the transition distribution. Thus, $q(x_t^{(i)}|x_{t-1}^{(i)}, y_t) = p(x_t^{(i)}|x_{t-1}^{(i)})$. Additionally, they resample particles at every frame, which removes the dependency on previous weights. Hence, the weight of each particle in these trackers is

$$\omega_t^{(i)} \propto p(y_t|x_t^{(i)}). \quad (4)$$

In these trackers, the particles are shifted to the peak of the correlation maps and the posterior distribution is then calculated by the particles' weights and the corresponding shifted locations

$$p(x_t|y_t) \approx \sum_{i=1}^N \tilde{\omega}_t^{(i)} \delta(x_t - \tilde{x}_t^{(i)}), \quad (5)$$

where $\tilde{x}_t^{(i)}$ is the peak of the correlation response map corresponding to the i -th particle.

However, the posterior distribution must take into consideration the weights corresponding to the shifted locations, not the original particles. As seen in the top and middle rows of Fig. 2, it is possible that multiple particles with different correlation maps and weights converge to the same location, which means the posterior distribution would then include multiple particles

at a common location but with different weights, which invalidates the assumption that the likelihood depends solely on $x_t^{(i)}$. In other words, the patch centered in the shifted location generates different features and consequently a different weight from the corresponding patch centered in the particle as shown in the middle and bottom rows of Fig. 2. Thus, the posterior distributions estimated by these trackers are not accurate.

Because the posterior distributions generated by these trackers are not reliable, they resort to resampling at every frame. While resampling is a suitable solution to avoid sample degeneracy that should be performed when necessary, resampling at every frame causes loss of information. It also causes sample impoverishment (i.e., loss of diversity among particles) and may cause all the particles to collapse to a single point within a few frames. In (Mozhdehi and Medeiros, 2020), we addressed the aforementioned problems by proposing a likelihood particle filter. Although the peaks of the correlation maps in (Mozhdehi and Medeiros, 2020) are used as the proposal and transition distributions, the weights of the peaks are still calculated based on the likelihood of the particles instead of the likelihood of the peaks. Our proposed iterative particle filter is a novel solution for correlation-convolutional trackers to calculate an accurate posterior without performing resampling at every frame.

4. Deep convolutional iterative particle filter

This section discusses our proposed strategy to generate particles that better reflect the actual position of the target while avoiding the change of support problem discussed in the previous section. Our approach is based on an iterative particle filter that gradually shifts the particle positions to locations that are closer to the peak of the correlation response map while also updating the response maps so that they become less sensitive to background clutter and better aligned with the target position. As the particles are updated, their corresponding weights are also recomputed based on the new correlation response maps.

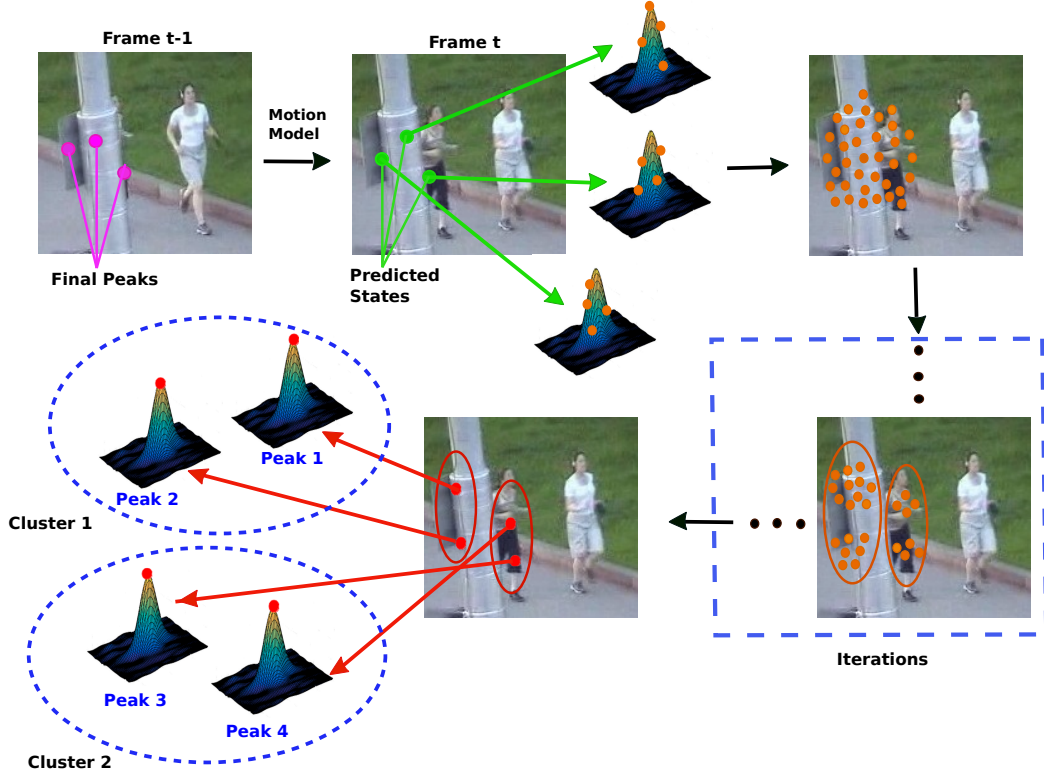


Fig. 4. Illustration of the particle selection process for $J_{t-1} = 3$. The particles are sampled from three distributions whose means are given by the previous correlation map peaks. At time t , the particles converge to four final peaks at the end of the iterations. Two clusters are found by applying K-means to the final particle locations. After selecting the best cluster, the peak of the correlation response map corresponding to the cluster with the highest number of particles is selected as the target state.

4.1. Iterative particle filter

As Fig. 3 illustrates, correlation filter-based trackers attempt to determine the position of the target by analyzing the displacement between the center of the correlation map, which corresponds to each particle in our tracker, and the peak of the map. A particle filter allows us to generate multiple samples around the predicted target state and hence increase our chances of finding maps with low displacement (Mozhdehi and Medeiros, 2017). If this displacement is sufficiently low, the features extracted from the CNN are similar to the model and the correlation response map is reliable. Our iterative particle filter considerably decreases this displacement and generates more reliable correlation maps for all the particles. As shown in Fig. 3, after generating the correlation response map for one particle, the distance between the particle position (yellow point) and the peak of the map (red point) is calculated. If the distance is larger than a small threshold ϵ , the correlation response map is not reliable enough to estimate the target position because it was generated based on an image patch centered at a position (yellow point) far from the ground truth location (black point). In such scenarios, the corresponding particle needs to be refined. To that end, the peak of the map is considered the new particle position and its corresponding correlation response map is calculated in a subsequent iteration. Although the peak of the new map (cyan point) is closer to the ground truth, the corresponding particle needs further refinement because the distance between the new particle position (which is now the red point)

and the peak of the new map is larger than ϵ . Finally, in the K_t -th iteration, the calculated distance is smaller than ϵ and the iterative refinement procedure terminates. The peak of the final map (purple point) is considered the estimated target position for this particle. Since no shifting is performed in the last iteration, the particle support problem discussed above is avoided. Our iterative particle filter is explained in greater detail in the following subsections.

4.1.1. Particle prediction model

As shown in the top row of Fig. 4, the posterior distribution of the target at time $t-1$ is modeled by a mixture of J_{t-1} normal distributions $\mathcal{N}(z_{t-1}^{(j)}, \sigma^2)$ whose means are given by

$$z_{t-1}^{(j)} = [x_{t-1}^{(j)}, \dot{x}_{t-1}^{(j)}]^T, \quad (6)$$

where $j = 1, \dots, J_{t-1}$, $x_{t-1}^{(j)} = [p_{t-1}^{(j)}, s_{t-1}^{(j)}]$ comprises the position $p_{t-1}^{(j)} \in \mathbb{R}^2$ and size $s_{t-1}^{(j)} \in \mathbb{R}^2$ of the target and $\dot{x}_{t-1}^{(j)} \in \mathbb{R}^4$ corresponds to its velocity. We use a constant velocity motion model to predict the means of the J_t distributions at the next time instant according to

$$\hat{z}_t^{(j)} = A z_{t-1}^{(j)}, \quad (7)$$

where A is the process matrix defined by

$$A = \begin{bmatrix} I_4 & I_4 \\ 0_{(4,4)} & I_4 \end{bmatrix}, \quad (8)$$

where I_4 is a 4×4 identity matrix and $0_{(4,4)}$ is a 4×4 zero matrix. Our transition distribution is then given by the mixture

$$p(x_{t,0}|x_{t-1}) = \frac{1}{J_{t-1}} \sum_{j=1}^{J_{t-1}} \mathcal{N}(\hat{z}_t^{(j)}, \sigma^2). \quad (9)$$

We then sample $N_t \times J_{t-1}$ new particles $x_{t,0}^{(i,j)} \sim p(x_{t,0}|x_{t-1})$, where $i = 1, \dots, N_t$. To increase the efficiency of our strategy, instead of sampling all the particles directly from the mixture distribution, we employ a stratified approach and sample N_t particles from each of the J_{t-1} predicted normal distributions. Fig. 4 also illustrates the processes of refining and clustering the particles, which are discussed in more detail in the following sections. Algorithm 1 summarizes our iterative particle filter algorithm. Lines 1-4 correspond to the sampling method described above. The remaining steps of the algorithm are also discussed in the following sections.

Algorithm 1 Deep Convolutional Correlation Iterative Particle (D2CIP).

Input: Current frame at time t , target models, previous final peaks $z_{t-1}^{(j)}$ and their normalized weights $\bar{\omega}_{t-1}^{(j)}$

Output: Estimated target state x_t^* , updated target models, current final peaks $z_{t,K_t}^{(j)}$ and their normalized weights $\bar{\omega}_{t,K_t}^{(j)}$

- 1: Find the predicted distributions $\mathcal{N}(\hat{z}_t^{(j)}, \sigma^2)$ according to Eqs. 6 to 8
 - 2: **for** each predicted state $\hat{z}_t^{(j)}$ **do**
 - 3: Sample N_t initial particles $x_{t,0}^{(i,j)} \sim \mathcal{N}(\hat{z}_t^{(j)}, \sigma^2)$
 - 4: **end for**
 - 5: **for** each initial particle $x_{t,0}^{(i,j)}$ **do**
 - 6: Find its corresponding final peak using Alg. 2
 - 7: **end for**
 - 8: **for** each final peak $x_{t,K_t}^{(i,j)}$ **do**
 - 9: Calculate the peak weight $\omega_{t,K_t}^{(i,j)}$
 - 10: **end for**
 - 11: Estimate the target state x_t^* based on the final peaks $x_{t,K_t}^{(i,j)}$ using Alg. 3
 - 12: Find the updated target models based on the final peaks $x_{t,K_t}^{(i,j)}$
 - 13: Resample if the effective sample size is lower than γ
-

4.1.2. Iterative particle refinement

Particle $x_{t,0}^{(i,j)}$ is used to sample a patch from the current frame at time t and to generate the corresponding convolutional features. These features are compared with the target models to calculate the correlation response map $R_{t,0}^{(i,j)}$ using the correlation filter proposed in (Danelljan et al., 2017a). We maintain one target model for each of the predicted distributions generated using Eq. 7, but to simplify the notation in this section, we refrain from explicitly differentiating the models. Let $p(y_t|x_{t,0}^{(i,j)})$ be the likelihood of $x_{t,0}^{(i,j)}$, which is given by the sum of the elements of $R_{t,0}^{(i,j)}$. We discard the samples for which $p(y_t|x_{t,0}^{(i,j)}) < L_{min}$, where L_{min} is the threshold to consider a correlation response map acceptable. As illustrated in Fig. 3, at

each iteration $k = 1, \dots, K_t$, the remaining samples are shifted to $x_{t,k}^{(i,j)}$, which is defined as

$$x_{t,k}^{(i,j)} = [p_{t,k}^{(i,j)}, s_{t,0}^{(i,j)}], \quad (10)$$

where $p_{t,k}^{(i,j)} = \text{argmax}(R_{t,k-1}^{(i,j)})$, i.e., the peak of the associated correlation response map at step $k-1$ of the iterative refinement process. We then have $d_{t,k}^{(i,j)} = \|p_{t,k}^{(i,j)} - p_{t,k-1}^{(i,j)}\|$ as the Euclidean distance between $p_{t,k}^{(i,j)}$ and $p_{t,k-1}^{(i,j)}$. For each particle, the refinement procedure continues until $d_{t,k}^{(i,j)} < \epsilon$. Since particles in close proximity tend to generate correlation response maps whose peaks share a common location, all the particles converge to a small number high-likelihood positions. These peaks determine the means of the normal distributions used to generate the prediction model described in Section 4.1.1.

As seen in Fig. 4, let $\mathcal{N}(x_{t,k}^{(i,j)}, \sigma^2)$ be the normal distributions after the convergence of the k -th iteration. We select the mean of these normal distributions as the particles for the next iteration if $d_{t,k}^{(i,j)} \geq \epsilon$. After the iterations, the particles reach the final peaks $x_{t,K_t}^{(i,j)} = [p_{t,K_t}^{(i,j)}, s_{t,0}^{(i,j)}]$, which do not need further refinement because $d_{t,K_t}^{(i,j)} < \epsilon$. Thus, we have

$$p_{t,K_t}^{(i,j)} = p_{t,0}^{(i,j)} + \sum_{k=1}^{K_t} d_{t,k}^{(i,j)}. \quad (11)$$

Algorithm 2 explains how to reach the final peaks in our iterative particle filter. Additionally, all the normal distributions $\mathcal{N}(x_{t,K_t}^{(i,j)}, \sigma^2)$ are used in the process of updating the target models as well. Our baseline tracker examines only the estimated target state to update the target models, while our iterative particle filter provides all $\mathcal{N}(x_{t,K_t}^{(i,j)}, \sigma^2)$ for the tracker to examine in the target model update process.

4.1.3. Weight update model

The posterior distribution for the particle filter is approximated by

$$p(x_t|y_t) \approx \sum \bar{\omega}_{t,K_t}^{(i,j)} \delta(x_t - x_{t,K_t}^{(i,j)}), \quad (12)$$

where $\bar{\omega}_{t,K_t}^{(i,j)}$ represents the normalized weights of the final correlation map peaks. Unlike previous convolution-correlation particle filters, since our approach keeps track of the particles that converge to a common location, it allows us to update the particle posterior distribution based on the likelihood of their final locations and their corresponding prior weights, i.e.,

$$\omega_{t,K_t}^{(i,j)} \propto p(y_t|x_{t,K_t}^{(i,j)}) \max_{\bar{\omega}_{t-1}^{(j)} \in \mathcal{X}_{t-1}^j} \bar{\omega}_{t-1}^{(j)}, \quad (13)$$

where $p(y_t|x_{t,K_t}^{(i,j)})$ is the likelihood of the final peak based on its correlation response map and $\mathcal{X}_{t-1}^j$ is the set of weights of $\mathcal{N}(z_{t-1}^{(j)}, \sigma^2)$ that converge to $x_{t,K_t}^{(i,j)}$. This approach allows us to refrain from unnecessarily resampling the particles at every frame. Instead, we perform resampling only when the effective sample size is lower than a threshold γ as in (Mozhdehi and Medeiros, 2017).

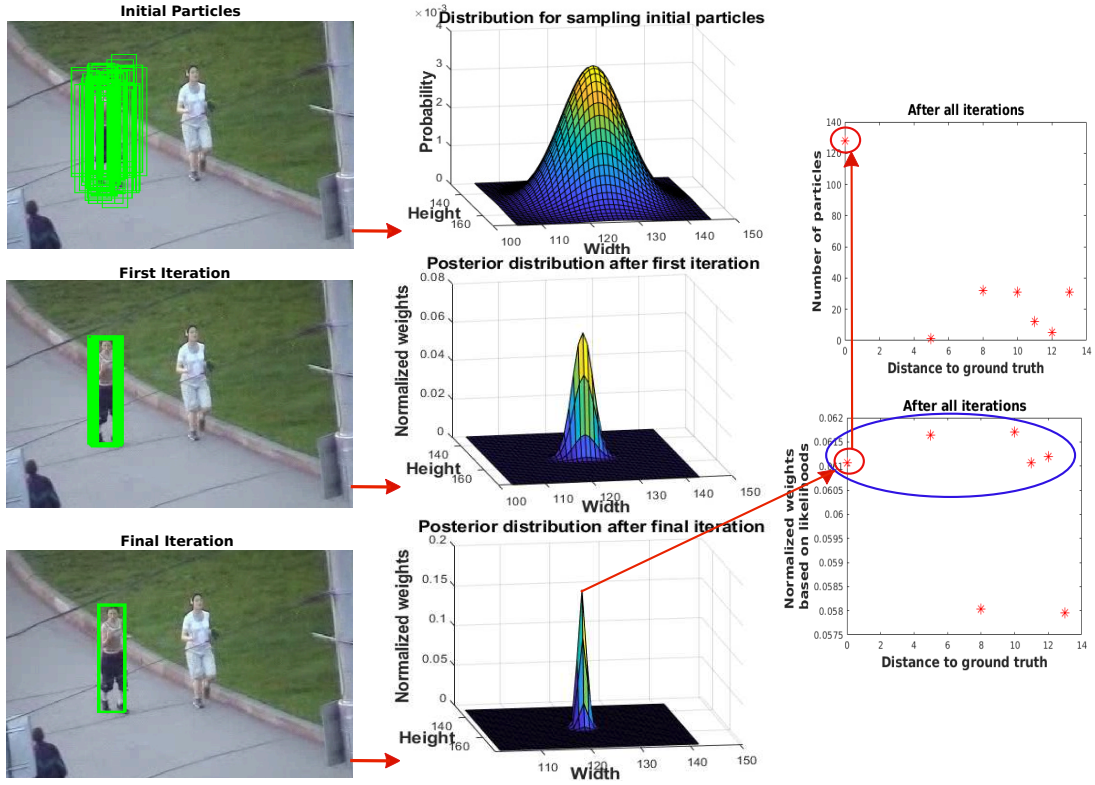


Fig. 5. The images in the left column show how our initial particles converge to the target after the iterations. The plots in the middle column illustrate how the particles reach a sharp final posterior distribution from the wide initial sampling distribution. In the right column, the plots show the normalized weights and the number of particles converging to the peaks after the iterations.

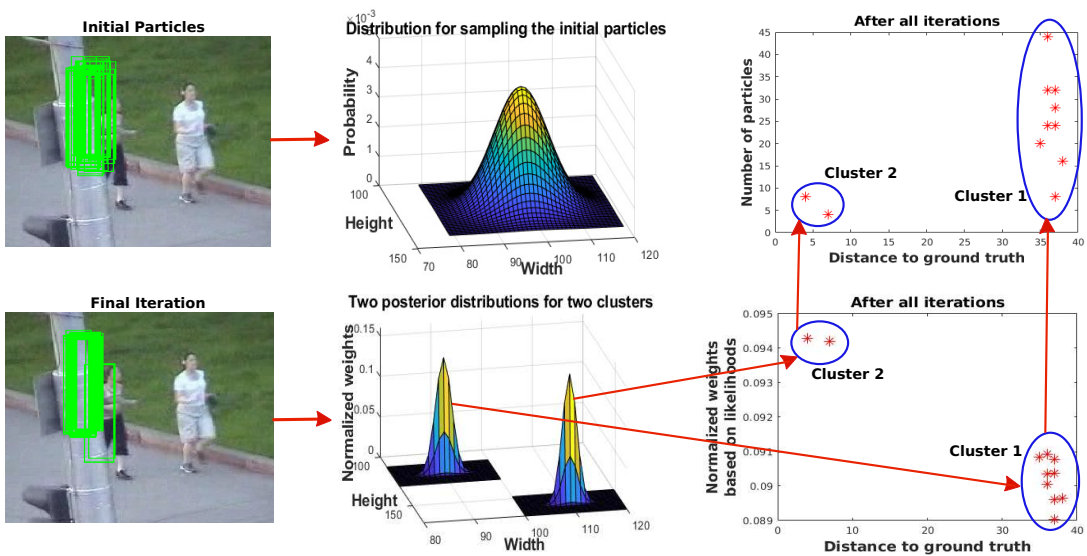


Fig. 6. The images in the left column show that the initial particles reach two clusters after the iterations in a challenging scenario. The plots in the middle column illustrate that the two clusters correspond to two sharp posterior distribution modes from the wide initial sampling distribution after applying K-means to the normalized particle weights. The right column shows that the weights are more reliable for distinguishing the clusters than the number of particles converging to their modes.

Algorithm 2 Iterative Particle Refinement.**Input:** Current frame t , initial particles $x_{t,0}^{(i,j)}$, target models**Output:** Final current peaks $x_{t,K_t}^{(i,j)}$

```

1: for each particle  $x_{t,0}^{(i,j)}$  do
2:    $d_{t,k}^{(i,j)} = \infty$ 
3:   while  $d_{t,k}^{(i,j)} > \epsilon$  do
4:     Generate the correlation response map  $R_{t,k-1}^{(i,j)}$ 
5:     Calculate the likelihood  $p(y_t|x_{t,k-1}^{(i,j)})$  based on  $R_{t,k-1}^{(i,j)}$ 
6:     if  $p(y_t|x_{t,k}^{(i,j)}) > L_{min}$  then
7:        $p_{t,k}^{(i,j)} = \text{argmax}(R_{t,k-1}^{(i,j)})$ 
8:        $d_{t,k}^{(i,j)} = \|p_{t,k}^{(i,j)} - p_{t,k-1}^{(i,j)}\|$ 
9:     else
10:      Discard particle  $x_{t,k-1}^{(i,j)}$ 
11:    end if
12:  end while
13:  Find  $J_{t,K}$  for each final peak
14: end for

```

4.2. Target state estimation

Using Eq. 13 in simple frames that do not involve any challenging scenario results in particle weights very similar to one another. Again, this is because the particles converge to a few nearby peaks after the iterations. Hence, the correlation maps related to these final peaks are similar as well. Thus, evaluation of the particles based on likelihoods calculated from the correlation maps is not sufficiently accurate in such simple frames. However, after the iterations, it is possible to determine the location to where most particles converge. As Fig. 5 illustrates, our initial particles gradually converge to a few peaks at the end of the iterative refinement procedure. The plots in the middle column of the figure show how the iterations decrease the area covered by the particles. As the plots indicate, after the iterations, the particles reach a sharp posterior distribution from a wide initial distribution. The plots in the left column of the figure show that the weights based on the correlation maps are similar to one another after the iterations. As seen in the bottom right plot, which shows the weights based on Eq. 13, the weight of the peak located exactly on the ground truth is slightly lower than the weights of farther peaks (shown within the blue ellipse). However, most particles converge to the peak closest to the ground truth location as shown in the plot at the top of the right column. Thus, the final state x_t^* is calculated by

$$x_t^* = \text{argmax}_{x_{t,K_t}^{(i,j)}} J_{t,K_t}, \quad (14)$$

where J_{t,K_t} is the number of particles $x_{t,0}^{(i,j)}$ that converge to the common final peak $x_{t,K_t}^{(i,j)}$.

However, when the tracker faces a challenging scenario, the area covered by the particles does not necessarily decrease after the iterations. This is because the particles may converge to different image regions. In the challenging scenario illustrated in Fig. 6, the particles converge to the pole and the jogger, which correspond to two clusters of particles. In such scenarios, we first determine the number of clusters and select the one

that best represents the posterior. Since the particles converge to distinct image regions, our proposed method can partition them using K-means clustering (Bishop, 1995). We determine the number of modes in the distribution using simplified silhouette analysis based on the Euclidean distances among particles (Wang et al., 2017a). The plots in the middle column of Fig. 6 illustrate how the particles form a posterior distribution with two sharp modes from the wide initial sampling distribution. This posterior distribution is calculated based on the particle weights according to Eq. 13. As seen in the top plot of the right column of Fig. 6, the number of particles converging to each cluster is not sufficiently accurate to find the image region corresponding to the target. As the plot indicates, only a few particles converge to the region surrounding the jogger. The bottom plot of the right column of the figure shows that the weights calculated by Eq. 13, on the other hand, provide an accurate method to distinguish the clusters. Because of the considerable distance between the clusters, the correlation response maps within different clusters are not similar to each other. Thus, particle evaluation based on the likelihoods according to Eq. 13 is reliable because correlation maps closer to the target generate higher likelihoods. Thus, we first find the clusters using K-means after performing the iterations, and the mode of each cluster is then selected based on the number of particles reaching the final peaks. The best cluster is then selected based on the correlation response maps corresponding to each mode according to Eq. 13. Algorithm 3 summarizes our mode clustering and state estimation method.

5. Results and Discussion

We evaluate our algorithm on three publicly available visual tracking benchmarks: the large-scale single object tracking benchmark (LaSOT) (Fan et al., 2019), the recently published TREK-150 dataset (Dunnhofer et al., 2021), and the visual tracker benchmark v1.1 beta (OTB100) (Wu et al., 2015). All the results shown in this section correspond to a particle filter with 200 particles.

5.1. LaSOT evaluation

LaSOT is currently the largest publicly available benchmark for object tracking. It provides high-quality dense manual annotations with 14 attributes representing challenging aspects of tracking. The benchmark consists of 1,400 videos with an average of 2,512 frames per sequence. The benchmark provides

Algorithm 3 Target State Estimation.**Input:** Final peaks $x_{t,K_t}^{(i,j)}$ at time t , their weights $\omega_{t,K_t}^{(i,j)}$ and J_{t,K_t} **Output:** Current target state x_t^*

-
- 1: Apply K-means to all final current peaks $x_{t,K_t}^{(i,j)}$ to find the clusters
 - 2: Find the mode of each cluster based on J_{t,K_t}
 - 3: Compare the weight of the calculated modes based on $\omega_{t,K_t}^{(i,j)}$ to select the best mode
 - 4: Consider the best mode as the current target state x_t^*
-

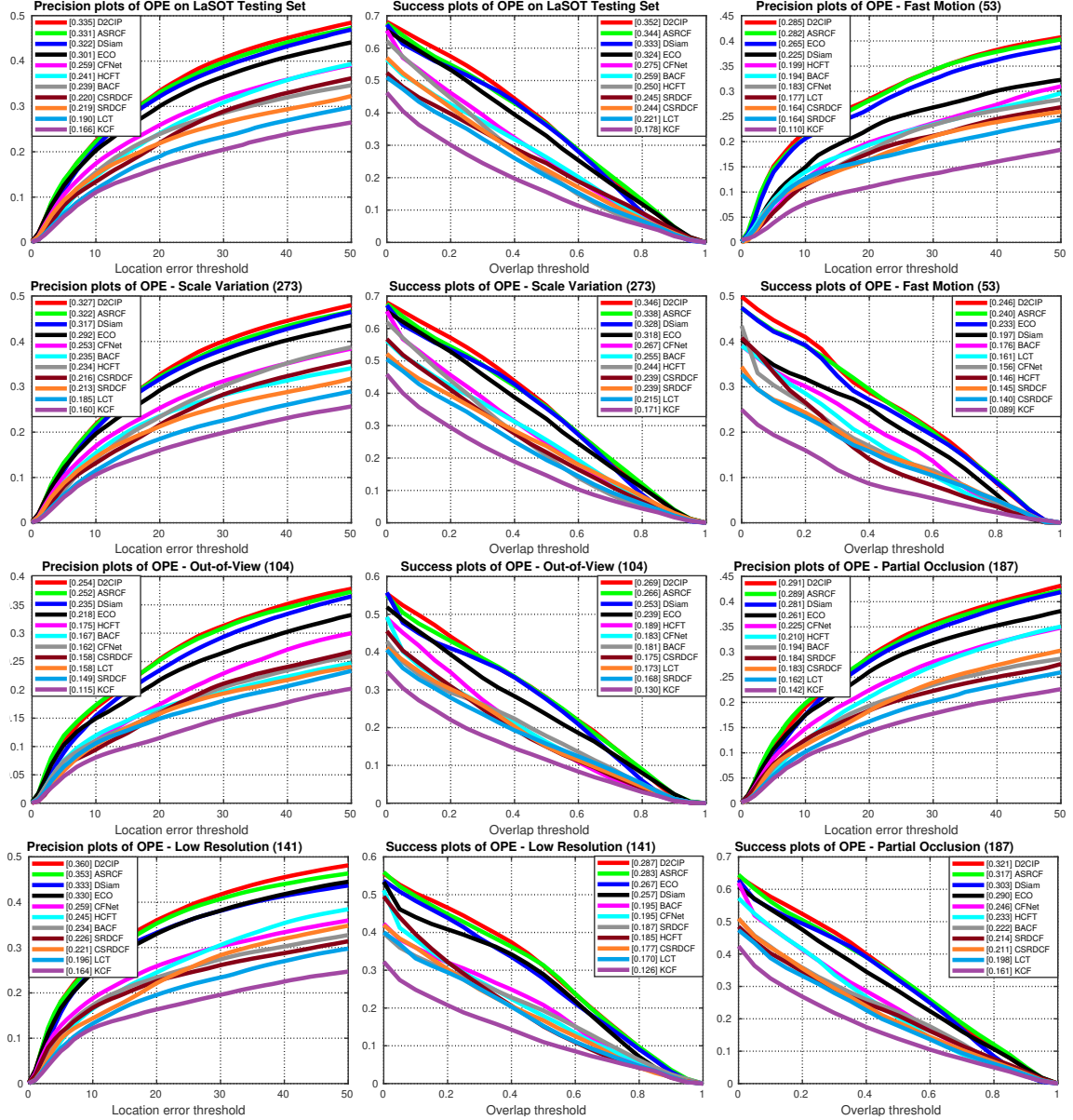


Fig. 7. Quantitative assessment of the performance of our tracker in comparison with state-of-the-art trackers using a one-pass evaluation on the LaSOT benchmark dataset.

two different metrics to evaluate visual trackers: precision and success. For more details on the evaluation metrics, we refer the reader to (Fan et al., 2019). Fig. 7 presents a quantitative assessment of our proposed approach using a one-pass evaluation (OPE) on LaSOT in comparison with 10 state-of-the-art trackers including ECO, ASRCF (Dai et al., 2019a), DSiam (Guo et al., 2017), CFNet (Valmadre et al., 2017), HCFT (Ma et al., 2015a), BACF (Galoogahi et al., 2017), CSRDCF (Lukezic et al., 2017), SRDCF (Danelljan et al., 2015), LCT (Ma et al., 2015b) and KCF (Henriques et al., 2015). In the one-pass evaluation, the tracker is initialized with the ground truth location of the target at the first frame of the image sequence and allowed to keep track of the target over the remaining frames without reinitialization. As seen in Fig. 7, our tracker outperforms all the other trackers in terms of overall precision and success. In particular, it outperforms ASRCF by 1.2% and 2.3%, respectively.

Similar to our proposed tracker, ASRCF is a recent state-of-the-art correlation-convolutional visual tracker that uses ECO as a baseline method. Our most significant improvements in comparison with ASRCF occur in low resolution and scale variation scenarios, which show improvements of 2% and 1.6% in precision and 1.4% and 2.4% in success, respectively. In comparison with our baseline tracker, our precision improvement reaches 9.1%, 12.0%, and 11.5% in low resolution, scale variation, and partial occlusion scenarios, respectively. In terms of the success metric, our improvement in such scenarios reaches 7.5%, 8.8%, and 10.7% in comparison with ECO.

5.2. TREK-150 evaluation

TREK-150 (Dunnhofer et al., 2021) is a recent visual tracking benchmark dataset that includes 150 densely annotated

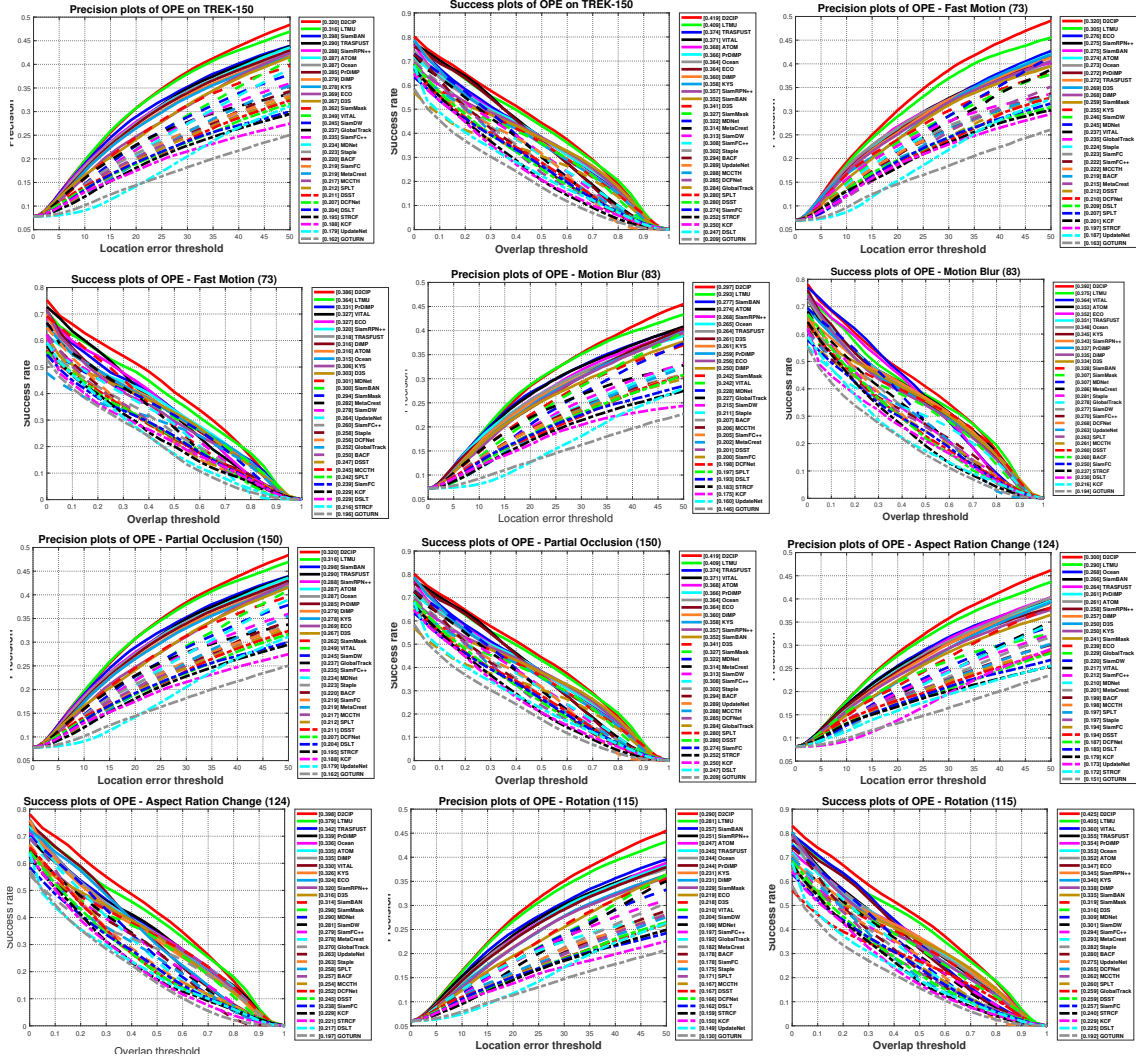


Fig. 8. Quantitative assessment of the performance of our tracker in comparison with state-of-the-art trackers using a one-pass evaluation on the TREK-150 benchmark dataset.

video sequences obtained from a first person perspective. Because no existing visual tracker has used this benchmark in its training process, it is a valuable resource to perform an accurate comparative evaluation of different tracking algorithms. Fig. 8 presents the OPE results of our algorithm on the TREK-150 dataset in comparison with 31 state-of-the-art trackers including methods based on deep Siamese networks such as SiamFC++ (Xu et al., 2020), SiamBAN (Chen et al., 2020), Ocean (Zhang et al., 2020), SiamMask (Wang et al., 2019), SiamRPN++ (Li et al., 2019), SiamDW (Zhang and Peng, 2019), UpdateNet (Zhang et al., 2019), DSLT (Lu et al., 2018), SiamFC (Bertinetto et al., 2016b), and GOTURN (Held et al., 2016), as well as correlation trackers (PrDiMP (Danelljan et al., 2020), KYS (Bhat et al., 2020), ECO, ATOM (Danelljan et al., 2019), DiMP (Bhat et al., 2019)), DSST (Danelljan et al., 2017b), KCF (Henriques et al., 2015), Staple (Bertinetto et al., 2016a), BACF, DCFNet (Wang et al., 2017b), STRCF (Li et al., 2018a), MCCTH (Ning Wang et al., 2018), MOSSE (Bolme et al., 2010)), tracking-by-detection methods (MDNet, VITAL (Song et al., 2018)), and trackers based on target seg-

mentation representations (D3S (WLukezic et al., 2020)), meta-learning (MetaCrest (Park and Berg, 2018)), fusion strategies (TRASFUST (Dunnhofer et al., 2020)), or long-term tracking mechanisms (SPLT (Yan et al., 2019), GlobalTrack (Huang et al., 2020), and LTMU (Dai et al., 2020)). As seen in Fig. 8, our tracker outperforms all the other trackers. In particular, it outperforms the state-of-the-art LTMU by 1.27% and 2.4% in terms of overall precision and success, respectively. It is worth noting that our method outperforms SiamBAN, the third best tracker in terms of success by a significant margin (7.38% relative improvement). As seen in Fig. 8, our improvement in comparison to LTMU is higher than 6.5% in attribute "Fast Motion". Furthermore, the long-term meta-update strategy proposed in LTMU could be integrated with our method, which would likely lead to further performance improvements.

5.3. OTB100 evaluation

The OTB100 benchmark contains 100 data sequences that are annotated with 11 attributes. In Fig. 9, we provide a quantitative OPE assessment of our proposed approach in comparison

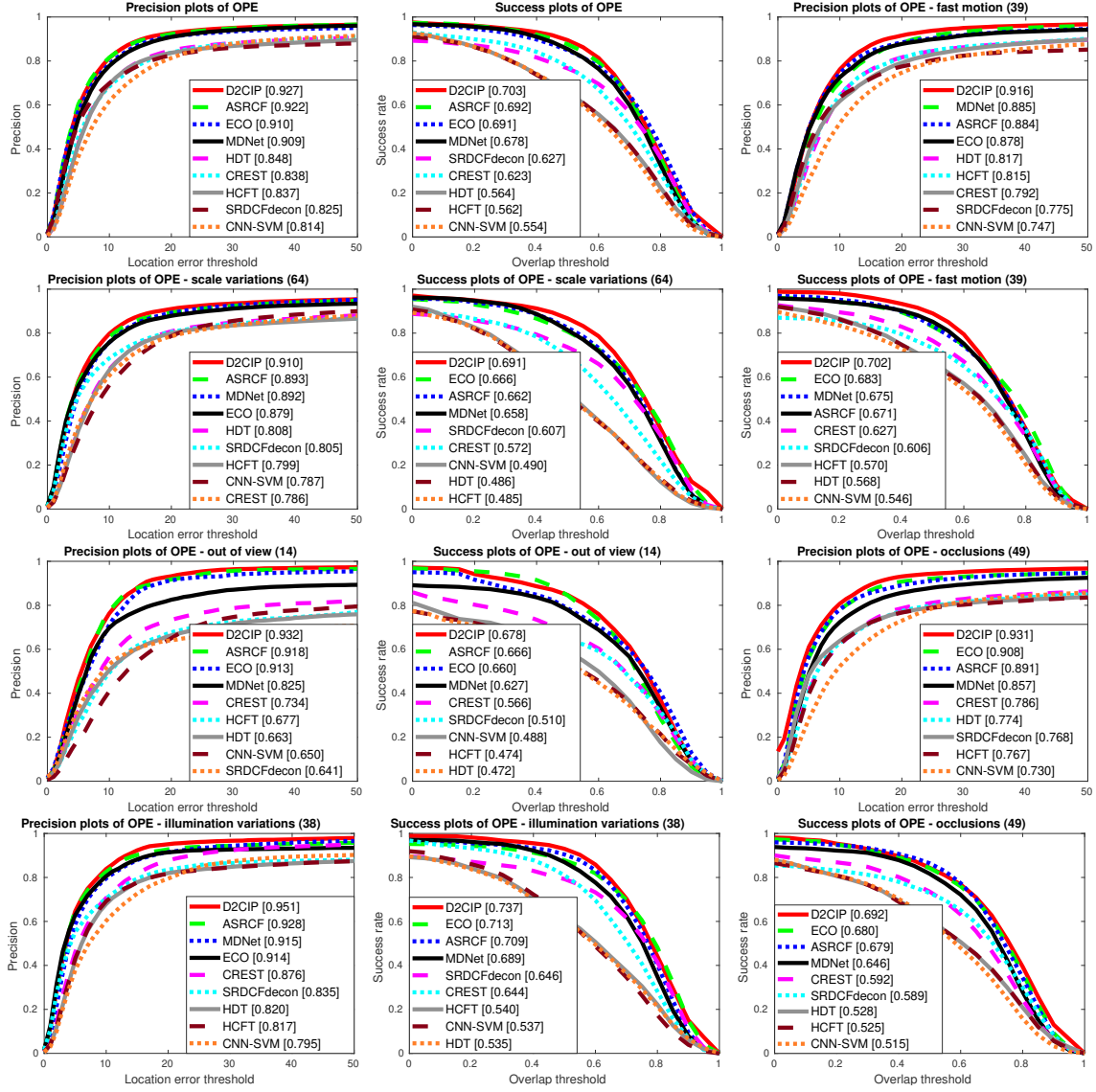


Fig. 9. Quantitative performance assessment of our tracker in comparison with eight state-of-the-art trackers using a one-pass evaluation on the OTB100 benchmark dataset.

with eight state-of-the-art trackers whose results in the OTB100 dataset are publicly available: ASRCF, ECO, MDNet (Nam and Han, 2016), HDT (Qi et al., 2016), HCFT, FRDCFdecon (Danelljan et al., 2016a), CREST (Song et al., 2017) and CNN-SVM (Hong et al., 2015). On the overall evaluation of the precision and success metrics considering the entire dataset, our tracker shows improvements of approximately 0.6% and 1.6% in comparison with the second best tracker ASRCF. Our precision and success improvements in comparison with ASRCF reach 3.6% and 4.6% in fast motion scenarios, 1.9% and 4.4% in scale variation scenarios, 1.5% and 1.8% in out of view conditions, 2.5% and 3.9% when illumination variations are present and 4.4% and 1.9% in scenes including occlusion. In some challenging scenarios in the OTB100 dataset, ASRCF is outperformed by ECO and MDNet. ECO is the second best tracker in the success metric for fast motion, scale variation and illumination variation scenarios as well as both metrics for occlusion scenarios. Our performance improvements with respect to ECO

in these scenarios are 2.8%, 3.8%, 3.4%, 2.5%, and 1.8%, respectively. MDNet only outperforms ASRCF in the precision metric for fast motion scenarios. In that case, our performance improvement with respect to MDNet is 3.5%.

Fig. 10 presents a qualitative assessment of our tracker in comparison with ASRCF, ECO, and HCFT. In the first row, the other trackers fail because of a relatively long occlusion period, which causes not only tracking loss but also incorrect model updates. Our tracker, on the other hand, finds multiple potential clusters during partial occlusion and maintains one model per cluster. When the target becomes visible again, this enables our tracker to sample from distributions closer to the target and assess the corresponding updated models. In the second row, the other trackers fail because of a period of fast motion by the target. Our particle filter enables our tracker to sample particles over a wider search area and the iterative particle refinement allows it to reach the best location for the target. In the third row, the presence of several similar objects causes failures in

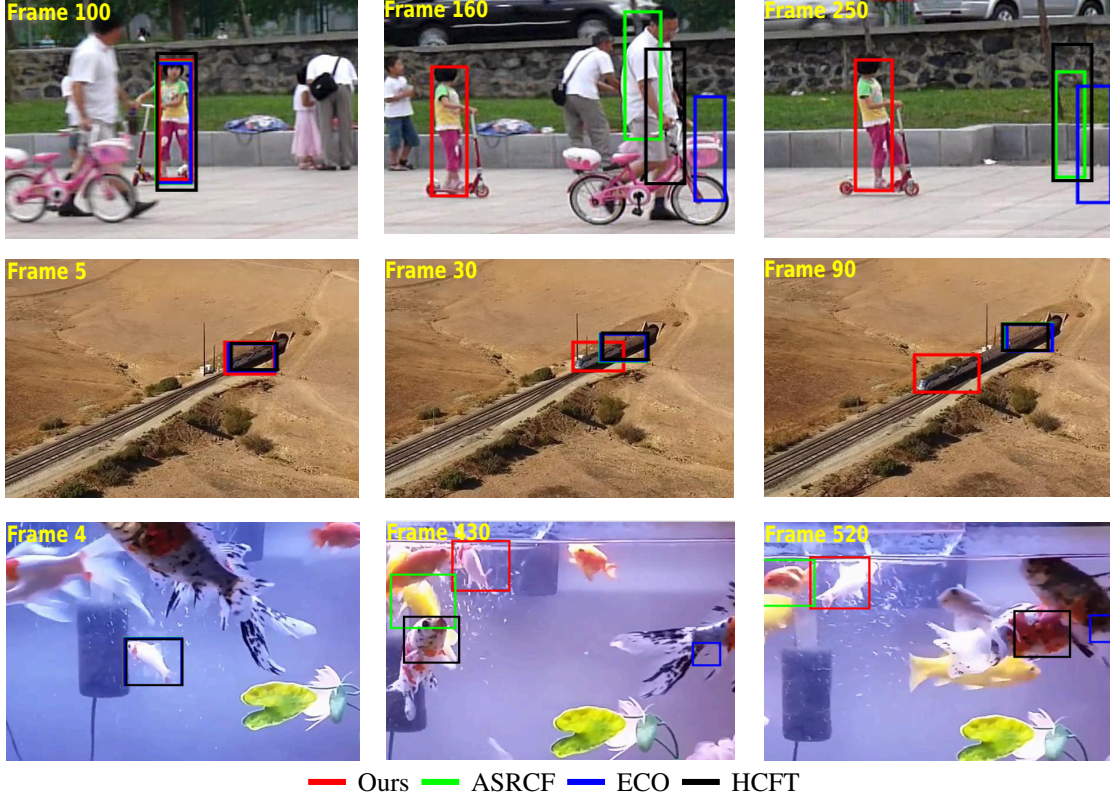


Fig. 10. Qualitative evaluation of our tracker in comparison with ASRCF, ECO and HCFT on three challenging sequences of the OTB100 (top row, *Girl2* sequence) and LaSOT datasets (middle and bottom rows, *Train-1* and *Goldfish-4* sequences, respectively).

Table 1. Ablative analysis of the components of the proposed visual tracker. PF corresponds to the integration of the baseline tracker with the particle filter model. IPF incorporates the iterative particle refinement procedure. IPFK includes the clustering of the likelihood distributions. D2CIP corresponds to our complete algorithm, which further incorporates maintaining one target model for each predicted mode. The numbers within parentheses indicate the relative performance gains over the baseline tracker.

Benchmark	PF	IPF	IPFK	D2CIP	Precision	Success
LaSOT	✓	✗	✗	✗	30.9% (+2.6%)	33.1% (+2.3%)
	✓	✓	✗	✗	32.3% (+7.3%)	34.5% (+6.4%)
	✓	✓	✓	✗	33.2% (+10.4%)	34.9% (+7.9%)
	✓	✓	✓	✓	33.5% (+11.3%)	35.2% (+8.7%)
TREK-150	✓	✓	✓	✓	32.0% (+18.96%)	41.9% (+15.11%)
OTB100	✓	✗	✗	✗	91.5% (+0.5%)	69.6% (+0.7%)
	✓	✓	✗	✗	92.2% (+1.3%)	69.9% (+1.2%)
	✓	✓	✓	✗	92.6% (+1.8%)	70.2% (+1.6%)
	✓	✓	✓	✓	92.7% (+1.9%)	70.3% (+1.7%)

the other trackers, whereas our novel method for evaluating the particle likelihoods locates the correct target accurately.

5.4. Ablative analysis and computation time

Table 1 shows the performance improvement introduced by the different components of our proposed method and compares them with ECO (Danelljan et al., 2017a), our baseline tracker. As the table indicates, the incorporation of a simple constant velocity particle filter with the baseline tracker leads to relative improvements of 2.6% and 2.3% in terms of precision and success on the LaSOT dataset. The iterative particle refinement procedure further improve these results by 4.7% and 4.1%. Clustering the particles to identify the modes of the

likelihood provide additional gains of 3.1% and 1.5%. Finally, maintaining one target model for each mode of the distribution leads to relative improvements for 0.9% and 0.8%. That is, our proposed method improves the overall performance of ECO by 11.3% in terms of precision and 8.7% on the success metric for LaSOT.

Since the OTB100 dataset contains less complex sequences, the performance of the baseline tracker is already relatively high for both metrics. For example, one of the main benefits of the iterative particle refinement procedure is that it generates more precise target estimates, reducing model drift. While this is easily observed in the longer videos comprising the LaSOT dataset, there are significantly fewer opportunities to cor-

rect model drift in the shorter videos contained in the OTB100 dataset. In addition, since the OTB100 dataset includes only 16 target categories, the convolutional features generated by the backbone CNN are much more discriminative than for the 85 categories present in LaSOT. As a consequence, the robustness to background clutter introduced by the multi-modal likelihood model are dramatically less pronounced in OTB100. Nonetheless, our tracker still provides up to 1.9% and 1.7% relative improvements in precision and success.

Our unoptimized MATLAB implementation of D2CIP runs at approximately 2 FPS on an RTX 2080Ti GPU. The main bottleneck in our current implementation is the computation of the convolutional features used in our likelihood model, which accounts for more than 80% of the processing time. However, since there are no inter-dependencies among particles, it is possible to compute their likelihoods in parallel. We have observed that the hardware resources are severely underutilized, which indicates room for reduction in the overall computation time.

6. Conclusion

This work proposes an iterative particle filter that works along with a deep convolutional neural network and a correlation filter to accurately track objects of interest in video sequences. In the proposed algorithm, after generating a set of initial particles around the predicted target sizes and positions and extracting their hierarchical convolutional features, a correlation map is calculated for each particle. These maps are used to refine the positions of the particles and generate the corresponding likelihoods. After discarding low likelihood particles, the displacement between each particle and the peak of its correlation response map is calculated. If the distance between the peak of the correlation response map and the particle location is larger than a threshold ϵ , the image patch centered at the peak of the response map is used to generate a new response map. This iterative procedure leads most particles to converge to only a few final positions. By iteratively updating the particle likelihoods, our method also addresses the problem of calculating the posterior distribution over the correct support points in particle-correlation trackers. Finally, a novel method is used to assess multi-modal likelihoods based on clustering the particles. The LaSOT, TREK-150, and OTB100 datasets are used for evaluating the proposed tracker's performance. The results show that our tracker substantially outperforms several state-of-the-art methods.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 2224591.

References

- Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing* 50.
- Bashiri, F.S., LaRose, E., Badger, J.C., D'Souza, R.M., Yu, Z., Peissig, P., 2018. Object detection to assist visually impaired people: A deep neural network adventure, in: 13th International Symposium on Visual Computing (ISVC).
- Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H., 2016a. Staple: Complementary learners for real-time tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H., 2016b. Fully-convolutional siamese networks for object tracking, in: *European Conference on Computer Vision (ECCV)*.
- Bhat, G., Danelljan, M., Gool, L.V., Timofte, R., 2019. Learning discriminative model prediction for tracking, in: *International Conference on Computer Vision*.
- Bhat, G., Danelljan, M., Gool, L.V., Timofte, R., 2020. Know your surroundings: Exploiting scene information for object tracking, in: *European Conference on Computer Vision (ECCV)*.
- Bishop, C., 1995. *Neural networks for pattern recognition*. Oxford university press.
- Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M., 2010. Feature pyramid networks for object detection, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Chen, Z., Zhong, B., Li, G., Zhang, S., Ji, R., 2020. Siamese box adaptive network for visual tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Choi, J., Chang, H.J., Jeong, J., Demiris, Y., Choi, J.Y., 2016. Visual tracking using attention-modulated disintegration and integration, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Dai, K., Wang, D., Huchuan, L., Sun, C., Li, J., 2019a. Visual tracking via adaptive spatially-regularized correlation filters, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Dai, K., Wang, D., Lu, H., Sun, C., Li, J., 2019b. Visual tracking via adaptive spatially-regularized correlation filters, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4670–4679.
- Dai, K., Zhang, Y., Wang, D., Li, J., Lu, H., Yang, X., 2020. High-performance longterm tracking with meta-updater, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M., 2019. Atom: Accurate tracking by overlap maximization, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M., 2017a. Eco: Efficient convolution operators for tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Danelljan, M., Gool, L.V., Timofte, R., 2020. Probabilistic regression for visual tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Danelljan, M., Häger, G., Khan, F.S., Felsberg, M., 2016a. Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Danelljan, M., Hager, G., Khan, F.S., Felsberg, M., 2017b. Discriminative scale space tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.
- Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M., 2015. Learning spatially regularized correlation filters for visual tracking, in: *IEEE International Conference on Computer Vision (ICCV)*.
- Danelljan, M., Robinson, A., Shahbaz Khan, F., Felsberg, M., 2016b. Beyond correlation filters: Learning continuous convolution operators for visual tracking, in: *European Conference on Computer Vision (ECCV)*.
- Du, F., Liu, P., Zhao, W., Tang, X., 2020. Joint channel reliability and correlation filters learning for visual tracking. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 1625–1638.
- Dunnhofer, M., Furnari, A., Farinella, G.M., Micheloni, C., 2021. Is first person vision challenging for object tracking?, in: *International Conference on Computer Vision (ICCV) Workshops*.
- Dunnhofer, M., Martinel, N., Micheloni, C., 2020. Tracking-by-trackers with a distilled and reinforced model, in: *Asian Conference on Computer Vision (ACCV)*.
- Fan, H., Lin, L., Fan, Y., Chu, P., Deng, G., Yu, S., Bai, H., 2019. LaSOT: A high-quality benchmark for large-scale single object tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Fan, H., Ling, H., 2017. SANet: Structure-aware network for visual tracking, in: *CVPR Workshop on DeepVision: Temporal Deep Learning*.
- Fan, Z., Ji, H., Zhang, Y., 2015. Iterative particle filter for visual tracking. *Image Communication* 36, 140–153.
- Fu, H., Zhang, Y., Wuneng, Z., Xiaofeng, W., Zhang, H., 2020. Learning reliable-spatial and spatial-variation regularization correlation filters for vi-

- sual tracking. *Image and Vision Computing* 94, 103869.
- Galoogahi, H.K., Fagg, A., Lucey, S., 2017. Learning background-aware correlation filters for visual tracking, in: *International Conference on Computer Vision (ICCV)*.
- Girshick, R., Donahue, J., Darrell, T., Malik, J., 2016. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 142–158.
- Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L., Wang, S., 2017. Learning dynamic siamese network for visual object tracking, in: *International Conference on Computer Vision (ICCV)*.
- Hare, S., Saffari, A., Torr, P.H.S., 2011. Struck: Structured output tracking with kernels, in: *IEEE International Conference on Computer Vision (ICCV)*, pp. 263–270.
- Held, D., Thrun, S., Savarese, S., 2016. Learning to track at 100 fps with deep regression networks, in: *European Conference on Computer Vision (ECCV)*.
- Henriques, J.F., Caseiro, R., Martins, P., Batista, J., 2015. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 583–596.
- Hong, S., You, T., Kwak, S., Han, B., 2015. Online tracking by learning discriminative saliency map with convolutional neural network, in: *International Conference on Machine Learning (ICML)*.
- Huang, L., Zhao, X., Huang, K., 2020. Globaltrack: A simple and strong baseline for long-term tracking, in: *Conference on Artificial Intelligence (AAAI)*.
- Kalal, Z., Mikolajczyk, K., Matas, J., 2012. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 1409–1422.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105.
- Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J., 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Li, F., Tian, C., Zuo, W., Zhang, L., Yang, M.H., 2018a. Learning spatial-temporal regularized correlation filters for visual tracking, in: *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4904–4913.
- Li, H., Li, Y., Porikli, F., 2016a. Convolutional neural net bagging for online visual tracking. *Computer Vision and Image Understanding* 153, 120–129.
- Li, H., Li, Y., Porikli, F., 2016b. Deeptrack: Learning discriminative feature representations online for robust visual tracking. *IEEE Transactions on Image Processing* 25, 1834–1848.
- Li, P., Wang, D., Wang, L., Lu, H., 2018b. Deep visual tracking: Review and experimental comparison. *Pattern Recognition* 76, 323–338.
- Lin, T.Y., Dollar, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J., 2017. Feature pyramid networks for object detection, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Liu, Y., Han, J., Qiang, Z., Caifeng, S., 2020. Deep salient object detection with contextual information guidance. *IEEE Transactions on Image Processing* 29, 360–374.
- Lu, X., Ma, C., Ni, B., Yang, X., Reid, I., Yang, M.H., 2018. Deep regression tracking with shrinkage loss, in: *European Conference on Computer Vision (ECCV)*.
- Lukezic, A., Vojir, T., Cehovin, L., Matas, J., Kristan, M., 2017. Discriminative correlation filter with channel and spatial reliability, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ma, C., Huang, J.B., Yang, X., Yang, M.H., 2015a. Hierarchical convolutional features for visual tracking, in: *IEEE International Conference on Computer Vision (ICCV)*.
- Ma, C., Yang, X., Zhang, C., Yang, M.H., 2015b. Long-term correlation tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5388–5396.
- Mozhdehi, R.J., Medeiros, H., 2017. Deep convolutional particle filter for visual tracking, in: *IEEE International Conference on Image Processing (ICIP)*.
- Mozhdehi, R.J., Medeiros, H., 2020. Deep convolutional likelihood particle filter for visual tracking, in: *24th International Conference on Image Processing, Computer Vision, & Pattern Recognition (IPCV)*.
- Mozhdehi, R.J., Reznichenko, Y., Siddique, A., Medeiros, H., 2018a. Convolutional adaptive particle filter with multiple models for visual tracking, in: *13th International Symposium on Visual Computing (ISVC)*.
- Mozhdehi, R.J., Reznichenko, Y., Siddique, A., Medeiros, H., 2018b. Deep convolutional particle filter with adaptive correlation maps for visual tracking, in: *IEEE International Conference on Image Processing (ICIP)*.
- Nam, H., Han, B., 2016. Learning multi-domain convolutional neural networks for visual tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ning Wang, N., Zhou, W., Tian, Q., Hong, R., Wang, M., Li, H., 2018. Multi-cue correlation filters for robust visual tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Park, E., Berg, A.C., 2018. Meta-tracker: Fast and robust online adaptation for visual object trackers, in: *European Conference on Computer Vision (ECCV)*.
- Qi, Y., Zhang, S., Qin, L., Yao, H., Huang, Q., Lim, J., Yang, M.H., 2016. Hedged deep tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4303–4311.
- Raju, P.M., Mishra, D., Mukherjee, P., 2021. DA-SACOT: Domain adaptive-segmentation guided attention for correlation based object tracking. *Image and Vision Computing*, 104215.
- Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks, pp. 1–10.
- Shinde, A., Sahu, A., Apley, D., Runger, G., 2014. Preimages for variation patterns from kernel PCA and bagging. *IEEE Transactions* 46, 429–456.
- Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: *International Conference on Learning Representations (ICLR)*.
- Song, Y., Ma, C., Gong, L., Zhang, J., Lau, R., Yang, M.H., 2017. CREST: Convolutional residual learning for visual tracking, in: *IEEE International Conference on Computer Vision (ICCV)*, pp. 2555 – 2564.
- Song, Y., Ma, C., Wu, X., Gong, L., Bao, L., Zuo, W., Shen, C., Lau, W.R., Yang, M.H., 2018. Vital: Visual tracking via adversarial learning, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Sun, C., Wang, H., Lu, H., Yang, M.H., 2018. Correlation tracking via joint discrimination and reliability learning, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tang, M., Feng, J., 2015. Multi-kernel correlation filter for visual tracking, in: *IEEE International Conference on Computer Vision (ICCV)*, pp. 3038–3046.
- Valmadre, J., Bertinetto, L., Henriques, J.F., Vedaldi, Andrea Torr, P.H.S., 2017. End-to-end representation learning for correlation filter based tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wang, F., Franco-Penya, H.H., Kelleher, J.D., Pugh, J., Ross, R., 2017a. An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity, in: *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pp. 291–305.
- Wang, N., Yeung, D.Y., 2013. Learning a deep compact image representation for visual tracking, in: *Conference on Neural Information Processing Systems (NIPS)*.
- Wang, N., Zhou, Q., Tian, R., Hong, R., Wang, M., Li, H., 2018. Multi-cue correlation filters for robust visual tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wang, Q., Gao, J., Xing, J., Zhang, M., Hu, W., 2017b. DCFNet: Discriminant correlation filters network for visual tracking. *arXiv preprint arXiv:1704.04057*.
- Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H., 2019. Fast online object tracking and segmentation: A unifying approach, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- WLukezic, A., Matas, J., Kristan, M., 2020. D3s – a discriminative single shot segmentation tracker, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wu, Y., Lim, J., Yang, M.H., 2015. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Xu, Y., Wang, Z., Li, Z., Yuan, Y., Yu, G., 2020. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines, in: *Conference on Artificial Intelligence (AAAI)*.
- Yan, B., Zhao, H., Wang, D., Lu, H., Yang, X., 2019. 'skimming-perusal' tracking: A framework for real-time and robust long-term tracking, in: *IEEE Conference on Computer Vision (ICCV)*.
- Yuan, D., Lu, X., Li, D., Liang, Y., Zhang, X., 2019. Particle filter re-detection for visual tracking via correlation filters. *Multimedia Tools and Applications* 78, 14277–14301.
- Zhang, J., Ma, S., Sclaroff, S., 2014. Robust tracking via multiple experts, in: *European Conference on Computer Vision (ECCV)*, pp. 188–203.
- Zhang, L., Gonzalez-Garcia, A., Van De Weijer, J., Danelljan, M., Khan, F.S., 2019. Learning the model update for siamese trackers, in: *IEEE Conference*

- on Computer Vision (ICCV).
- Zhang, M., Wang, Q., Xing, J., Gao, J., Peng, P., Hu, W., Maybank, S., 2018a. Visual tracking via spatially aligned correlation filters network, in: European Conference on Computer Vision (ECCV), pp. 469–485.
- Zhang, T., Liu, S., Xu, C., 2018b. Correlation particle filter for visual tracking. *IEEE Transactions on Image Processing* 27, 2676–2687.
- Zhang, T., Xu, C., Yang, M.H., 2017. Multi-task correlation particle filter for robust object tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Zhang, Z., Peng, H., 2019. Deeper and wider siamese networks for real-time visual tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Zhang, Z., Peng, Z., Fu, J., Li, B., Hu, W., 2020. Ocean: Object-aware anchor-free tracking, in: European Conference on Computer Vision (ECCV).
- Zhong, W., Lu, H., Yang, M.H., 2014. Robust object tracking via sparse collaborative appearance model. *IEEE Transactions on Image Processing* 23, 2356–2368.
- Zhu, Z., Wu, W., Zou, W., Yan, J., 2018. End-to-end flow correlation tracking with spatial-temporal attention, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 548–557.