

Automatic Failure Recovery and Re-Initialization for Online UAV Tracking with Joint Scale and Aspect Ratio Optimization

Fangqiang Ding¹, Changhong Fu^{1,*}, Yiming Li¹, Jin Jin¹ and Chen Feng²

Abstract—Current unmanned aerial vehicle (UAV) visual tracking algorithms are primarily limited with respect to: (i) the kind of size variation they can deal with, (ii) the implementation speed which hardly meets the real-time requirement. In this work, a real-time UAV tracking algorithm with powerful size estimation ability is proposed. Specifically, the overall tracking task is allocated to two 2D filters: (i) translation filter for location prediction in the space domain, (ii) size filter for scale and aspect ratio optimization in the size domain. Besides, an efficient two-stage re-detection strategy is introduced for long-term UAV tracking tasks. Large-scale experiments on four UAV benchmarks demonstrate the superiority of the presented method which has computation feasibility on a low-cost CPU.

I. INTRODUCTION

Equipped with visual perception capability, robots can have flourishing real-world applications, *e.g.*, visual object tracking has stimulated broad practical utilities like human-robot collaboration [1], robotic arm manipulation [2], and aerial filming [3]. Tracking onboard unmanned aerial vehicle (UAV) has many advantages over general object tracking, for instance, broad view scope, high flexibility, and mobility. Yet more difficulties are introduced such as aspect ratio change (ARC)¹, out-of-view, exiguous calculation resources, *etc.* Hence, a robust-against-ARC, low-cost, and energy-efficient tracking algorithm applicable in long short-term tasks is highly desirable for UAV tracking.

In literature, although deep feature [4]–[6] or deep architecture [7]–[9] can exceedingly boost the tracking robustness, the complex convolution operations have hampered their practical utility. Another research direction in visual tracking is discriminative correlation filters (DCF) [10]–[13]. With only hand-crafted features, DCF-based trackers mostly have real-time speed on a single CPU thanks to their transforming intractable spatial convolution into element-wise multiplication in the Fourier domain. While most researches focus on location and scale estimation, scarce of them focus on aspect ratio. Current DCF-based trackers commonly fix the aspect ratio of the object during tracking. Consequently, in UAV tracking scenarios with extensive ARC, erroneous

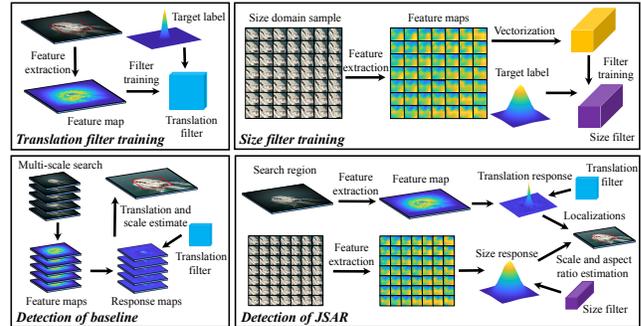


Fig. 1. Comparison of overall flow in baseline [14] (left) and JSAR (right). Baseline only trains a translation filter for translation estimation in the detection phase and updates object scale by brute-force multi-scale search strategy. JSAR proposes to train size filter in size domain by multi-size sampling and jointly estimate the scale and aspect ratio of the object in detection stage.

appearance is frequently introduced in filter training because of the imprecise size estimation, leading to filter degradation.

Inspired by 1D scale filter [15] aiming to handle inefficiency of brute-force multi-scale search [12], [16]–[18], this work proposes a joint scale and aspect ratio optimization tracker (JSAR) to achieve accurate scale and aspect ratio estimation. As displayed in Figure 1, the training procedure is two-fold: (i) training a translation filter with a single patch [14] for location prediction, and (ii) training a size filter with exponentially-distributed samples for scale and aspect ratio estimation. Consequently, location, scale, and aspect ratio are calculated simultaneously, *i.e.*, the object bounding box can be estimated in the 4-DoF (degree of freedom) space, promoting the tracking accuracy without losing much speed.

Recently, combining visual tracking with re-detection framework has raised precision in the long-term tracking scenarios where objects frequently suffer from out-of-view or full occlusion [19], [20]. Yet the speed is mostly sacrificed due to the intractable object detection methods. In this work, a CPU-friendly re-detection strategy is proposed to enable long-term tracking. An effective tracking failure monitoring mechanism and an efficient re-initialization method based on EdgeBoxes [21] collaboratively contribute to the smooth long-term tracking. Our main contributions are three-fold:

- A novel robust tracking method with real-time speed is proposed with joint scale and aspect ratio optimization.
- A new CPU-friendly re-detection framework is developed to accomplish long-term tracking tasks efficiently.
- Large-scale experiments conducted on three short-term UAV benchmarks and one long-term benchmark validate the outstanding performance of the proposed

*Corresponding author

¹Fangqiang Ding, Changhong Fu, Yiming Li and Jin Jin are with the School of Mechanical Engineering, Tongji University, 201804 Shanghai, China. changhongfu@tongji.edu.cn

²Chen Feng is with the Tandon School of Engineering, New York University, NY 11201 New York, United States. cfeng@nyu.edu

The code and tracking video are respectively released on <http://github.com/vision4robotics/JSAR-Tracker> and <https://youtu.be/wNFhkpH6osk>.

¹Caused by rapid attitude alteration and strong motion of UAV, ARC is generally brought with the form of large viewpoint variation, intense rotation, deformation, to name a few.

method.

II. RELATED WORKS

A. Discriminative correlation filter tracking algorithm

In literature, discriminative tracking algorithms train a classifier to differentiate the tracked object from the background by maximizing the classification score. Recent investigations focus on discriminative correlation filters since D. S. Bolme *et al.* [22] proposed to learn robust filters by mapping the training samples to the desired output. J. F. Henriques *et al.* [10] presented to solve the rigid regression equation in the Fourier domain, and established the basic structure of modern DCF methods. Afterwards, several attempts are made to promote tracking performance within DCF framework, *e.g.*, spatial penalization [17], [23], multi-feature fusion [11], [24], and real negative sampling [12], [25]. However, most research highlight the improvement of localization accuracy rather than amelioration in the size estimation.

B. Prior works in object size estimation

Pioneer DCF trackers [10], [22], [26] fix the object size and only estimate the trajectory in the 2D space. Presetting a scaling pool, [16] sampled on different scales to find the optimal one in the detection phase. [27] proposed a separate scale correlation filter to estimate scale variance in the 1D scale domain. To enable aspect ratio estimation, [28] tackled scale and aspect ratio variation by embedding detection proposals generator in tracking pipeline. [5] enforced near-orthogonality constraint on center and boundary filters. Despite bringing more freedoms in object tracking, these two methods bring heavy computation burden for DCF trackers, and are hence not satisfactory alternatives for UAV tracking.

C. Re-detection in object tracking

Tracking-learning-detection (TLD) [29] is proposed to validate tracking results and decide whether to enable learning and detection. Among DCF trackers, [20] introduced an online random fern to generate candidates and score each of them for re-detection. Despite the effectiveness, it is time-consuming due to the scanning window strategy. [19] presented a novel multi-threading framework in which an offline-trained Siamese network is used as a verifier. However, speed is largely decreased. This work utilizes EdgeBoxes [21] to quickly generate proposals, and then a decision filter is applied to select the most possible bounding box for tracker's re-initialization. The proposed two-stage re-detection strategy is more efficacious and light-weight.

D. UAV tracking

In UAV tracking scenarios, the tracked objects possess higher motion flexibility than in tracking based on hand-held or fixed surveillance cameras. Therefore, UAV tracking is confronted with more difficulties. In literature, aberrance repression [30] and intermittent context learning [31], [32] are proposed to improve tracking precision. Despite obtaining appealing results, they cannot estimate aspect ratio variation. Adaptive to ARC, JSAR has better robustness and real-time speed, and hence superior to other trackers in UAV tracking.

III. PROPOSED METHOD

A. Discriminative correlation filter

In frame t , a multi-channel correlation filter $\mathbf{W}_t \in \mathbb{R}^{M \times N \times D}$ is trained by restricting its correlation result with training samples $\mathbf{X}_t \in \mathbb{R}^{M \times N \times D}$ to the given target label $\mathbf{g} \in \mathbb{R}^{M \times N}$. The minimized objective $\mathcal{E}(\mathbf{W}_t)$ is formulated as the sum of least square term and regularization term:

$$\mathcal{E}(\mathbf{W}_t) = \left\| \sum_{d=1}^D \mathbf{w}_t^d \star \mathbf{x}_t^d - \mathbf{g} \right\|_2^2 + \lambda \sum_{d=1}^D \left\| \mathbf{w}_t^d \right\|_2^2, \quad (1)$$

where $\mathbf{x}_t^d \in \mathbb{R}^{M \times N}$ and $\mathbf{w}_t^d \in \mathbb{R}^{M \times N}$ respectively indicate the d -th channel feature representation and filter, and \star denotes the cyclic correlation operator. M and N denote the width and height of a single channel sample while D denotes the number of feature channels. λ is a hyper parameter for avoiding over-fitting. Minimizing the objective in the Fourier domain, a closed-form solution of filter \mathbf{W}_t is obtained:

$$\tilde{\mathbf{w}}_t^d = \frac{\tilde{\mathbf{g}} \odot \tilde{\mathbf{x}}_t^{d*}}{\sum_{d=1}^D (\tilde{\mathbf{x}}_t^d \odot \tilde{\mathbf{x}}_t^{d*}) + \lambda}, \quad (2)$$

where \odot and \div denote the element-wise multiplication and division, respectively. $\tilde{\cdot}$ means discrete Fourier transform (DFT) and \cdot^* means complex conjugation. The appearance model is updated by linear interpolation with a predefined learning rate θ . Use \mathcal{F}^{-1} to denote inverse discrete Fourier transform and \mathbf{m}_t^d as the d -th feature representation of the search region, the response map \mathbf{R}_t is obtained by:

$$\mathbf{R}_t = \mathcal{F}^{-1} \left(\sum_{d=1}^D \tilde{\mathbf{w}}_{t-1}^{d*} \odot \tilde{\mathbf{m}}_t^d \right). \quad (3)$$

B. Translation estimation

For translation estimation, most trackers [10], [11], [24] learn a 2D translation filter $\mathbf{W}_{t,trans}$ in the space domain by Eq. (2). In the training phase, the region of interest (ROI) is cropped centered at object location with a fixed proportion to the object scale. In the detection phase, the feature of ROI centered at the location of the last frame $\mathbf{M}_{t,trans}$ is extracted. The object is localized by finding the peak position of the response map generated by Eq. (3). For scale estimation, classical brute-force search in a multi-scale hierarchical structure is inefficient due to repetitive feature extraction on large image patches.

C. Size estimation

Motivated by [27] which trains a 1D scale filter in the scale domain for scale-adaptive tracking, we propose to train a 2D size filter $\mathbf{W}_{t,size}$. Different to the 2D translation filter learned in the space domain which is composed of a horizontal and vertical axis, the samples are extracted in the size domain consisting of a scale axis and aspect ratio axis.

1) *Sampling in size domain:* During size filter training, we crop $S \times A$ patches centered at current object location, where A and S represent the number of aspect ratios and scales in training sample. The size of these patches is calculated by:

$$\{W_t^{s,a}, H_t^{s,a}\} = \{W_t \gamma^{N_s} \phi^{N_a}, H_t \gamma^{N_s} / \phi^{N_a}\} \quad (s = 1, 2, \dots, S, a = 1, 2, \dots, A), \quad (4)$$

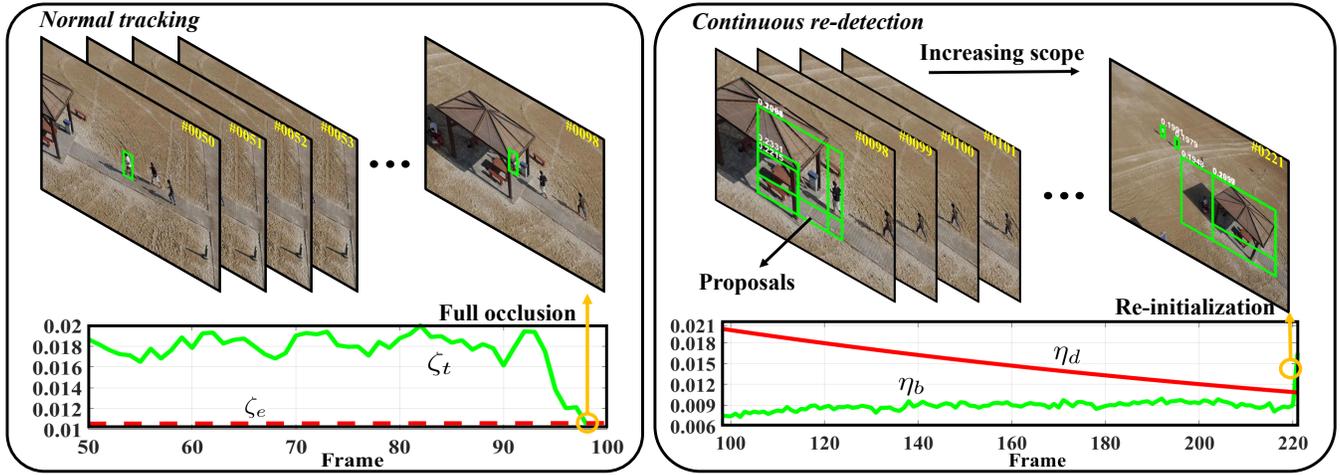


Fig. 2. Overall flow of re-detection strategy. When the maximum value of the response map in frame t (ζ_t) is larger than the threshold value (ζ_e), tracking procedure is normally implemented, if not, re-detection mechanism is activated. When the peak value of the response map η_b generated by selected proposal and decision filter exceeds the descending threshold η_d , the bounding box is re-initialized and re-starts to be tracked normally. It is note that the values remarked above proposals are confidence scores k^i and the search scope is increasing in continuous re-detection.

where W_t and H_t are the object width and height in frame t , and $\{s, a\}$ denotes the index of patches with various scale and aspect ratio. To maintain sample symmetry, we set $N^s = -\frac{S+1}{2} + s$ and $N^a = -\frac{A+1}{2} + a$. Otherwise, γ and ϕ is hyper parameters to control sampling step. To make sure the dimension consistency of cropped patches and reduce computation burden, all patches are downsampled to a presetting model size $\{W_{model}, H_{model}\}$. Afterwards, feature map $\mathbf{V}_t^{s,a} \in \mathbb{R}^{\frac{W_{model}}{C} \times \frac{H_{model}}{C} \times K}$ is extracted on each patch with K feature channels. Here, C denotes the side length of single cell for feature extraction. Different to translation filter which utilizes original feature map for training, for each patch, the extracted feature representations are vectorized to 1D vector $\mathbf{v}_t^{s,a} = \text{vec}(\mathbf{V}_t^{s,a}) \in \mathbb{R}^{\frac{W_{model}H_{model}K}{C^2}}$, as exhibited in Fig. 1. In this process, the number of feature channel changes from original K to $C = \frac{W_{model}H_{model}K}{C^2}$. By stacking column vectors from different patches, the final sample $\mathbf{X}_{t,size} \in \mathbb{R}^{S \times A \times C}$ can be denoted by:

$$\mathbf{X}_{t,size} = \begin{bmatrix} \mathbf{v}_t^{1,1} & \mathbf{v}_t^{1,2} & \mathbf{v}_t^{1,3} & \dots & \mathbf{v}_t^{1,A} \\ \mathbf{v}_t^{2,1} & \mathbf{v}_t^{2,2} & \mathbf{v}_t^{2,3} & \dots & \mathbf{v}_t^{2,A} \\ \mathbf{v}_t^{3,1} & \mathbf{v}_t^{3,2} & \mathbf{v}_t^{3,3} & \dots & \mathbf{v}_t^{3,A} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{v}_t^{S,1} & \mathbf{v}_t^{S,2} & \mathbf{v}_t^{S,3} & \dots & \mathbf{v}_t^{S,A} \end{bmatrix}. \quad (5)$$

2) *Size estimation*: After sample extraction, Eq. (2) is applied to learn size filter $\mathbf{W}_{t,size}$. In the estimation stage, we assume the size is unchanged and estimate the location translation at first when a new frame comes. Centering at the predicted location, the feature representation of the search region in size domain $\mathbf{M}_{t,size}$ is extracted for size estimation, as shown in Fig. 1. It is noted that $\mathbf{M}_{t,size}$ has the same dimension as training sample $\mathbf{X}_{t,size}$. By Eq. (3), the current scale and aspect ratio are obtained by maximizing the response score, and then the object size is optimized. In a word, our method can be generally applicable in the DCF framework, and work in the 4-DoF space.

D. Re-detection strategy

As displayed in Fig. 2, the re-detection will be implemented when tracking failure is observed. The proposed re-detection strategy has two stages, *i.e.*, object proposals generation, and candidates scoring. Object proposal method EdgeBoxes [22] is applied to generate candidates and decision filter scores them for object re-initialization. The detailed illustration is as follows.

1) *Tracking failure monitoring mechanism*: Ideally, the re-detection is enabled when the object is lost or the output deviates greatly against real object location. Related to the tracking confidence, the peak value $\zeta_t = \max(\mathbf{R}_{t,trans})$ of response map generated in translation estimation of frame t is adopted to decide whether to activate the re-detection. Presetting a threshold ζ_e , the re-detection mechanism is activated when $\zeta_t < \zeta_e$.

2) *Object proposal generation*: When re-detection begins, EdgeBoxes [21] is utilized to generate class-agnostic object proposals within surrounding square area at the first stage. The side length of this surrounding area is $\omega\sqrt{W_tH_t}$ in this work. Each proposal \mathbf{b}^i generated by EdgeBoxes has five variables, *i.e.*, $\mathbf{b}^i = [x^i, y^i, w^i, h^i, k^i]$. The first four variables denote the location and size of the proposal while the last value k^i is the confidence score. Depending on the confidence score, we choose the top N_e proposals for re-detection.

3) *Object proposals scoring*: During the second stage, for each proposal, the feature from ROI in frame t is extracted with K feature channels, which is denoted by $\mathbf{P}_t^i (i = 1, 2, \dots, N_e)$. To make a final decision for re-initialization, a decision filter \mathbf{M}_{deci} is trained along with translation filter using selected pure samples². After feature extraction of pro-

²The selection of pure samples also depends on the peak value of response map in frame t : if the peak value $\zeta_t > \zeta_s$ (ζ_s is a predefined threshold), the sample for translation filter training is adopted to update the decision filter because larger peak value indicates better tracking quality.

Algorithm 1: JSAR-Re

Input: Object location and size at the first frame
Subsequent images in the video sequence
Output: Location and size of object in frame t

```
1 if  $t = 1$  or re-detection enabled then
2   Extract training samples  $\mathbf{X}_{i,trans}$  and  $\mathbf{X}_{i,size}$ 
3   Use Eq. (2) to initialize  $\mathbf{W}_{i,trans}$  and  $\mathbf{W}_{i,size}$ 
4   Initialize  $\mathbf{M}_{deci}$  by  $\mathbf{X}_{i,trans}$ , disable re-detection
5 else
6   Extract search region feature maps  $\mathbf{M}_{t,trans}$ 
7   Generate  $\mathbf{R}_{t,trans}$  by Eq. (3) and find  $\zeta_t$ 
8   if  $\zeta_t > \zeta_e$  then
9     Estimate object translation and extract  $\mathbf{M}_{i,size}$ 
10    Estimate object size using Eq. (3)
11    Use Eq. (2) to update  $\mathbf{W}_{t,trans}$  and  $\mathbf{W}_{t,size}$ 
12    if  $\zeta_t > \zeta_s$  then
13      Update filter  $\mathbf{M}_{deci}$  using  $\mathbf{M}_{t,trans}$ 
14    end
15  end
16  Generate proposals with search area
17  Scoring proposals using  $\mathbf{M}_{deci}$  by Eq. (3)
18  Find the largest peak value  $\eta_b$ 
19  if  $\eta_b > \eta_d$  then
20    Enable re-detection, initialize the object
21  else
22    Increase  $\omega$  and reduce  $\eta_d$ 
23    Continue to re-detect next frame
24  end
25 end
26 end
```

posals $\mathbf{P}_i^i (i = 1, 2, \dots, N_e)$, the corresponding N_e response maps are calculated through the correlation of decision filter and feature maps by Eq. (3), and then the proposal with the largest peak value is selected. However, in the scenarios of out-of-view and full occlusion, the selected proposal is generally fallacious. To this end, we set a threshold η_d to decide whether to re-initialize: if the selected proposal's peak value $\eta_b > \eta_d$, the re-initialization will be enabled, or else, re-detection is continued.

In this work, the scale of the search area is increased and the re-initialization threshold η_d is reduced frame-by-frame in re-detection failure cases to make sure the re-initialization ultimately works. The overall flow of the proposed method is presented in Algorithm 1.

IV. EXPERIMENTS

In this section, the proposed method is evaluated on three challenging short-term UAV benchmarks, *i.e.*, UAVDT [33], UAV123@10fps [34], DTB70 [35] and one long-term benchmark UAV20L [34], including over 149K images overall captured by drone camera in all kinds of harsh aerial scenarios. The experimental result of our method is compared with 30 state-of-the-art (SOTA) approaches including 14 deep trackers, *i.e.*, SiameseFC [7], DSiam [9], IBCCF [5], ECO [6], C-COT [4], GOTURN [38], PTAV [19], DeepSTRCF [5],

CFNet [39], ASRCF [23], MCCT [24], ADNet [40], TADT [8], UDT+ [41], and 16 hand-crafted trackers, *i.e.*, MCCT-H [24], KCF [10], DSST [27], fDSST [15], ECO-HC [6], DCF [10], BACF [12], ARCF [30], SRDCF [17], STAPLE-CA [25], ARCF-H [30], STAPLE [11], SRDCFdecon [37], CSR-DCF [36], KCC [42], STRCF [14].

A. Implementation details

To test the size estimation ability of JSAR, first of all, experiments are conducted on three short-term UAV benchmarks [33]–[35], compared with both deep and hand-crafted trackers. Then, the re-detection module is added to cope with tracking failure, generating JSAR-Re. We evaluate JSAR-Re with SOTA trackers on UAV20L dataset [34].

1) **Platform:** All experiments are implemented with MATLAB R2018a and all experimental results are obtained on a computer with a single i7-8700K (3.70GHz) CPU, 32GB RAM, and an NVIDIA RTX 2080 GPU for fair comparisons.

2) **Baseline:** In this work, spatial-temporal regularized correlation filter (STRCF) [14] is selected as our baseline tracker which adds spatio-temporal regularized term to training objective for improving robustness and adopts a multi-scale search strategy for scale adaptivity. Discarding the hierarchical scale searching, JSAR separately trains a size filter using Eq. (2) to estimate the scale and aspect ratio variations and follows the translation estimation in [14].

3) **Features:** To guarantee real-time performance on a low-cost CPU, we only apply hand-crafted features to our tracker for experiments. Gray-scale, histogram of oriented gradient (HOG) [10] and color name (CN) [43] are employed in translation filter, while size filter only uses HOG.

4) **Hyper Parameters:** The main parameters in this work are listed in Table I. For impartial comparison, all the parameters are fixed in the experiments.

5) **Criteria:** Following one-pass evaluation (OPE) [44] protocol, we evaluate all trackers by two measures, *i.e.*, precision and success rate. Precision plots can exhibit the percentage of all input images in which the distance of predicted location with ground truth one is smaller than various

TABLE I
FOR IMPARTIAL COMPARISON, THESE PARAMETERS ARE FIXED
IN ALL EVALUATION OF OUR TRACKERS

Symbol	Value	Meaning
S	13	The number of scales
A	13	The number of aspect ratios
γ	1.03	Scale sampling step
ϕ	1.02	Aspect ratio sampling step
θ_{size}	0.014	The learning rate of size filter
W_{model}	16	The width of model size
H_{model}	32	The height of mode size
C	4	The side length of feature cell
ζ_e	0.0105	Re-detection enablement threshold
ζ_s	0.013	Decision filter update threshold
η_d	0.02	Re-initialization threshold
ω	5	The side length factor of re-detection area
N_e	30	The number of proposals for re-detection

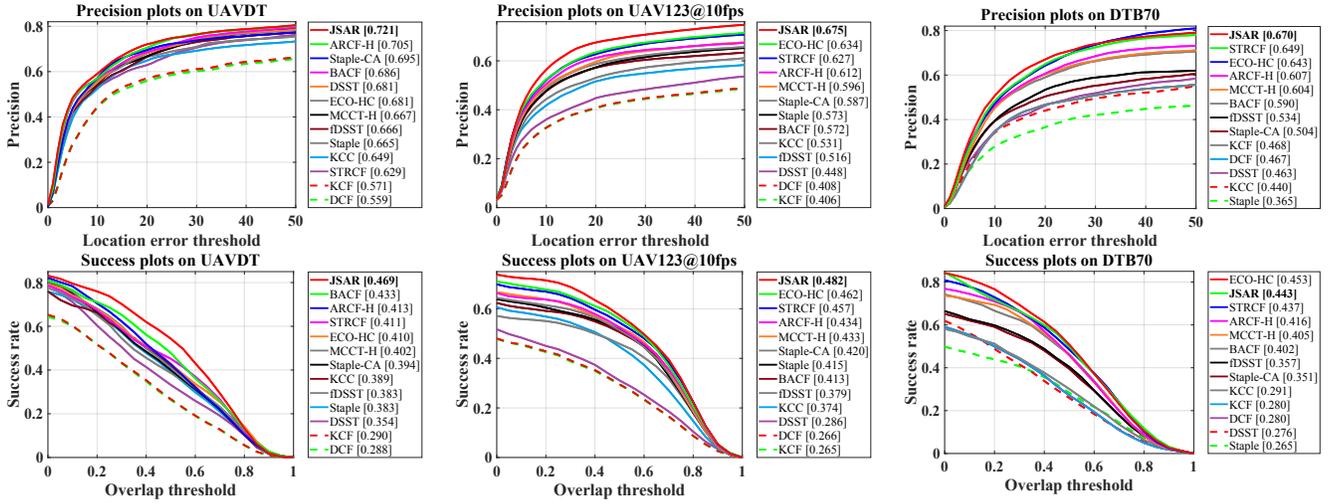


Fig. 3. Overall performance of hand-crafted real-time trackers on (a) UAVDT [33] (b) UAV123@10fps [34] (c) DTB70 [35]. JSAR has a notable improvement of 8.3% and 4.3% in terms of AUC on UAVDT and UAV123@10fps compared with the second best trackers, respectively.

TABLE II

AVERAGE PRECISION, AUC AND SPEED COMPARISON OF TOP 10 HAND-CRAFTED TRACKERS ON UAVDT [33], UAV123@10FPS [34] AND DTB70 [35]. RED, GREEN AND BLUE RESPECTIVELY MEAN THE FIRST, SECOND AND THIRD PLACE.

Tracker	Real time					Non-real time				
	JSAR	MCCT-H [24]	STRCF [14]	ARCF-H [30]	BACF [12]	ECO-HC [6]	CSR-DCF [36]	SRDCF [17]	ARCF-HC [30]	SRDCFdecon [37]
AUC	0.465	0.413	0.435	0.421	0.416	0.442	0.426	0.416	0.468	0.397
Precision	0.689	0.622	0.635	0.641	0.616	0.653	0.654	0.616	0.693	0.577
Speed(fps)	32.2	59.7	28.5	51.2	56.0	69.3	12.1	14.0	15.3	7.5
Conference	This work.	CVPR'18	CVPR'18	ICCV'19	CVPR'17	CVPR'17	CVPR'17	ICCV'15	ICCV'19	CVPR'16

TABLE III

PRECISION, AUC AND SPEED COMPARISON BETWEEN 14 RECENT DEEP TRACKERS ON UAVDT [33]. RED, GREEN, BLUE AND ORANGE RESPECTIVELY MEAN THE FIRST, SECOND, THIRD AND FOURTH PLACE.

Tracker	AUC	Precision	Speed(fps)	CPU/GPU	Conference
JSAR	0.469	0.721	35	CPU	This work.
GOTURN [38]	0.451	0.702	17	GPU	ECCV'16
IBCCF [5]	0.388	0.603	3	GPU	CVPR'17
TADT [8]	0.431	0.677	35	GPU	CVPR'19
DSiam [9]	0.457	0.704	16	GPU	ICCV'17
PTAV [19]	0.384	0.675	27	GPU	ICCV'17
ECO [6]	0.454	0.700	16	GPU	CVPR'17
ASRCF [23]	0.437	0.700	24	GPU	CVPR'19
MCCT [24]	0.437	0.671	9	GPU	CVPR'18
CFNet [39]	0.428	0.680	41	GPU	CVPR'17
C-COT [4]	0.406	0.656	1	GPU	ECCV'16
ADNet [40]	0.429	0.683	8	GPU	CVPR'17
UDT+ [41]	0.416	0.697	60	GPU	CVPR'19
SiameseFC [7]	0.465	0.708	38	GPU	ECCV'16
DeepSTRCF [14]	0.437	0.667	6	GPU	CVPR'18

thresholds, and success plots can reflect the proportion of frames in which the intersection over union (IoU) between the estimated bounding box and the ideal one is greater than distinctive thresholds. The score at 20 pixel and area under curve (AUC) are respectively used to rank the trackers.

B. JSAR vs. deep trackers

We first compare the tracking performance of JSAR with 14 recently proposed SOTA deep trackers, *i.e.*, deep features based trackers and deep convolution neural networks (DCNN) based trackers, on UAVDT benchmark [33]. As shown in Table III, JSAR has taken the first place in both precision and AUC, while coming fourth in tracking speed running on a low-cost CPU. Without robust deep features, the remarkable improvement (7.3% and 8.1% than DeepSTRCF in terms of AUC and precision) can be attributed to the ARC adaption, because UAVDT mainly addresses vehicle tracking and the viewpoint change can easily lead to ARC of the tracked vehicle in the image, as shown in Figure 4.

C. JSAR vs. hand-crafted trackers

1) Overall evaluation: Restricted by scarce computation resources, deep trackers have difficulties meeting real-time tracking speed on UAV. Hand-crafted trackers, *i.e.*, using hand-crafted features in DCF framework, are ideal choices in UAV tracking for its calculation efficiency. In this subsection, twelve SOTA and real-time hand-crafted trackers are used for comparison with JSAR at first. JSAR outperforms other real-time trackers in terms of precision and AUC, as displayed in Fig. 3. Notably, compared with the baseline STRCF [14], JSAR respectively improves the



Fig. 4. Display of tracking results from eight hand-crafted trackers on twelve UAV video, *i.e.*, *S0301*, *S0602*, *S0701* of UAVDT [33], *boat3*, *boat4*, *wakeboard1*, *wakeboard5* of UAV123@10fps [34] and *ChasingDrones*, *SnowBoarding4*, *Surfing03*, *Surfing06*, *SpeedCar4* of DTB70 [35].

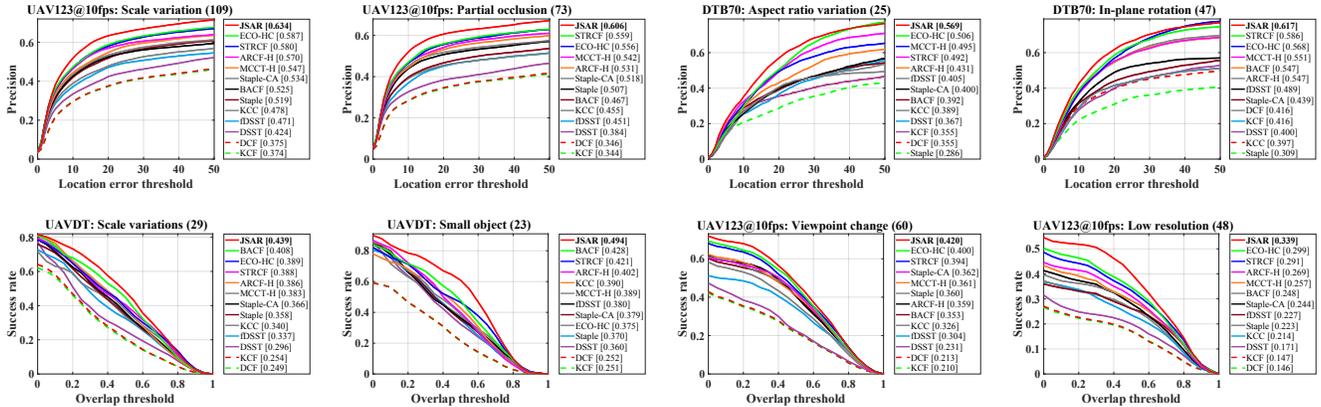


Fig. 5. Attribute-oriented comparison with hand-crafted real-time trackers. Precision plots of four attributes, *i.e.*, scale variation, partial occlusion, aspect ratio variation, in-plane rotation, and success plots of four attribute, *i.e.*, scale variations, small object, viewpoint change and low resolution are presented.

AUC by 14.1%, 5.5%, and the precision by 14.6%, 7.7% on UAVDT and UAV123@10fps. We further compare the average performance of best 10 hand-crafted trackers on three benchmarks [33]–[35], as shown in Table II. It can be seen JSAR obtained the second place in both AUC and precision, however, JSAR has a tiny gap compared with the best tracker ARCF-HC [30] (0.003 and 0.004 in AUC and precision), and it has remarkably improved the speed by 110%. Hence, compared to ARCF-HC, JSAR performs comparably with much higher efficiency. Averagely, JSAR

gains an improvement of 6.9% in AUC and 8.5% in precision compared with the baseline method, *i.e.*, STRCF.

2) *Attribute-oriented evaluation*: Fig. 5 exhibits the precision and success plots of real-time trackers on eight challenging attributes from UAV123@10fps [34], UAVDT [33] and DTB70 [35]. It can be seen that JSAR has respectively improved the precision by 8%, 12.5%, and 5.3% compared with the second-best trackers in the attributes of scale variation, aspect ratio variation, and in-plane rotation. As for AUC, JSAR gains a 7.6% improvement in scale variations and

a 5.0% improvement in viewpoint change. The remarkable improvements demonstrate the effectiveness of the size filter in scale and aspect ratio change cases. Besides, in partial occlusion, small object, and low resolution, JSAR still outperforms other real-time trackers dramatically, exhibiting its excellent generality in various aerial scenarios.

D. Hyper parameters analysis

We analyze the impacts of five core hyper parameters in the proposed size filter, including the sampling step γ and ϕ , the learning rate of size filter θ , and the number of scale as well as aspect ratio (S/A). The impacts on AUC and precision of the first three parameters are exhibited in Figure 6, from which it can be seen they have a relatively small influence on tracking performance (with precision from 0.672 to 0.721, AUC from 0.418 to 0.469), which demonstrates the strong robustness of JSAR. The comparison of tracking performance and speed of various S/A configurations are displayed in Table IV. From 9 to 21, the number of scales/aspect ratios has little influence on both AUC (ranging from 0.452 to 0.469) and precision (ranging from 0.704 to 0.721). Yet the results rapidly fall off when the value of S/A is 7. This situation can be explained by insufficient samples for size filter training.

E. Re-detection evaluation

To testify the effectiveness of our proposed re-detection strategy, we conduct experiments on JSAR-Re and JSAR with eleven SOTA trackers on long-term UAV20L benchmarks, which consists of 20 long-term image sequences with over 2.9K frames per sequence averagely. The precision plot is reported in Figure 7. JSAR-Re ranks No.1 and improves

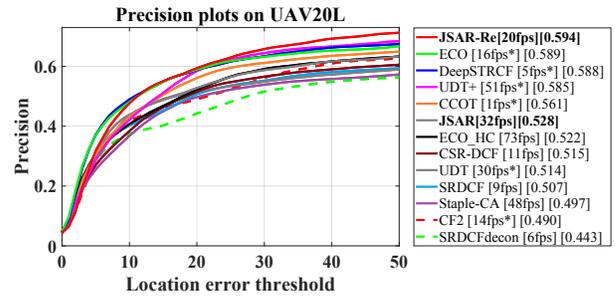


Fig. 7. Precision plots with tracking speed of JSAR-Re, JSAR and eleven SOTA trackers on UAV20L dataset [34]. * denotes this tracker is tested on GPU.

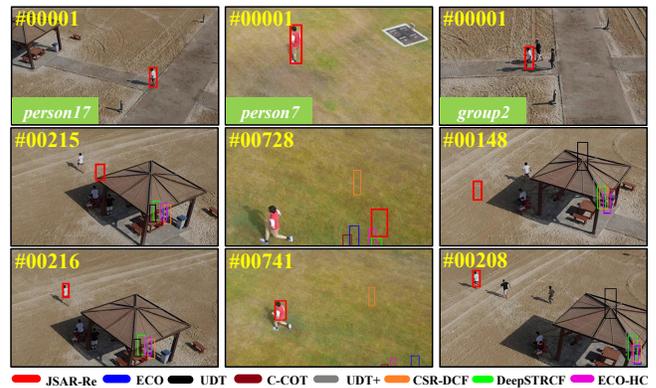


Fig. 8. Qualitative tracking performance of JSAR and seven SOTA trackers on *person14*, *person7* and *group2* of UAV20L dataset [34].

the tracking precision by 11.3% compared with JSAR, with a speed of 20fps on a low-cost CPU. Some qualitative results are exhibited in Figure 8.

V. CONCLUSIONS

In this work, a novel UAV tracking framework of joint scale and ARC estimation is proposed. Also, an object proposal based re-detection algorithm is introduced to achieve long-term tracking. Experimental comparison with 30 SOTA trackers exhibits the superiority of our method. Most tellingly, our method can outperform SOTA deep trackers on UAVDT [33] with only hand-crafted features. Using C++ implementation can further raise the tracking speed for real-world as well as real-time UAV applications.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 61806148).

REFERENCES

- [1] O. Palinko, F. Rea, G. Sandini, and A. Sciutti, “Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration,” in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016, pp. 5048–5054.
- [2] M. Hofer, L. Spannagl, and R. D’Andrea, “Iterative Learning Control for Fast and Accurate Position Tracking with an Articulated Soft Robotic Arm,” in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019, pp. 6602–6607.

TABLE IV

DEMOGRAPHIC PREDICTION PERFORMANCE COMPARISON BY THREE EVALUATION METRICS.

S/A	5	7	9	11	13	15	17	19	21
AUC	0.373	0.397	0.452	0.465	0.469	0.454	0.455	0.452	0.454
Precision	0.696	0.685	0.715	0.719	0.721	0.704	0.712	0.709	0.713
Speed(fps)	50.2	46.9	43.1	38.5	35.1	31.1	27.7	24.0	21.7

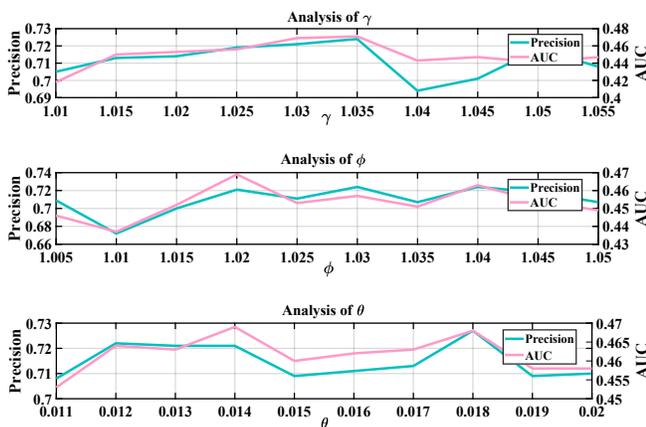


Fig. 6. Sensitivity analysis of four parameters (γ , ϕ , θ and S/A) on UAVDT [33]. It is noted that we fixed the untested parameters in analysis.

- [3] R. Bonatti, C. Ho, W. Wang, S. Choudhury, and S. Scherer, "Towards a Robust Aerial Cinematography Platform: Localizing and Tracking Moving Targets in Unstructured Environments," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and System*, 2019, pp. 229–236.
- [4] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proceedings of European Conference on Computer Vision*. Springer, 2016, pp. 472–488.
- [5] F. Li, Y. Yao, P. Li, D. Zhang, W. Zuo, and M. Yang, "Integrating Boundary and Center Correlation Filters for Visual Tracking with Aspect Ratio Variation," in *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2001–2009.
- [6] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient Convolution Operators for Tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6931–6939.
- [7] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Proceedings of European Conference on Computer Vision*. Springer, 2016, pp. 850–865.
- [8] X. Li, C. Ma, B. Wu, Z. He, and M. Yang, "Target-Aware Deep Tracking," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1369–1378.
- [9] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning Dynamic Siamese Network for Visual Object Tracking," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 1781–1789.
- [10] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 583–596, 2015.
- [11] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary Learners for Real-Time Tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1401–1409.
- [12] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning Background-Aware Correlation Filters for Visual Tracking," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 1135–1143.
- [13] Y. Li, C. Fu, F. Ding, Z. Huang, and G. Lu, "Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11923–11932.
- [14] F. Li, C. Tian, W. Zuo, L. Zhang, and M. Yang, "Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4904–4913.
- [15] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative Scale Space Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1561–1575, 2017.
- [16] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proceedings of European Conference on Computer Vision*. Springer, 2014, pp. 254–265.
- [17] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Learning Spatially Regularized Correlation Filters for Visual Tracking," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 4310–4318.
- [18] F. Li, C. Fu, F. Lin, Y. Li, and P. Lu, "Training-set distillation for real-time uav object tracking," in *Proceedings of IEEE International Conference on Robotics and Automation*, 2020, pp. 1–8.
- [19] H. Fan and H. Ling, "Parallel Tracking and Verifying: A Framework for Real-Time and High Accuracy Visual Tracking," in *Proceedings of IEEE International Conference on Computer Vision*, 2017.
- [20] C. Ma, X. Yang, Chongyang Zhang, and M. Yang, "Long-term correlation tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5388–5396.
- [21] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proceedings of European Conference on Computer Vision*. Springer, 2014, pp. 391–405.
- [22] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2544–2550.
- [23] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual Tracking via Adaptive Spatially-Regularized Correlation Filters," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4665–4674.
- [24] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue Correlation Filters for Robust Visual Tracking," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4844–4853.
- [25] M. Mueller, N. Smith, and B. Ghanem, "Context-Aware Correlation Filter Tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1387–1395.
- [26] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proceedings of European Conference on Computer Vision*. Springer, 2012, pp. 702–715.
- [27] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proceedings of British Machine Vision Conference*, 2014, pp. 1–5.
- [28] D. Huang, L. Luo, W. Wen, Z. Chen, and C. Zhang, "Enable scale and aspect ratio adaptability in visual tracking with detection proposals," in *Proceedings of British Machine Vision Conference*, 2015.
- [29] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.
- [30] Z. Huang, C. Fu, Y. Li, F. Lin, and P. Lu, "Learning Aberrance Repressed Correlation Filters for Real-time UAV Tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2891–2900.
- [31] Y. Li, C. Fu, Z. Huang, Y. Zhang, and J. Pan, "Keyfilter-aware real-time uav object tracking," in *Proceedings of IEEE International Conference on Robotics and Automation*, 2020, pp. 1–8.
- [32] Y. Li, C. Fu, Z. Huang, Y. Zhang, and J. Pan, "Intermittent contextual learning for keyfilter-aware uav object tracking using deep convolutional feature," *IEEE Transactions on Multimedia*, 2020.
- [33] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 370–386.
- [34] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," in *Proceedings of European Conference on Computer Vision*. Springer, 2016, pp. 445–461.
- [35] S. Li and D.-Y. Yeung, "Visual object tracking for unmanned aerial vehicles: A and new motion models," in *Proceedings of Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 4140–4146.
- [36] A. Lukežič, T. Vojtík, L. C. Zajc, J. Matas, and M. Kristan, "Discriminative Correlation Filter with Channel and Spatial Reliability," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4847–4856.
- [37] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Adaptive Decontamination of the Training Set: A Unified Formulation for Discriminative Visual Tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1430–1438.
- [38] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 fps with deep regression networks," in *Proceedings of European Conference on Computer Vision*. Springer, 2016, pp. 749–765.
- [39] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for Correlation Filter based tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2805–2813.
- [40] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Y. Choi, "Action-Decision Networks for Visual Tracking with Deep Reinforcement Learning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1349–1358.
- [41] N. Wang, Y. Song, C. Ma, W. Zhou, W. Liu, and H. Li, "Unsupervised Deep Tracking," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1308–1317.
- [42] C. Wang, L. Zhang, L. Xie, and J. Yuan, "Kernel cross-correlator," in *Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 4179–4186.
- [43] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer, "Adaptive Color Attributes for Real-Time Visual Tracking," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1090–1097, 2014.
- [44] Y. Wu, J. Lim, and M. Yang, "Online Object Tracking: A Benchmark," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2411–2418.