Combining Local and Global Models for Robust Re-detection

Goutam Bhat, Martin Danelljan, Fahad Khan and Michael Felsberg

Conference article

Cite this conference article as:

Bhat, G., Danelljan, M., Khan, F., Felsberg, M. Combining Local and Global Models for Robust Re-detection, In Proceedings of AVSS 2018. 2018 IEEE International Conference on Advanced Video and Signal-based Surveillance, Auckland, New Zealand, 27-30 November 2018, Institute of Electrical and Electronics Engineers (IEEE); 2018, pp. 25-30. ISBN: 978-1-5386-9294-3 (online), 978-1-5386-9293-6 (online), 978-1-5386-9295-0 (Print-on-demand).

DOI: <https://doi.org/10.1109/AVSS.2018.8639159>

Copyright: Institute of Electrical and Electronics Engineers (IEEE)

The self-archived postprint version of this conference article is available at Linköping University Institutional Repository (DiVA):

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-158403>

Combining Local and Global Models for Robust Re-detection

Goutam Bhat¹, Martin Danelljan¹, Fahad Shahbaz Khan^{1,2}, Michael Felsberg¹

¹Computer Vision Laboratory, Department of Electrical Engineering, Linköping University, Sweden

²Inception Institute of Artificial Intelligence, Abu Dhabi, UAE

{martin.danelljan, goutam.bhat, fahad.khan, michael.felsberg}@liu.se

Abstract

Discriminative Correlation Filters (DCF) have demonstrated excellent performance for visual tracking. However, these methods still struggle in occlusion and out-of-view scenarios due to the absence of a re-detection component. While such a component requires global knowledge of the scene to ensure robust re-detection of the target, the standard DCF is only trained on the local target neighborhood.

In this paper, we augment the state-of-the-art DCF tracking framework with a re-detection component based on a global appearance model. First, we introduce a tracking confidence measure to detect target loss. Next, we propose a hard negative mining strategy to extract background distractors samples, used for training the global model. Finally, we propose a robust re-detection strategy that combines the global and local appearance model predictions. We perform comprehensive experiments on the challenging UAV123 and LTB35 datasets. Our approach shows consistent improvements over the baseline tracker, setting a new state-of-the-art on both datasets.

1. Introduction

Generic object tracking refers to the problem of estimating the trajectory of a target, given its location in the first frame. It is one of the fundamental problems in computer vision with numerous applications in *e.g.* surveillance, robotics, and autonomous driving. Object tracking is particularly challenging due to limited availability of training data. Furthermore, the target can undergo significant appearance changes caused by *e.g.* deformation, blur, and illumination variation. Therefore, the tracker must learn a robust appearance model that can generalize to such appearance changes from only a single annotated example.

In recent years, Discriminative Correlation Filters (DCF)-based tracking approaches have shown excellent performance on benchmark tracking datasets [1, 5, 8, 15]. The DCF-based tracking methods train a correlation filter on examples containing all circular shifts of the target re-

Figure 1: A comparison of our approach with the baseline ECO tracker [5] on three example sequences from the UAV123 dataset. ECO fails to track when the target is occluded (top and middle row) or goes out of frame (bottom row) due to the absence of a re-detection component. Instead, our approach is able to detect tracking failures and performs successful re-detection.

gion [3]. Previously, hand-crafted features were employed for image representation within DCF trackers [1, 6, 9]. However, recently major performance gains have been achieved with the integration of deep features [5, 8, 18]. While DCF-based trackers have demonstrated robustness to complex appearance changes, they typically struggle when the target object is not visible for longer durations (figure 1). This is due to the absence of explicit failure detection and re-detection components. Such capabilities are of critical importance in many tracking applications, including UAV based surveillance where the target may frequently go out-of-frame due to camera motion or undergo long occlusions.

In recent years, several attempts have been made to integrate re-detection capabilities within the DCF based methods. Ma *et al.* [19] train a random ferns classifier online to re-detect the target in case of tracking failure. Hong *et al.* [13] employ keypoint matching to handle target loss. Fan *et al.* [10] propose a parallel tracking and verification frame-

work, in which a deep network verifies the output of the tracker and corrects it in case of failure. Recently, Lukezic *et al.* [16] employ a DCF for re-detection by padding the filter with zeros and applying it on the full image. Their approach exploits the appearance model that is trained for localization, when performing re-detection. However, such an approach is likely to struggle when distinguishing the target from background distractors, since the appearance model is trained only using the *local* neighborhood of the target. These cases occur frequently when tracking non-unique object types, such as humans or vehicles. Here, we argue that the appearance model needs to incorporate the existence of such distractors to enable robust re-detection capabilities.

A straightforward procedure of handling distractors is to train the target appearance model *globally* on the entire image. However, such a naive model training strategy is cumbersome and ineffective, since most of the background content typically corresponds to easy training samples. Thus, only the hard negative training samples, classified incorrectly by the original local model, are of interest. This approach, generally referred to as "hard negative mining", is often employed in the field of object detection as a means of robustifying the classifier [4, 11]. While similar strategies have been employed in a few non-DCF tracking methods [21, 25], it remains largely unexplored in visual tracking. Instead, the vast majority of tracking models rely on local information. This motivates us to investigate the use of hard negative mining in state-of-the-art DCF trackers to enable robust re-detection.

In this paper, we propose an approach to incorporate re-detection capabilities in a state-of-the-art DCF tracking framework. Our approach aims to combine both the local appearance model, trained for target localization, as well as a global appearance model, trained to handle background distractors, for robust re-detection. First, we introduce a tracking confidence measure to detect target loss. Next, we propose a hard negative mining strategy to extract background distractors used for training a global appearance model. Finally, we propose a re-detection strategy which uses the predictions of both the local as well as the global model to perform robust re-detection. We perform comprehensive experiments on the challenging UAV123 [20] and LTB35 [17] datasets. Our approach shows consistent improvements over the baseline tracker, setting a new state-of-the-art on both datasets.

2. Baseline Tracker

In this work, we employ the recently introduced ECO tracker [5] as our baseline. The ECO tracker has demonstrated excellent results on several tracking benchmarks [14]. It learns a continuous convolution filter f and a projection matrix P based on a set of sample feature maps $\{x_j\}_1^M$ and their corresponding desired convolution responses $\{y_j\}_1^M$.

The target location is then predicted by the score operator

$$S_{f,P}\{x\} = \sum_{c=1}^C f^c * \left(\sum_{d=1}^D P_{dc} J_d \{x^d\} \right). \quad (1)$$

Here, the d -th channel x^d of the input feature map x is first interpolated to the continuous domain using the operator J_d . The sample is then mapped to a lower dimensional space using the $D \times C$ matrix P . The projected sample is finally convolved with the learnt filter f to get the detection score.

The filter f and the projection matrix P are both learnt jointly by minimizing the following objective,

$$E(f) = \sum_{j=1}^M \alpha_j \|S_{f,P}\{x_j\} - y_j\|^2 + \sum_{c=1}^C \|w f^c\|^2 + \lambda \|P\|_F^2. \quad (2)$$

The second term represents a spatial regularization controlled by the weight function w , needed to alleviate the boundary effects induced by circular convolution [7]. The last term regularizes the matrix P using the Frobenius norm. The individual training samples are weighted by α_j . The label function y_j is set to a Gaussian centered at the target location in the corresponding sample x_j .

Using Parseval's formula, the objective (2) is transformed to the Fourier domain to achieve the equivalent loss,

$$E(f) = \sum_{j=1}^M \alpha_j \|\widehat{S_{f,P}\{x_j\}} - \hat{y}_j\|^2 + \sum_{c=1}^C \|\hat{w} * \hat{f}^c\|^2 + \lambda \|P\|_F^2. \quad (3)$$

Here $\hat{\cdot}$ denotes the Fourier coefficients. The resulting normal equations are solved efficiently using the conjugate gradient method. For more details, we refer to [8, 5].

In this work, we further adopt the recommendations proposed in [2] by training two independent tracking models. One is based on features extracted from a very deep network [12], while the other employs low-level hand-crafted features [4, 22]. The deep feature based model is trained using extensive data augmentation and a wide label function y_j to achieve high robustness, while the shallow feature based model is trained with a narrow label function to obtain high accuracy. The scores obtained from these models are fused using a maximum margin based formulation [2] to combine the advantages of both models.

3. Our Approach

In this section we describe our tracking approach. We use the baseline DCF tracker as the *local* appearance model. It is trained using only a local neighborhood around the target, and therefore designed primarily for accurate target localization. We complement the local model with a DCF-based *global* appearance model, that is trained on entire image frames. The purpose of the global model is to distinguish the target object from distractor objects, not correctly

Figure 2: An illustration of the hard negative mining strategy used to extract distractor samples. Left: Image used to perform hard negative mining. Middle: The response of the local model on the input image. The background regions with high score are identified as distractors (drawn as red boxes in the left image). Right: Response of the global model trained using the hard negative samples. We see that the global model gives high score only at the target location.

classified by the locally trained model. The training data for the global model is extracted efficiently using the “hard negative mining” strategy, as described in section 3.2.

In case of target loss, we apply both the global and local models on the full image to robustly and accurately re-detect the target object, as detailed in section 3.3. We also tackle the problem of detecting a target loss by introducing a tracking confidence measure in section 3.1. It is used in every frame to determine if the target is present or lost, due to *e.g.* an occlusions. If the target is not present, the re-detection procedure is invoked in every frame until the target is localized again. Note that in the proposed re-detection procedure as well as the confidence measure, we only use the deep feature based model due to its high robustness

3.1. Tracking Confidence

In this section, we propose a measure for determining loss of the tracked target, which is the first step in designing a re-detection strategy. This is a challenging task in the generic tracking scenario since the model is updated dynamically using predictions of the tracker. If the tracking output is wrong, training samples collected from the corresponding frame will be incorrectly annotated, leading to a risk of model drift. If hard negative mining is performed, the actual target might be considered a negative training sample and added to the distractor set, further deteriorating the discriminative power of the model. Therefore, a robust measure indicating the certainty of the current prediction is crucial. This confidence measure can then be used to filter out incorrect samples and determine tracking failures.

In the ECO tracking framework, the maximum value of the score function 1, denoted r_p , can be used as a measure of tracking confidence. However, the magnitude of r_p can vary significantly between different tracking sequences. For instance, if the target is moving slowly and is easily distinguishable, the maximum score r_p is generally high. However, if the target is undergoing fast motion with significant deformations, the value of r_p will be low even if the tracking is correct. Thus, the value of r_p needs to be normalized in

order to incorporate the difficulty of the tracking sequence. We use these insights to define a normalized tracking confidence t_{conf} as,

$$t_{\text{conf}} = \frac{r_p}{\bar{r}_p} \quad \bar{r}_p = \eta r_p + (1 - \eta) \bar{r}_{p, \text{prev}} \quad (4)$$

Here, $\bar{r}_p$ is the running average of r_p , capturing the general magnitude of the maximum score. If the tracking is successful, the value of t_{conf} is high (≈ 1). However, if the target is occluded or out of frame, the value of t_{conf} drops sharply, indicating a tracking failure. In order to prevent the corruption of the training set, we do not update the model, or perform hard negative mining when the confidence t_{conf} is less than a threshold τ_{min} . Moreover, if this condition remains true for a certain number of consecutive frames (N_{loss}), we declare the target lost and enter the re-detection mode.

3.2. Hard Negative Mining

Generally, DCF based trackers train the appearance model based only on samples from local target neighborhood. This is sufficient in scenarios where the target motion is moderate and occlusions are brief. In practice however, the target can remain occluded or out-of-frame for substantial durations, making such motion priors ineffective. In these cases, it is necessary to search the full image in order to robustly re-detect the target. It is therefore crucial to learn a global appearance model, based on the entire image, in order to perform re-detection. However, using entire frames as training samples is computationally cumbersome. In fact, such a strategy is unnecessary. Instead, we only need to add the background regions that are incorrectly classified by the local tracking model. We term such regions *hard negatives*.

In our approach, we perform hard negative mining every N_{hn} frame by collecting samples of distractor objects not correctly classified by the tracker. To this end, we apply the local appearance model on the full image and select background regions with tracker confidence t_{conf} greater than a threshold τ_{hn} (see figure 2). Furthermore, if a target loss is detected as described in the previous section, we perform an additional round of hard negative mining in the most recent frame that the target was confidently tracked (*i.e.* when $t_{\text{conf}} > \tau_{\text{min}}$). The purpose is to find the most recent distractors in view before the target was lost. As detailed in the next section, we then train the global appearance model based on the collected hard negative examples.

3.3. Re-detection

In this section, we describe the re-detection mechanism used in case of tracking failure. Whenever a target loss is detected, the DCF-based global appearance model is first trained to handle background distractors. This is done by minimizing the spatially regularized DCF loss 2. We use two types of training samples $\{x_j\}_j$ to learn the global

Figure 3: An illustration of the different behaviour of the local model and the global model on two example frames. The target is occluded in the top row, but is clearly visible in the bottom row. Left: Input image. Middle: Response of local model. Right: Response of the global model. The local model, while giving a high score at the correct target location, also gives high scores at several background regions. The global model on the other hand can correctly distinguish the target from background distractors.

model: (i) samples centered at the target location and (ii) samples extracted using the hard negative mining procedure described in section 3.2. In the former case (i), we employ a standard Gaussian label function y_j centered at the target, with the aim of producing a distinct positive score at the target. For the hard negative samples (ii), a negatively scaled Gaussian function is used. The negative peak of y_j is centered at the distractor object itself to emphasize the suppression of such samples during training. Note that the local appearance model is trained using only samples of the first type (i). The global model is trained by first initializing it with the local model and then performing a fixed number of conjugate gradient iterations to minimize the loss 2.

To re-detect the target in a frame, we first generate predictions of both the global and local model over the entire image. These models have complementary advantages to be exploited during re-detection. The local model generally localizes the target more accurately, while the global model is designed for high target detection precision when applied on the entire image (see figure 3). We therefore extract candidate target locations by requiring both the local and global detection scores to be larger than a threshold $\bar{r}_p * \tau_{\text{redetection}}$. We then aim to find the most accurate can-

didate location by selecting the one with the highest local model score as our re-detected target. If no location satisfies the above mentioned criteria, then the target is declared to not be visible in the corresponding frame. Re-detection is attempted every $N_{\text{redetection}}$ frames until the target is successfully re-detected. Once the target is re-detected, we resume conventional tracking using only the local model.

4. Experiments

4.1. Implementation Details

Our tracker is implemented in MATLAB using MatConvNet [23]. We use the activations from the fourth convolutional block of ResNet-50 [12] as our deep features, and the Histogram of Oriented Gradients (HOG) [4] and Color Names [22] as our shallow features. The learning rate η in estimation of the target confidence $\bar{r}_p$ in (4) is set to 0.05. When determining if the tracking output is reliable (section 3.1), the threshold τ_{min} is set to 0.6. For the hard negative mining strategy presented in section 3.2, we use a period $N_{\text{hn}} = 15$, while the threshold τ_{hn} is set to 0.3. For the re-detection mechanism (section 3.3), we use a period $N_{\text{redetection}} = 15$ and a threshold $\tau_{\text{redetection}} = 0.7$. Note that all parameters are kept fixed for all videos and all datasets.

4.2. Baseline Comparison

Here, we analyze the impact of our contributions on the aerial video tracking benchmark, UAV123 [20]. The UAV123 dataset consists of 123 aerial videos with more than 110K frames. Further, since tracking humans is especially important in surveillance applications, the performance of the tracker on human videos is interesting in itself. Thus we extract all the videos from UAV123 where the target is a human, here called the UAV-Human dataset, consisting of 45 videos. Table 1 shows the results of our analysis on full UAV123 as well as UAV-Human in terms of area-under-the-curve (AUC) score [24]. The baseline tracker with no re-detection component achieves as AUC score of 55.5 on UAV123 and 62 on UAV-Human. Extending the baseline tracker with the proposed failure detection component and performing a naive re-detection using only the learnt local model improves the results by 2.5% on UAV123 and 5.4% on UAV-Human. Using the proposed hard negative mining strategy to handle background distractors further improves the performance by 0.9% on UAV123 and 1.8% on UAV-Human. Note that the improvement obtained by using hard negative mining is much larger on the UAV-Human set, since these videos often contain other distractor humans in the background which can confuse the naive re-detection method. This results validate our hypothesis that using only the local model to perform re-detection is insufficient and that a global model trained using the distractor samples is necessary to perform robust re-detection.

	Baseline	+Re-detection	+Hard Negative Mining
UAV123	55.5	58.0	58.9
UAV-Human	62.0	67.4	69.2

Table 1: Analysis of our approach on UAV123 and UAV-Human datasets. The performance is reported in terms of area-under-the-curve (AUC) score. Introducing an explicit re-detection component provides a significant boost in tracking performance, compared to the baseline tracker. The results are further improved by our hard-negative mining strategy.

Figure 4: Qualitative results showing the importance of using a global model for re-detection on two example sequences. The baseline method lacks a failure detection components, and hence drifts into the background when the target is occluded (second column). The naive re-detection strategy using only the local model, while correctly determining target loss, fails to distinguish the target from background distractors (fourth column). Our approach on the other hand augments the local model with a global model trained to handle these distractors, and hence can successfully re-detect the target.

Figure 4 provides some qualitative examples showing the importance of using a global model. In the top row, since there is another person in the background wearing a white shirt, similar to the tracked person, the naive re-detection mechanism fails to re-detect correctly. Meanwhile, in the second row, the naive re-detection falsely classifies a background distractor as the target. Our approach on the other hand is able to track successfully in both cases as the distractor samples are added to the training set due to hard negative mining. These results show that the global filter can learn to better distinguish the target from these distractors.

4.3. State-of-the-art Comparison

Here, we compare our approach with state-of-the-art tracking methods on the UAV123 dataset and the recently released Long-term dataset (LTB35) [17]. LTB35 consists of 35 challenging videos with the target frequently going out of frame, or undergoing long occlusions. The trackers are ranked using the F-score, computed from the average precision-recall plot. We refer to [17] for more details.

Figure 5 shows the comparison of different tracking methods, on the UAV123 dataset, using precision and success plots [24], illustrating the mean distance and overlap precision. We report the average DP score at 20 pixels for each tracker in the legend of the precision plot. In the legend of the success plot, tracking methods are ranked using the area-under-the-curve (AUC) score. Our approach outperforms the best existing method (UDT [2]) in both precision and success plots, demonstrating the superior robustness *and* accuracy of our tracker. Table 2 shows state-of-the-art comparison, in terms of F-score, on the LTB35

Figure 5: State-of-the-art comparison, using precision and success plots, on the UAV123 dataset. Our approach significantly outperforms existing methods in both cases.

	BACF	PTAV	CREST	CSRDCF	ECOhc	ECO	SiamFC	MDNet	FCLT	Ours
F-Score	0.31	0.31	0.33	0.33	0.33	0.35	0.4	0.41	0.48	0.52

Table 2: State-of-the-art comparison on the Long-term benchmark (LTB35). The tracking performance is reported in terms of F-score. The proposed approach achieves the best performance with a F-score of 0.52

dataset. Among the existing methods, FCLT approach [16] that has a re-detection component achieves an F-score of 0.48. Our approach, significantly outperforms the FCLT approach, achieving an F-score of 0.52.

4.4. Attribute based analysis

In UAV123, each video is annotated with attributes which denote the presence of a certain tracking challenge, *e.g.* scale variation, fast motion, in the given sequence. We use this attribute annotation to evaluate our method on the subset of videos where target disappearance is one of the

Figure 6: Attribute-based analysis of our approach, using success plots, on UAV123 dataset. Results are shown only for the attributes relevant for re-detection, namely full occlusion (left) and out-of-view (right). The title of each plot shows the name of the attribute, as well as the number of videos labelled with that attribute. Our approach achieves the best results for both the attributes.

main challenges. Figure 6 shows the success plots for the attributes full occlusion and out-of-view. Our approach provides significant improvement in both the cases, showing the effectiveness of the proposed re-detection mechanism.

5. Conclusion

We proposed a re-detection approach in the state-of-the-art DCF tracking framework. A confidence measure is introduced to detect the target loss in the standard DCF tracker. We then introduced a hard negative mining scheme to extract background distractors and proposed a strategy to train a global filter using these hard negative samples in order to perform robust re-detection. Our approach is comprehensively evaluated on the challenging UAV123 and LTB35 datasets. The results showed that a consistent improvement is obtained by the proposed approach over the baseline, leading to state-of-the-art performance on both datasets.

Acknowledgments: This work was supported by SSF (SymbiCloud), VR (EMC², starting grant 2016-05543), CENIIT grant (18.14), SNIC, and WASP.

References

- [1] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr. Staple: Complementary learners for real-time tracking. In *CVPR*, 2016.
- [2] G. Bhat, J. Johnander, M. Danelljan, F. S. Khan, and M. Felsberg. Unveiling the power of deep tracking. In *ECCV*, 2018.
- [3] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, 2010.
- [4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
- [5] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg. ECO: efficient convolution operators for tracking. In *CVPR*, 2017.
- [6] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Discriminative scale space tracking. *TPAMI*, 39(8):1561–1575, 2017.
- [7] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg. Learning spatially regularized correlation filters for visual tracking. In *ICCV*, 2015.
- [8] M. Danelljan, A. Robinson, F. Khan, and M. Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *ECCV*, 2016.
- [9] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. van de Weijer. Adaptive color attributes for real-time visual tracking. In *CVPR*, 2014.
- [10] H. Fan and H. Ling. Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. *CoRR*, abs/1708.00153, 2017.
- [11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [13] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In *CVPR*, 2015.
- [14] M. Kristan, A. Leonardis, J. Matas, R. Felsberg, Pflugfelder, M., L. Čehovin, G. Vojír, T. and Hger, and et al. The visual object tracking vot2017 challenge results. In *ICCV workshop*, 2017.
- [15] A. Lukezic, T. Vojír, L. C. Zajc, J. Matas, and M. Kristan. Discriminative correlation filter tracker with channel and spatial reliability. *IJCV*, 126(7):671–688, 2018.
- [16] A. Lukezic, L. C. Zajc, T. Vojír, J. Matas, and M. Kristan. FCLT - A fully-correlational long-term tracker. *CoRR*, abs/1711.09594, 2017.
- [17] A. Lukezic, L. C. Zajc, T. Vojír, J. Matas, and M. Kristan. Now you see me: evaluating performance in long-term visual tracking. *CoRR*, abs/1804.07056, 2018.
- [18] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang. Hierarchical convolutional features for visual tracking. In *ICCV*, 2015.
- [19] C. Ma, X. Yang, C. Zhang, and M.-H. Yang. Long-term correlation tracking. In *CVPR*, 2015.
- [20] M. Mueller, N. Smith, and B. Ghanem. A benchmark and simulator for uav tracking. In *ECCV*, 2016.
- [21] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, 2016.
- [22] J. van de Weijer, C. Schmid, J. J. Verbeek, and D. Larlus. Learning color names for real-world applications. *TIP*, 18(7):1512–1524, 2009.
- [23] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. *CoRR*, abs/1412.4564, 2014.
- [24] Y. Wu, J. Lim, and M.-H. Yang. Object tracking benchmark. *TPAMI*, 37(9):1834–1848, 2015.
- [25] G. Zhu, F. Porikli, and H. Li. Not all negatives are equal: Learning to track with multiple background clusters. *IEEE Trans. Circuits Syst. Video Techn.*, 28(2):314–326, 2018.