D3S – A Discriminative Single Shot Segmentation Tracker

Alan Lukežič[1], Jiří Matas[2], Matej Kristan[1]

[1]Faculty of Computer and Information Science, University of Ljubljana, Slovenia

[2]Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

alan.lukezic@fri.uni-lj.si

Abstract

Template-based discriminative trackers are currently the dominant tracking paradigm due to their robustness, but are restricted to bounding box tracking and a limited range of transformation models, which reduces their localization accuracy. We propose a discriminative single-shot segmentation tracker – D3S, which narrows the gap between visual object tracking and video object segmentation. A single-shot network applies two target models with complementary geometric properties, one invariant to a broad range of transformations, including non-rigid deformations, the other assuming a rigid object to simultaneously achieve high robustness and online target segmentation. Without per-dataset finetuning and trained only for segmentation as the primary output, D3S outperforms all trackers on VOT2016, VOT2018 and GOT-10k benchmarks and performs close to the state-of-the-art trackers on the TrackingNet. D3S outperforms the leading segmentation tracker SiamMask on video object segmentation benchmarks and performs on par with top video object segmentation algorithms, while running an order of magnitude faster, close to real-time.

1. Introduction

Visual object tracking is one of core computer vision problems. The most common formulation considers the task of reporting target location in each frame of the video given a single training image. Currently, the dominant tracking paradigm, performing best in evaluations [22, 24], is correlation bounding box tracking [11, 3, 33, 2, 54, 28] where the target represented by a multi-channel rectangular template is localized by cross-correlation between the template and a search region.

State-of-the-art template-based trackers apply an efficient brute-force search for target localization. Such strategy is appropriate for low-dimensional transformations like translation and scale change, but becomes inefficient for more general situations e.g. such that induce an aspect ratio change and rotation. As a compromise, modern trackers combine approximate exhaustive search with sampling
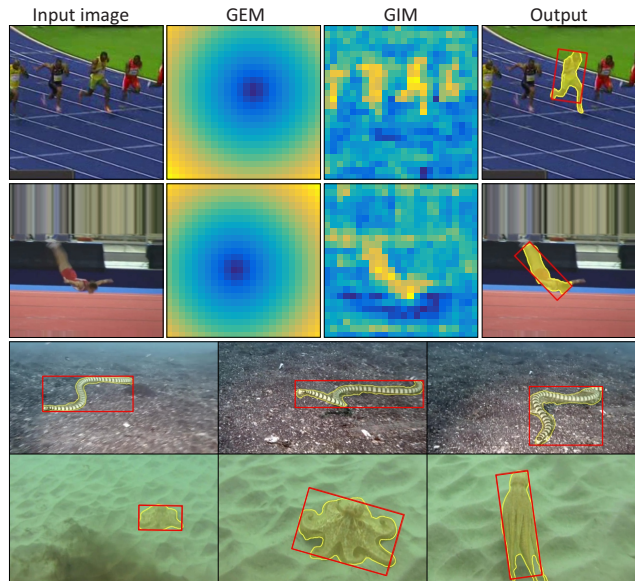

Figure 1. The D3S tracker represents the target by two models with complementary geometric properties, one invariant to a broad range of transformations, including non-rigid deformations (GIM - geometrically invariant model), the other assuming a rigid object with motion well approximated by an euclidean transformation (GEM - geometrically constrained Euclidean model). The D3S, exploiting the complementary strengths of GIM and GEM, provides both state-of-the-art localisation and accurate segmentation, even in the presence of substantial deformation.

and/or bounding box refinement/regression networks [10, 27] for aspect ratio estimation. However, these approaches are restricted to axis-aligned rectangles.

Estimation of high-dimensional template-based transformation is unreliable when a bounding box is a poor approximation of the target [31]. This is common – consider e.g. elongated, rotating, deformable objects, or a person with spread out hands. In these cases, the most accurate and well-defined target location model is a binary per-pixel segmentation mask. If such output is required, tracking becomes the video object segmentation task recently popularized by DAVIS [38, 40] and YoutubeVOS [51] challenges.

Unlike in tracking, video object segmentation challenges typically consider large targets observed for less than 100

frames with low background distractor presence. Top video object segmentation approaches thus fare poorly in short-term tracking scenarios [24] where the target covers a fraction of the image, substantially changes its appearance over a longer period and moves over a cluttered background. Best trackers apply visual model adaptation, but in the case of segmentation errors it leads to an irrecoverable tracking failure [41]. Because of this, in the past, segmentation has played only an auxiliary role in template-based trackers [1], constrained DCF learning [33] and tracking by 3D model construction [20].

Recently, the SiamRPN [28] tracker has been extended to produce high-quality segmentation masks in two stages [50] – the target bounding box is first localized by SiamRPN branches and then a segmentation mask is computed only within this region by another branch. The two-stage processing misses the opportunity to treat localization and segmentation jointly to increase robustness. Another drawback is that a fixed template is used that cannot be discriminatively adapted to the changing scene.

We propose a new single-shot discriminative segmentation tracker, D3S, that addresses the above-mentioned limitations. The target is encoded by two discriminative visual models – one is adaptive and highly discriminative, but geometrically constrained to an Euclidean motion (GEM), while the other is invariant to broad range of transformation (GIM, geometrically invariant model), see Figure 1.

GIM sacrifices spatial relations to allow target localization under significant deformation. On the other hand, GEM predicts only position, but discriminatively adapts to the target and acts as a selector between possibly multiple target segmentations inferred by GIM. In contrast to related trackers [50, 27, 10], the primary output of D3S is a segmentation map computed in a single pass through the network, which is trained end-to-end for segmentation only (Figure 2).

Some applications and most tracking benchmarks require reporting the target location as a bounding box. As a secondary contribution, we propose an effective method for interpreting the segmentation mask as a rotated rectangle. This avoids an error-prone greedy search and naturally addresses changes in location, scale, aspect ratio and rotation.

D3S outperforms all state-of-the-art trackers on most of the major tracking benchmarks [23, 24, 19, 35] despite not being trained for bounding box tracking. In video object segmentation benchmarks [38, 40], D3S outperforms the leading segmentation tracker [50] and performs on par with top video object segmentation algorithms (often tuned to a specific domain), yet running orders of magnitude faster. Note that D3S is not re-trained for different benchmarks – a single pre-trained version shows remarkable generalization ability and versatility[1].

---

[1]PyTorch implementation will be made available.

2. Related Work

Robust localization crucially depends on the discrimination capability between the target and the background distractors. This property has been studied in depth in discriminative template trackers called discriminative correlation filters (DCF) [4]. The template learning is formulated as a (possibly nonlinear) ridge regression problem and solved by circular correlation [4, 12, 17, 30]. While trackers based purely on color segmentation [8, 41] are inferior to DCFs, segmentation has been used for improved DCF tracking of non-rectangular targets [1, 31]. Lukežič et al. [33] used color segmentation to constrain DCF learning and proposed a real-time tracker with hand-crafted features which achieved performance comparable to trackers with deep features. The method was extended to long-term [32] and RGB-depth tracking [20] using color and depth segmentation. Further improvements in DCF tracking considered deep features: Danelljan et al. [11] used features pre-trained for detection, Valmadre et al. [46] proposed pre-training features for DCF localization and recently Danelljan et al. [10] proposed a deep DCF training using backpropagation.

Another class of trackers, called Siamese trackers [2, 44, 15], has evolved in direction of generative templates. Siamese trackers apply a backbone pre-trained offline with general targets such that object-background discrimination is maximized by correlation between the search region and target template extracted in the first frame [2]. The template and the backbone are fixed during tracking, leading to an excellent real-time performance [24]. Several multi-stage Siamese extensions have been proposed. These include addition of region proposal networks for improved target localization accuracy [28, 27] and addition of segmentation branches [50] for accurate target segmentation. Recently a template adaptation technique by backprop has been proposed [29] to improve tracking robustness.

Segmentation of moving objects is a central problem in the emerging field of video object segmentation (VOS) [38, 51]. Most recent works [47, 5, 48, 7, 53] achieve impressive results, but involve large deep networks, which often require finetuning and are slow. Hu et al. [18] and Chen et al. [6] concurrently proposed segmentation by matching features extracted in the first frame, which considerably reduces the processing time. However, the VOS task considers segmentation of large objects with limited appearance changes in short videos. Thus, these methods fare poorly on the visual object tracking task with small, fast moving objects. The work proposed in this paper aims at narrowing the gap between visual object tracking and video object segmentation.
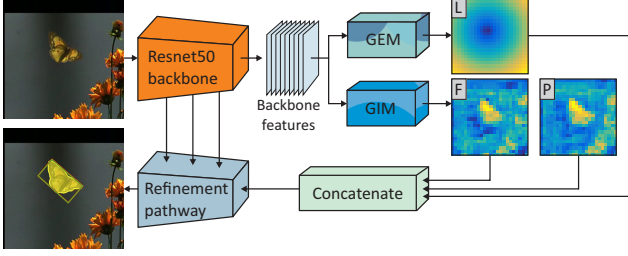
Figure 2. The D3S segmentation architecture. The backbone features are processed by the GEM and GIM pathways, producing the target location ($\mathbf{L}$), foreground similarity ($\mathbf{F}$) and target posterior ($\mathbf{P}$) channels. The output of the three channels are concatenated and refined into a detailed segmentation map.
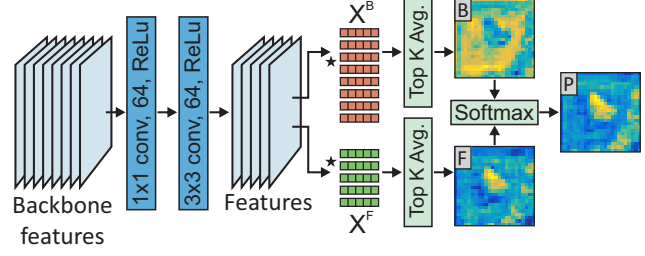

Figure 3. GIM – the geometrically invariant model – features are matched to the features in the foreground-background model $\{\mathbf{X}^F, \mathbf{X}^B\}$ to obtain the target ($\mathbf{F}$) and background ($\mathbf{B}$) similarity channels. The posterior channel ($\mathbf{P}$) is the softmax of $\mathbf{F}$ and $\mathbf{B}$.

3. Discriminative segmentation network

Two models are used in D3S to robustly cope with target appearance changes and background discrimination: a geometrically invariant model (GIM) presented in Section 3.1, and a geometrically constrained Euclidean model (GEM) presented in Section 3.2. These models process the input in parallel pathways and produce several coarse target presence channels, which are fused into a detailed segmentation map by a refinement pathway described in Section 3.3. See Figure 2 for the architecture outline.

3.1. Geometrically invariant model pathway

Accurate segmentation of a deformable target requires loose spatial constraints in the discriminative model. Our geometrically invariant model (GIM) is thus composed of two sets of deep feature vectors corresponding to the target and the background, i.e., $\mathbf{X}_{\mathrm{GIM}} = \{\mathbf{X}^F, \mathbf{X}^B\}$.

Since the pre-trained backbone features are sub-optimal for accurate segmentation, these are first processed by a 1×1 convolutional layer to reduce their dimensionality to 64, which is followed by a 3×3 convolutional layer (a ReLU is placed after each convolutional layer). Both these layers are adjusted in the network training stage to produce optimal features for segmentation. The target/background models are created in the first frame by extracting the segmentation feature vectors at pixel locations corresponding to the target ($\mathbf{X}^F$) and from the immediate neighborhood for the background ($\mathbf{X}^B$).

During tracking, the pixel-level features extracted from the search region are compared to those of GIM ($\mathbf{X}_{\mathrm{GIM}}$) to compute foreground and background similarity channels $\mathbf{F}$ and $\mathbf{B}$ following [18]. Specifically, for the $\mathbf{F}$ channel computation, each feature $\mathbf{y}_i$ extracted at pixel i is compared to all features $\mathbf{x}_j^F \in \mathbf{X}^F$ by a normalized dot product

$$s_{ij}^F(\mathbf{y}_i, \mathbf{x}_j^F) = \langle \tilde{\mathbf{y}}_i, \tilde{\mathbf{x}}_j^F \rangle, \tag{1}$$

where $\tilde{(\cdot)}$ indicates an L_2 normalization. The final per-pixel

foreground similarity at pixel i, $\mathbf{F}_i$, is obtained by average of top-K similarities at that pixel, i.e.,

$$\mathbf{F}_i = \mathrm{TOP}(\{s_{ij}^F\}_{j=1:N_F}, K), \tag{2}$$

where $\mathrm{TOP}(\cdot, K)$ is a top-K averaging operator over the set of N_F similarities. Computation of the background similarity channel $\mathbf{B}$ follows the same principle, but with similarities computed with the background model feature vectors, i.e., $\mathbf{x}_j^B \in \mathbf{X}^B$. Finally, a softmax layer is applied to produce a target posterior channel $\mathbf{P}$. The GIM pathway architecture is shown in Figure 3.

3.2. Geometrically constrained model pathway

While GIM produces an excellent target-background separation, it cannot well distinguish the target from similar instances, leading to a reduced robustness (see Figure 1, first line). Robust localization, however, is a well-established quality of the discriminative correlation filters. Although they represent the target by a geometrically constrained model (i.e., a rectangular filter), efficient techniques developed to adapt to the target discriminative features [13, 33, 10] allow tracking reliably under considerable appearance changes.

We thus employ a recent deep DCF formulation [10] in the geometrically constrained Euclidean model (GEM) pathway. Following [10], the backbone features are first reduced to 64 channels by 1×1 convolutional layer. The reduced features are correlated by a 64 channel DCF followed by a PeLU nonlinearity [45]. The reduction layer and DCF are trained by an efficient backprop formulation (see [10] for details).

The maximum of the correlation response is considered as the most likely target position. The D3S output (i.e., segmentation), however, requires specifying a belief of target presence at each pixel. Therefore, a target location channel is constructed by computing a (Euclidean) distance transform from the position of the maximum in the correlation map to the remaining pixels in the search region. The GEM pathway is shown in Figure 4.
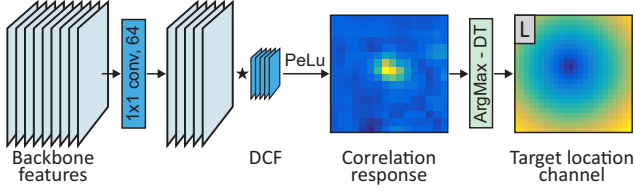
Figure 4. GEM – the geometrically constrained Euclidean model – reduces the backbone features dimensionality and correlates them with a DCF. The target localisation channel (**L**) is the distance transform to the maximum correlation response, representing the per-pixel confidence of target presence.

3.3. Refinement pathway

The GIM and GEM pathways provide complementary information about the pixel-level target presence. GEM provides a robust, but rather inaccurate estimate of the target region, whereas the output channels from GIM show a greater detail, but are less discriminative (Figure 1). Furthermore, the individual outputs are low-resolution due to the backbone encoding. A refinement pathway is thus designed to combine the different information channels and upscale the solution into an accurate and detailed segmentation map.

The refinement pathway takes the following inputs: the target location channel (**L**) from GEM and the foreground similarity and posterior channels (**F** and **P**) from the GIM. The channels are concatenated and processed by a 3×3 convolutional layer followed by a ReLU, resulting in a tensor of 64 channels. Three stages of upscaling akin to [42, 39] are then applied to refine the details by considering the features in different layers computed in the backbone. An upscaling stage consists of doubling the resolution of the input channels, followed by two 3×3 convolution layers (each followed by a ReLU). The resulting channels are summed with the adjusted features from the corresponding backbone layer. Specifically, the backbone features are adjusted for the upscaling task by a 3×3 convolution layer, followed by a ReLU. The last upscaling stage (which contains only resolution doubling, followed by a single 3×3 convolution layer) is followed by a softmax to produce the final segmentation probability map. The refinement pathway is shown in Figure 5.

4. Discriminative Segmentation Tracker

This section outlines application of the discriminative segmentation network from Section 3 to online general object tracking. Given a single supervised training example from the first frame, the network produces target segmentation masks in all the remaining frames. However, some applications and most tracking benchmarks require target location represented by a bounding box. For most benchmarks, the bounding box is trivially obtained by fitting an
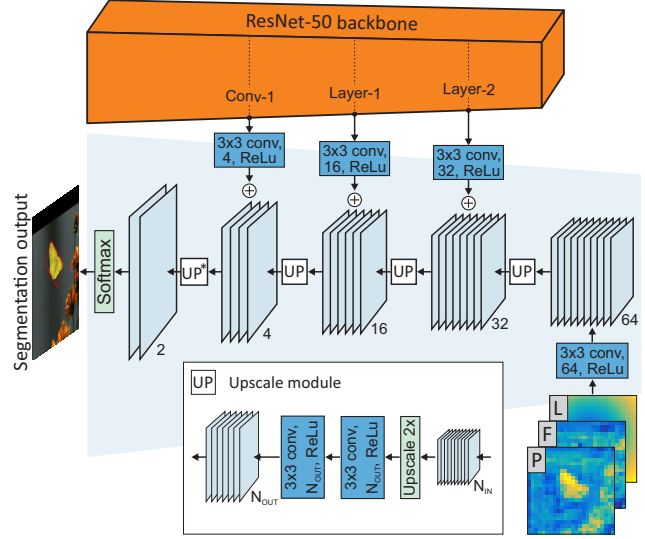

Figure 5. The refinement pathway combines the GIM and GEM channels and gradually upscales them by using adjusted features from the backbone. The UP* is a modified UP layer (see the text).

axis-aligned bounding box that tightly fits a segmentation mask. However, for the benchmark requiring a rotated bounding box, we propose a simple fitting procedure in Section 4.1. The tracking steps are outlined in Section 4.2.

4.1. Bounding box fitting module

The segmentation probability map from the discriminative segmentation network (Section 3) is thresholded at 0.5 probability to yield a binary segmentation mask. Only the largest connected component within the mask is kept and an ellipse is fitted to its outline by least squares [14]. The ellipse center, major and minor axis make up an initial estimate of the rotated bounding box. This is typically the most liberal solution with oversized rectangles, preferring most of the target pixels lying within its area, but accounts poorly for the presence of the background pixels within the region. We therefore further reduce the rectangle sides in direction of the major axes by optimizing the following modified overlap cost function IoU^{MOD} between the predicted segmentation mask and fitted rectangle using a coordinate descent:

$$IoU^{\mathrm{MOD}} = \frac{N_{\mathrm{IN}}^{+}}{\alpha N_{\mathrm{IN}}^{-} + N_{\mathrm{IN}}^{+} + N_{\mathrm{OUT}}^{+}}, \qquad (3)$$

where N_{IN}^{+} and N_{OUT}^{+} denote the number of foreground pixels within and outside the rectangle, respectively, and N_{IN}^{-} denotes the number of background pixels within the rectangle. The scalar α controls the contribution of N_{IN}^{-}. The bounding box fitting method is very fast and takes on average only 2ms.

4.2. Tracking with D3S

Initialization. D3S is initialized on the first frame using the ground truth target location. The GEM and GIM initialization details depend on whether the target ground truth is presented by a bounding box or a segmentation mask. If a ground truth bounding box is available, the GEM follows the initialization procedure proposed in [10], which involves training both the dimensionality reduction network and the DCF by backprop on the first frame by considering the region four times the target size. On the other hand, if a segmentation mask is available, the ground truth target bounding box is first approximated by an axis-aligned rectangle encompassing the segmented target.

In case a segmentation mask is available, the GIM is initialized by extracting foreground samples from the target mask and background samples from the neighborhood four times the target size. However, if only a bounding box is available, an approximate ground truth segmentation mask is constructed first. Foreground samples are extracted from within the bounding box, while the background samples are extracted from a four times larger neighborhood. A tracking iteration of D3S is then run on the initialization region to infer a *proxi* ground truth segmentation mask. The final foreground and background samples are extracted from this mask. This process might be iterated a few times (akin to GrabCut [43]), however, we did not observe improvements and chose only a single iteration for initialization speed and simplicity.

Tracking. During tracking, when a new frame arrives, a region four times the target size is extracted at previous target location. The region is processed by the discriminative segmentation network from Section 3 to produce the output segmentation mask. A rotated bounding box is fitted to the mask (Section 4.1) if required by the evaluation protocol. The DCF in the GEM is updated on the estimated target location following the backprop update procedure [10].

5. Experiments

5.1. Implementation details

The backbone network in D3S is composed of the first four layers of ResNet50, pre-trained on ImageNet for object classification. The backbone features are extracted from the target search region resized to 384×384 pixels. The background tradeoff parameter from (3) is set to $\alpha = 0.25$ and the top $K = 3$ similarities are used in GIM (2). We verified in a preliminary analysis that performance is insensitive to exact values of these parameters, and we therefore keep the same values in all experiments.

Network pre-training. The GIM pathway and the refinement pathway are pre-trained on 3471 training segmentation sequences from Youtube-VOS [51]. A training sample is constructed by uniformly sampling a pair of images

and the corresponding segmentation masks from the same sequence within a range of 50 frames. To increase the robustness to possibly inaccurate GEM localization, the target location channel was constructed by perturbing ground truth locations uniformly from $[-\frac{1}{8}\sigma, \frac{1}{8}\sigma]$, where σ is target size. The network was trained by 64 image pairs batches for 40 epochs with 1000 iterations per epoch using the ADAM optimizer [21] with learning rate set to 10^{-3} and with 0.2 decay every 15 epochs. The training loss was a crossentropy between the predicted and ground truth segmentation mask. The training takes 20 hours on a single GPU.

Speed. A Pytorch implementation of D3S runs at 25fps on a single NVidia GTX 1080 GPU, while 1.3s is required for loading the network to GPU and initialization.

5.2. Evaluation on Tracking Datasets

D3S was evaluated on four major short-term tracking datasets: VOT2016 [23], VOT2018 [24], GOT-10k [19] and TrackingNet [35]. In the following we discuss the results obtained on each of the datasets.

VOT2016 and **VOT2018** datasets each consist of 60 sequences. Targets are annotated by rotated rectangles to enable a more thorough localization accuracy evaluation compared to the related datasets. The standard VOT evaluation protocol [26] is used in which the tracker is reset upon tracking failure. Performance is measured by accuracy (average overlap over successfully tracked frames), robustness (failure rate) and the EAO (expected average overlap), which is a principled combination of the former two measures [25].

The following state-of-the-art (sota) trackers are considered on **VOT2016**: the VOT2016 top performers CCOT [13] and TCNN [36], a sota segmentation-based discriminative correlation filter CSR-DCF [33], and most recently published sota deep trackers SiamRPN [28], SPM [49], ASRCF [9], SiamMask [50] and ATOM [10].

Results reported in Table 1 show that D3S outperforms all tested trackers on all three measures by a large margin. In EAO measure, D3S outperforms the top sota tracker SPM by 14%, and simultaneously outperforms the top robust sota ATOM by 25% in robustness. The top sota performer in accuracy is the segmentation-based tracker SiamMask. D3S outperforms this tracker by over 3% in accuracy and approximately by 50% in robustness.

The VOT2016 dataset contains per-frame target segmentation masks which can be used to evaluate segmentation performance on the small and challenging targets present. We have thus compared D3S with the most recent segmentation tracker SiamMask by computing the average IoU between the ground truth and predicted segmentation masks during periods of successful tracks (i.e., segmentation accuracy). D3S achieves a 0.66 average IoU, while SiamMask IoU is 0.63. A nearly 5% improvement speaks of a considerable accuracy of the D3S segmentation mask prediction.

	D3S	SPM	SiamMask	ATOM	ASRCF	SiamRPN	CSRDCF	CCOT	TCNN
EAO ↑	① **0.493**	② 0.434	③ 0.433	0.430	0.391	0.344	0.338	0.331	0.325
Acc. ↑	① **0.66**	③ 0.62	② 0.64	0.61	0.56	0.56	0.51	0.54	0.55
Rob. ↓	① **0.131**	0.210	0.214	② 0.180	③ 0.187	0.302	0.238	0.238	0.268

Table 1. VOT2016 – comparison with state-of-the-art trackers.

	D3S	SiamRPN++	ATOM	LADCF	DaSiamRPN	SiamMask	SPM	ASRCF
EAO ↑	① **0.489**	② 0.414	③ 0.401	0.389	0.383	0.380	0.338	0.328
Acc. ↑	① **0.64**	③ 0.60	0.59	0.51	0.59	② 0.61	0.58	0.49
Rob. ↓	① **0.150**	0.234	③ 0.204	② 0.159	0.276	0.276	0.300	0.234

Table 2. VOT2018 – comparison with state-of-the-art trackers.

	D3S	ATOM	SiamMask	SiamFCv2	SiamFC	GOTURN	CCOT	MDNet
AO ↑	① **59.7**	② 55.6	③ 51.4	37.4	34.8	34.2	32.5	29.9
$SR_{0.75}$ ↑	① **46.2**	② 40.2	③ 36.6	14.4	9.8	12.4	10.7	9.9
$SR_{0.5}$ ↑	① **67.6**	② 63.5	③ 58.7	40.4	35.3	37.5	32.8	30.3

Table 3. GOT-10k test set – comparison with state-of-the-art trackers .

	D3S	SiamRPN++	SiamMask	ATOM	MDNet	CFNet	SiamFC	ECO
AUC ↑	② 72.8	① **73.3**	③ 72.5	70.3	60.6	57.8	57.1	55.4
Prec. ↑	② 66.4	① **69.4**	② 66.4	③ 64.8	56.5	53.3	53.3	49.2
$Prec._N$ ↑	76.8	① **80.0**	② 77.8	③ 77.1	70.5	65.4	66.3	61.8

Table 4. TrackingNet test set – comparison with state-of-the-art trackers.

On the **VOT2018** dataset, D3S is compared with the following sota trackers: the top VOT2018 performer LADCF [52] and the most recent sota trackers DaSiamRPN [54], SiamRPN++ [27], ATOM [10], SPM [49], ASRCF [9] and SiamMask [50]. Results are reported in Table 2. Again, D3S outperforms all sota trackers in all measures. The top sota trackers in EAO, accuracy and robustness are SiamRPN++, SiamMask and LADCF, respectively. D3S outperforms the SiamRPN++ in EAO by 18%, SiamMask in accuracy by over 5% and LADCF by over 6% in robustness. Note that SiamMask is a segmentation tracker, which explains the top accuracy among sota. D3S outperforms this tracker by over 45% in robustness, which is attributed to the discriminative formulation within the single-pass segmentation mask computation.

GOT-10k is a recent large-scale high-diversity dataset consisting of 10k video sequences with targets annotated by axis-aligned bounding boxes. The trackers are evaluated on 180 test sequences with 84 different object classes and 32 motion patterns, while the rest of the sequences form a training set. A tracker is initialized on the first frame and let to track to the end of the sequence. Trackers are ranked according to the average overlap, but success rates ($SR_{0.5}$ and $SR_{0.75}$) are reported at two overlap thresholds

0.5 and 0.75, respectively, for detailed analysis[2]. The following top-performing sota trackers are used in comparison [19]: SiamFCv2 [46], SiamFC [2], GOTURN [16], CCOT [13], MDNet [37] and the most-recent ATOM [10] and SiamMask [50]. We emphasize that D3S is not fine-tuned on the training set, while some of the top-performing sota trackers we use in comparison do utilize the GOT-10k training set. Results on GOT-10k are reported in Table 3. D3S outperforms all top-performing sota by a large margin in all performance measures, and achieves approximately 60% boost in average overlap compared to the SiamFCv2, which is a top-performer on [19] benchmark. It also outperforms the most recent ATOM and SiamMask trackers by over 7% and over 15% in average overlap, respectively. This demonstrates considerable generalization ability over a diverse set of target types.

TrackingNet is another large-scale dataset for training and testing trackers. The training set consists of over 30k video sequences, while the testing set contains 511 sequences. A tracker is initialized on the first frame and let to track to the end of the sequence. Trackers are ranked according to the area under the success rate curve (AUC), precision (Prec.) and normalized precision ($Prec._N$). The

---

[2]Success rate denotes percentage of frames where predicted region overlaps with the ground-truth region more than the threshold.

reader is referred to [35] for further details about the performance measures. The performance of D3S is compared with the top-performing sota trackers according to [35]: ECO [11], SiamFC [2], CFNet [46], MDNet [37] and most recent sota trackers ATOM [10], SiamMask [50] and SiamRPN++ [27]. D3S significantly outperforms the sota reported in [35] and is on par with SiamRPN++, SiamMask and ATOM. Note that D3S is trained only on 3471 sequences from YouTube-VOS [51], while both, ATOM and SiamRPN++ are finetuned on much larger datasets (31k, and over 380k sequences, respectively), which include the TrackingNet training set. This further supports a considerable generalization capability of D3S, which is primarily trained for segmentation, not tracking.

5.3. Ablation Study

An ablation study was performed on VOT2018 using the reset-based protocol [26] to expose the contributions of different components of the D3S architecture. The following variations of D3S were created: (i) D3S without the GIM foreground similarity channel $\mathbf{F}$ (D3S$^{\bar{F}}$); (ii) D3S without the GIM target posterior channel $\mathbf{P}$ (D3S$^{\bar{P}}$) ; (iii) D3S with only the GEM output channel and without GIM channels $\mathbf{F}$ and $\mathbf{P}$ (D3S$^{\bar{F}\bar{P}}$); (iv) D3S without the GEM output channel $\mathbf{L}$ (D3S$^{\bar{L}}$); (v) D3S in which the DCF is not updated from the position estimated by D3S, but rather from the position estimated by the DCF in GEM (D3S$^{\bar{U}}$). Two additional D3S versions with different bounding box fitting methods were included: a minimal area rotated bounding box that contains all foreground pixels (D3S^{MA}) and a min-max axis-aligned bounding box (D3S^{MM}). All variations were re-trained on the same dataset as the original D3S.

Results of the ablation study are presented in Table 5. Removal of the foreground similarity channel from GIM (D3S$^{\bar{F}}$) causes a 4.5% performance drop, while removal of the target posterior channel (D3S$^{\bar{P}}$) reduces the performance by 13.5%. The accuracy of both variants is comparable to the original D3S, while the number of failures increases. In conclusion, each, foreground similarity and posterior channel individually contribute to robust target localization.

Removal of the entire GIM module i.e., $\mathbf{F}$ and $\mathbf{P}$ (D3S$^{\bar{F}\bar{P}}$) reduces the overall tracking performance by 27%. The accuracy drops by 14%, while the number of failures increases by 56%. This speaks of crucial importance of the GIM module for accurate segmentation as well as tracking robustness.

Removal of the GEM module (D3S$^{\bar{L}}$) reduces the tracking performance by nearly 50%. This is primarily due to significant reduction of the robustness – the number of failures increases by over 270%. Thus the GEM module is crucial for robust target selection in the segmentation process.

Finally, updating the DCF in GEM module by its own

estimated position rather than the position estimated by the final segmentation (D3S$^{\bar{U}}$) reduces the overall performance by 7.5%, primarily at a cost of significant increase in the number of failures (over 15%). Thus, accurate target position estimation from D3S crucially affects the learning of the DCF in GEM and consequently the overall tracking performance.

Replacing the proposed bounding box fitting method (Section 4.1) with the minimal area rotated bounding box (D3S^{MA}) results in a 9% reduction in EAO and a 6% reduction in accuracy. This is still a state-of-the-art result, which means that the D3S performance boost can be primarily attributed to the segmentation mask quality. The min-max bounding box fitting method (D3S^{MM}) leads to a 19% EAO and 14% accuracy reduction. Thus D3S does benefit from the rotated bounding box estimation.

5.4. Evaluation on Segmentation Datasets

Segmentation capabilities of D3S were analyzed on two popular video object segmentation benchmarks DAVIS16 [38] and DAVIS17 [40]. Under the DAVIS protocol, the segmentation algorithm is initialized on the first frame by a segmentation mask. The algorithm is then required to output the segmentation mask for all the remaining frames in the video. Performance is evaluated by two measures averaged over the sequences: mean Jaccard index ($\mathcal{J}_{\mathcal{M}}$) and mean F-measure ($\mathcal{F}_{\mathcal{M}}$). Jaccard index represents a per-pixel intersection over union between the ground-truth and the predicted segmentation mask. The F-measure is a harmonic mean of precision and recall calculated between the contours extracted from the ground-truth and the predicted segmentation masks. For further details on these performance measures, the reader is referred to [38, 34].

D3S is compared to several sota video object segmentation methods specialized to the DAVIS challenge setup: OS-VOS [5], OnAVOS [48], OSMN [53], FAVOS [7], VM [18] and PML [6]. In addition, we include the most recent segmentation-based tracker SiamMask [50], which is the only published method that performs well on both, short-term tracking as well as on video object segmentation benchmarks.

Results are shown in Table 6. D3S performs on par with most of the video object segmentation top performers on DAVIS. Compared to top performer on DAVIS2016, the performance of D3S is 12% and 14% lower in the average Jaccard index and the F-measure, respectively. On DAVIS2017 this difference is even smaller – a 6% drop in Jaccard index and 8% drop in F-measure compared to the top-performer OnAVOS. This is quite remarkable, considering that D3S is 200 times faster. Furthermore, D3S delivers a comparable segmentation accuracy as pure segmentation methods ASMN and PML, while being orders of magnitude faster and achieving a near-realtime video object segmenta-

D3S	$\bar{\text{F}}$	$\bar{\text{U}}$	$\bar{\text{P}}$	$\bar{\text{F}}\bar{\text{P}}$	$\bar{\text{L}}$	MA	MM
EAO	0.489	0.467	0.452	0.423	0.357	0.251	0.444	0.398
Acc.	0.64	0.65	0.63	0.60	0.55	0.60	0.60	0.55
Rob.	0.150	0.187	0.173	0.211	0.234	0.567	0.160	0.173

Table 5. VOT2018 – ablation study. Removing: the GIM foreground similarity channel ($\bar{\text{F}}$), the GIM foreground probability channel ($\bar{\text{P}}$), both GIM channels ($\bar{\text{F}}\bar{\text{P}}$) and the GEM channel ($\bar{\text{L}}$). The DCF in GEM is updated from its own position estimation rather than position estimated by D3S ($\bar{\text{U}}$). D3S with a minimal area rotated bounding box (MA) and a min-max axis-aligned bounding box (MM).

	$\mathcal{J_M}^{16}$	$\mathcal{F_M}^{16}$	$\mathcal{J_M}^{17}$	$\mathcal{F_M}^{17}$	FPS
D3S	75.4	72.6	57.8	63.8	25.0
SiamMask [50]	71.7	67.8	54.3	58.5	55.0
OnAVOS [48]	86.1	84.9	61.6	69.1	0.1
FAVOS [7]	82.4	79.5	54.6	61.8	0.8
VM [18]	81.0	-	56.6	-	3.1
OSVOS [5]	79.8	80.6	56.6	63.9	0.1
PML [6]	75.5	79.3	-	-	3.6
OSMN [53]	74.0	72.9	52.5	57.1	8.0

Table 6. State-of-the-art comparison on the DAVIS16 and DAVIS17 segmentation datasets. Average Jaccard index and F-measure are denoted as $\mathcal{J_M}^{16}$ and $\mathcal{F_M}^{16}$ on DAVIS16 dataset and $\mathcal{J_M}^{17}$ and $\mathcal{F_M}^{17}$ on DAVIS17 dataset, respectively.

tion, which is particularly important for many video editing applications.

D3S also outperforms the only tracking and segmentation method SiamMask with respect to all measures. On average the segmentation is improved by over 5% in the Jaccard index and the contour accuracy-based F-measure. See Figure 6 for further qualitative comparison of D3S and SiamMask on challenging targets.

6. Conclusion

A deep single-shot discriminative segmentation tracker – D3S – was introduced. The tracker leverages two models from the extremes of the spectrum: a geometrically invariant model and a geometrically restricted Euclidean model. The two models localize the target in parallel pathways and complement each other to achieve high segmentation accuracy of deformable targets and robust discrimination of the target from distractors. The end-to-end trainable network architecture is the first single-shot pipeline with on-line adaptation that tightly connects discriminative tracking with accurate segmentation.

D3S outperforms state-of-the-art trackers on the VOT2016, VOT2018 and GOT-10k benchmarks and performs on par with top trackers on TrackingNet, regardless of the fact that some of the tested trackers were re-trained for specific datasets. In contrast, D3S was trained once on Youtube-VOS (for segmentation only) and the same version was used in all benchmarks. Tests on DAVIS16 and DAVIS17 segmentation benchmarks show performance close to top segmentation methods while running up to

Figure 6. D3S vs. SiamMask segmentation quality. *Bolt2*: SiamMask drifts to a similar object, D3S leverages the discriminative learning in GEM to robustly track the selected target. *Hand*: the rigid template in SiamMask fails on a deforming target, the GIM model in D3S successfully tracks despite a significant deformation. *Paragliding*: clutter causes drift and failure of SiamMask while in D3S, the combination of the GIM and GEM models leads to accurate and robust segmentation.

200× faster, close to real-time. D3S significantly outperforms recent top segmentation tracker SiamMask on all bechmarks in all metrics and contributes towards narrowing the gap between two, currently separate, domains of short-term tracking and video object segmentation, thus blurring the boundary between the two.

References

[1] Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip H. S. Torr. Staple: Complementary learners for real-time tracking. In *Comp. Vis. Patt. Recognition*, pages 1401–1409, June 2016. 2

[2] Luca Bertinetto, Jack Valmadre, João F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCV VOT Workshop*, 2016. 1, 2, 6, 7

[3] Goutam Bhat, Joakim Johnander, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Unveiling the power of deep tracking. In *Proc. European Conf. Computer Vision*, pages 493–509, 2018. 1

[4] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *Comp. Vis. Patt. Recognition*, pages 2544–2550, 2010. 2

[5] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixe, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Comp. Vis. Patt. Recognition*, 2017. 2, 7, 8

[6] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *Comp. Vis. Patt. Recognition*, 2018. 2, 7, 8

[7] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. In *Comp. Vis. Patt. Recognition*, 2018. 2, 7, 8

[8] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(5):564–577, May 2003. 2

[9] Kenan Dai, Dong Wang, Huchuan Lu, Chong Sun, and Jianhua Li. Visual tracking via adaptive spatially-regularized correlation filters. In *Comp. Vis. Patt. Recognition*, 2019. 5, 6

[10] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: Accurate tracking by overlap maximization. In *Comp. Vis. Patt. Recognition*, June 2019. 1, 2, 3, 5, 6, 7

[11] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ECO: Efficient convolution operators for tracking. In *Comp. Vis. Patt. Recognition*, pages 6638–6646, 2017. 1, 2, 7

[12] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost van de Weijer. Adaptive color attributes for real-time visual tracking. In *Comp. Vis. Patt. Recognition*, pages 1090–1097, 2014. 2

[13] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. Beyond correlation filters: learning continuous convolution operators for visual tracking. In *Proc. European Conf. Computer Vision*, pages 472–488, 2016. 3, 5, 6

[14] Andrew Fitzgibbon and Robert B. Fisher. A buyer's guide to conic fitting. In *Proc. British Machine Vision Conference*, pages 513–522, 1995. 4

[15] Anfeng He, Chong Luo, Xinmei Tian, and Wenjun Zeng. A twofold siamese network for real-time object tracking. In *Comp. Vis. Patt. Recognition*, 2018. 2

[16] David Held, Sebastian Thrun, and Silvio Savarese. Learning to track at 100 fps with deep regression networks. In *Proc. European Conf. Computer Vision*, pages 749–765, 2016. 6

[17] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(3):583–596, 2015. 2

[18] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G. Schwing. VideoMatch: Matching based Video Object Segmentation. In *Proc. European Conf. Computer Vision*, 2018. 2, 3, 7, 8

[19] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 2, 5, 6

[20] Ugur Kart, Alan Lukezic, Matej Kristan, Joni-Kristian Kamarainen, and Jiri Matas. Object tracking by reconstruction with view-specific discriminative correlation filters. In *Comp. Vis. Patt. Recognition*, 2019. 2

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[22] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Cehovin Zajc, Tomas Vojir, Gustav Hager, Alan Lukezic, Abdelrahman Eldesokey, and Gustavo Fernandez. The visual object tracking VOT2017 challenge results. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[23] Matej Kristan, Aleš Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Čehovin, Tomas Vojir, Gustav Häger, Alan Lukežič, and Gustavo et al. Fernandez. The visual object tracking VOT2016 challenge results. In *Proc. European Conf. Computer Vision*, 2016. 2, 5

[24] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Luka Čehovin Zajc, Tomáš Vojír, Goutam Bhat, Alan Lukežič, Abdelrahman Eldesokey, Gustavo Fernández, and et al. The sixth visual object tracking VOT2018 challenge results. In *Computer Vision – ECCV 2018 Workshops*, pages 3–53, 2018. 1, 2, 5

[25] Matej Kristan, Jiri Matas, Aleš Leonardis, Michael Felsberg, Luka Čehovin, Gustavo Fernandez, Tomas Vojir, Gustav Häger, Georg Nebehay, and Roman et al. Pflugfelder. The visual object tracking VOT2015 challenge results. In *Int. Conf. Computer Vision*, 2015. 5

[26] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016. 5, 7

[27] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In *Comp. Vis. Patt. Recognition*, 2019. 1, 2, 6, 7

[28] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Comp. Vis. Patt. Recognition*, 2018. 1, 2, 5

[29] Peixia Li, Boyu Chen, Wanli Ouyang, Dong Wang, Xiaoyun Yang, and Huchuan Lu. GradNet: Gradient-guided network for visual object tracking. In *Int. Conf. Computer Vision*, 2019. 2

[30] Yang Li and Jianke Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *Proc. European Conf. Computer Vision*, pages 254–265, 2014. 2

[31] A. Lukežič, L. Č. Zajc, and M. Kristan. Deformable parts correlation filters for robust visual tracking. *IEEE Trans. on Cyber.*, 48(6):1849–1861, 2017. 1, 2

[32] Alan Lukezic, Luka Cehovin Zajc, Tomas Vojir, Jiri Matas, and Matej Kristan. FuCoLoT - a fully-correlational long-term tracker. In *ACCV*, 2018. 2

[33] Alan Lukežič, Tomáš Vojíř, Luka Čehovin Zajc, Jiří Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *Comp. Vis. Patt. Recognition*, pages 6309–6318, 2017. 1, 2, 3, 5

[34] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(5):530–549, 2004. 7

[35] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. In *Proc. European Conf. Computer Vision*, September 2018. 2, 5, 7

[36] Hyeonseob Nam, Mooyeol Baek, and Bohyung Han. Modeling and propagating CNNs in a tree structure for visual tracking. *arXiv preprint arXiv:1608.07242*, 2016. 5

[37] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *Comp. Vis. Patt. Recognition*, pages 4293–4302, June 2016. 6, 7

[38] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Comp. Vis. Patt. Recognition*, 2016. 1, 2, 7

[39] Pedro O. Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *Proc. European Conf. Computer Vision*, pages 75–91, 2016. 4

[40] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv:1704.00675*, 2017. 1, 2, 7

[41] Horst Possegger, Thomas Mauthner, and Horst Bischof. In defense of color-based model-free tracking. In *Comp. Vis. Patt. Recognition*, June 2015. 2

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015. 4

[43] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut-interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (SIGGRAPH)*, August 2004. 5

[44] Ran Tao, Efstratios Gavves, and Arnold W M Smeulders. Siamese instance search for tracking. In *Comp. Vis. Patt. Recognition*, pages 1420–1429, 2016. 2

[45] Ludovic Trottier, Philippe Gigu, Brahim Chaib-draa, et al. Parametric exponential linear unit for deep convolutional neural networks. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 207–214, 2017. 3

[46] Jack Valmadre, Luca Bertinetto, Joao Henriques, Andrea Vedaldi, and Philip H. S. Torr. End-to-end representation learning for correlation filter based tracking. In *Comp. Vis. Patt. Recognition*, pages 2805–2813, 2017. 2, 6, 7

[47] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. FEELVOS: Fast end-to-end embedding learning for video object segmentation. In *Comp. Vis. Patt. Recognition*, 2019. 2

[48] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *Proc. British Machine Vision Conference*, 2017. 2, 7, 8

[49] Guangting Wang, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. SPM-Tracker: Series-parallel matching for real-time visual object tracking. In *Comp. Vis. Patt. Recognition*, 2019. 5, 6

[50] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H.S. Torr. Fast online object tracking and segmentation: A unifying approach. In *Comp. Vis. Patt. Recognition*, June 2019. 2, 5, 6, 7, 8

[51] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. YouTube-VOS: A large-scale video object segmentation benchmark. *arXiv:1809.03327*, 2018. 1, 2, 5, 7

[52] T. Xu, Z. Feng, X. Wu, and J. Kittler. Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. *IEEE Trans. Image Proc.*, 28(11):5596–5609, 2019. 6

[53] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K. Katsaggelos. Efficient video object segmentation via network modulation. In *Comp. Vis. Patt. Recognition*, 2018. 2, 7, 8

[54] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proc. European Conf. Computer Vision*, pages 103–119, 2018. 1, 6