

Siam R-CNN: Visual Tracking by Re-Detection

Paul Voigtlaender¹ Jonathon Luiten^{1,2,†}

¹RWTH Aachen University

{voigtlaender, luiten, leibe}@vision.rwth-aachen.de

Philip H.S. Torr² Bastian Leibe¹

²University of Oxford

phst@robots.ox.ac.uk

Abstract

We present Siam R-CNN, a Siamese re-detection architecture which unleashes the full power of two-stage object detection approaches for visual object tracking. We combine this with a novel tracklet-based dynamic programming algorithm, which takes advantage of re-detections of both the first-frame template and previous-frame predictions, to model the full history of both the object to be tracked and potential distractor objects. This enables our approach to make better tracking decisions, as well as to re-detect tracked objects after long occlusion. Finally, we propose a novel hard example mining strategy to improve Siam R-CNN’s robustness to similar looking objects. The proposed tracker achieves the current best performance on ten tracking benchmarks, with especially strong results for long-term tracking.

1. Introduction

We approach Visual Object Tracking using the paradigm of Tracking by Re-Detection. We present a powerful novel re-detector, Siam R-CNN, an adaptation of Faster R-CNN [74] with a Siamese architecture, which re-detects a template object anywhere in an image by determining if a region proposal is the same object as a template region, and regressing the bounding box for this object. Our two-stage re-detection architecture is robust against changes in object size and aspect ratio as the proposals are aligned to the same size, which is in contrast to the popular cross-correlation-based methods [49].

Tracking by re-detection has a long history, reaching back to the seminal work of Avidan [1] and Grabner *et al.* [28]. Re-detection is challenging due to the existence of distractor objects that are very similar to the template object. In the past, the problem of distractors has mainly been approached by strong spatial priors from previous predictions [4, 49, 48], or by online adaptation [1, 28, 2, 76, 30, 77, 42]. Both of these strategies are prone to drift.

We instead approach the problem of distractors by making two novel contributions beyond our Siam R-CNN re-

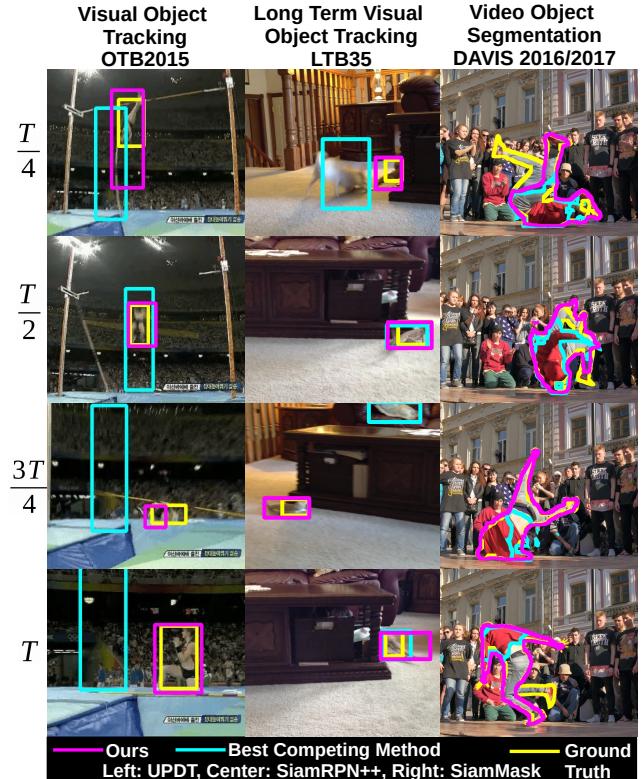


Figure 1: Example results of Siam R-CNN on 3 different tracking tasks where it obtains new state-of-the-art results.

detector design. Firstly we introduce a novel hard example mining procedure which trains our re-detector specifically for difficult distractors. Secondly we propose a novel Tracklet Dynamic Programming Algorithm (TDPA) which simultaneously tracks all potential objects, including distractor objects, by re-detecting all object candidate boxes from the previous frame, and grouping boxes over time into tracklets (short object tracks). It then uses dynamic programming to select the best object in the current timestep based on the complete history of all target object and distractor object tracklets in a video. By explicitly modeling the motion and interaction of all potential objects and pooling similarity information from detections grouped into tracklets, Siam R-CNN is able to effectively perform long-term tracking, while being resistant to tracker drift, and being

[†]Work performed both while at the RWTH Aachen and on a research visit at the University of Oxford.

able to immediately re-detect objects after disappearance. Our TDPA requires only a small set of new re-detections in each timestep, updating its tracking history iteratively online. This allows Siam R-CNN to run at 4.7 frames per second (FPS) with a ResNet-101 backbone, and at more than 15 FPS with a ResNet-50 backbone, fewer input proposals, and smaller input image size.

We present evaluation results on a large number of datasets. Siam R-CNN outperforms all previous methods on six short-term tracking benchmarks, OTB2015 [99], TrackingNet [66], GOT-10k [38], NFS [43], VOT2015 [46] and OTB50 [99] as well as on four long-term tracking benchmarks, LTB35 [62], UAV20L [65], LaSOT [23] and Ox-UVA [86], where it achieves especially strong results, up to 10 percentage points higher than previous methods. By obtaining segmentation masks using an off-the-shelf box-to-segmentation network, Siam R-CNN also outperforms all previous Video Object Segmentation methods that only use the first-frame bounding box (without the mask) on DAVIS 2017 (both val and test-dev) [72], YouTube-VOS 2018 [101] and DAVIS 2016 [71]. All code and models will be made available.

2. Related Work

Visual Object Tracking (VOT). VOT is the task of tracking an object through a video given the first-frame bounding box of the object. VOT is commonly evaluated on benchmarks such as OTB [98, 99], the yearly VOT challenges [47, 45], and many more [66, 38, 117, 65, 43]. Recently a number of long-term tracking benchmarks have been proposed [62, 86, 23] which extend VOT to a more difficult and realistic setting, where objects must be tracked over many frames, with objects disappearing and reappearing.

Many classical methods use an online learned classifier to re-detect the object of interest over the full image [1, 28, 2, 76, 30, 77, 42]. In contrast, Siam R-CNN learns the expected appearance variations by offline training instead of learning a classifier online.

Like our Siam R-CNN, many recent methods approach VOT using Siamese architectures. Siamese region proposal networks (SiamRPN [49]) use a single-stage RPN [74] detector adapted to re-detect a template by cross-correlating the deep template features with the deep features of the current frame. Here, single-stage means directly classifying anchor boxes [57] which is in contrast to two-stage architectures [74] which first generate proposals, and then align their features and classify them in the second stage. Recent tracking approaches improve upon SiamRPN, making it distractor aware (DaSiamRPN [118]), adding a cascade (C-RPN [25]), producing masks (SiamMask [94]), using deeper architectures (SiamRPN+ [114] and SiamRPN++ [48]) and maintaining a set of diverse templates (THOR [78]). These (and many more [7, 35, 63]) only search for

the object within a small window of the previous prediction. DiMP [5] follows this paradigm while meta-learning a robust target and background appearance model.

Other recent developments in VOT include using domain specific layers with online learning [67], learning an adaptive spatial filter regularizer [17], exploiting category-specific semantic information [85], using continuous [20] or factorized [18] convolutions, and achieving accurate bounding box predictions using an overlap prediction network [19]. Huang *et al.* [39] propose a framework to convert any detector into a tracker. Like Siam R-CNN, they also apply two-stage architectures, but their methods relies on meta-learning and achieves a much lower accuracy.

Long-term tracking is mainly addressed by enlarging the search window of these Siamese trackers when detection confidence is low [118, 48]. In contrast, we use a two-stage Siamese re-detector which searches over the whole image, producing stronger results than current methods across many benchmarks, and especially for long-term tracking.

Video Object Segmentation (VOS). VOS is an extension of VOT where a set of template segmentation masks are given, and segmentation masks need to be produced in each frame. Many methods perform fine-tuning on the template masks [8, 64, 89, 52, 3, 60], which leads to strong results but is slow. Recently, several methods have used the first-frame masks without fine-tuning [12, 105, 13, 37, 100, 101, 87, 69], running faster but often not performing as well.

Very few methods [94, 108] tackle the harder problem of producing mask tracking results while only using the given template bounding box and not the mask. We adapt our method to perform VOS in this setting by using a second network to produce masks for our box tracking results.

3. Method

Inspired by the success of Siamese trackers [45, 99, 47], we use a Siamese architecture for our re-detector. Many recent trackers [118, 94, 48, 49, 5] adopt a single-stage detector architecture. For the task of single-image object detection, two-stage detector networks such as Faster R-CNN [74] have been shown to outperform single-stage detectors. Inspired by this, we design our tracker as a Siamese two-stage detection network. The second stage can directly compare a proposed Region of Interest (RoI) to a template region by concatenating their RoI aligned features. By aligning proposals and reference to the same size, Siam R-CNN achieves robustness against changes in object size and aspect ratio, which is hard to achieve by when using the popular cross-correlation operation [49]. Fig. 2 shows an overview of Siam R-CNN including the Tracklet Dynamic Programming Algorithm (TDPA).

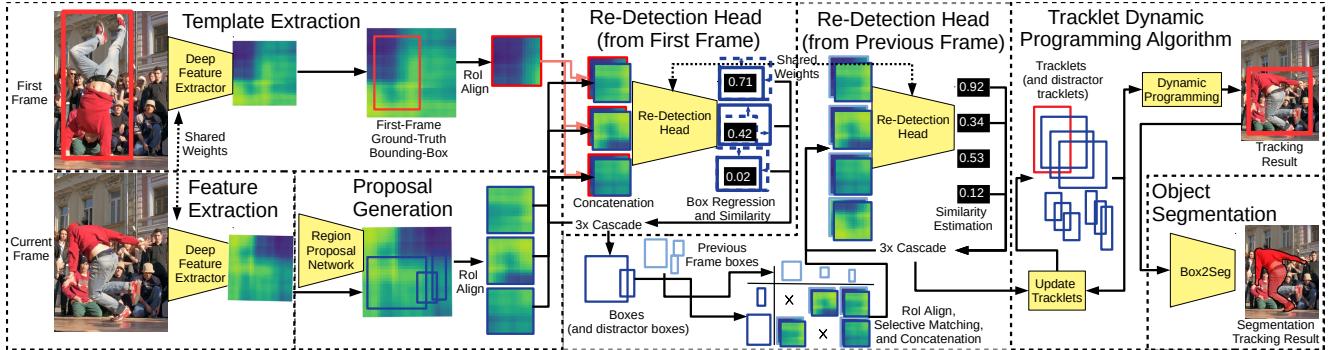


Figure 2: Overview of Siam R-CNN. A Siamese R-CNN provides re-detections of the object given in the first-frame bounding box, which are used by our Tracklet Dynamic Programming Algorithm along with re-detections from the previous frame. The results are bounding box level tracks which can be converted to segmentation masks by the Box2Seg network.

3.1. Siam R-CNN

Siam R-CNN is a Siamese re-detector based on a two-stage detection architecture. Specifically, we take a Faster R-CNN network that has been pre-trained on the COCO [56] dataset for detecting 80 object classes. This network consists of a backbone feature extractor followed by two detection stages; first a category-agnostic RPN, followed by a category-specific detection head. We fix the weights of the backbone and the RPN and replace the category-specific detection head with our re-detection head.

We create input features for the re-detection head for each region proposed by the RPN by performing RoI Align [33] to extract deep features from this proposed region. We also take the RoI Aligned deep features of the initialization bounding box in the first frame, and then concatenate these together and feed the combined features into a 1×1 convolution which reduces the number of features channels back down by half. These joined features are then fed into the re-detection head with two output classes; the proposed region is either the reference object or it is not. Our re-detection head uses a three-stage cascade [9] without shared weights. The structure of the re-detection head is the same as the structure of the detection head of Faster R-CNN, except for using only two classes and for the way the input features for the re-detection head are created by concatenation. The backbone and RPN are frozen and only the re-detection head (after concatenation) is trained for tracking, using pairs of frames from video datasets. Here, an object in one frame is used as reference and the network is trained to re-detect the same object in another frame.

3.2. Video Hard Example Mining

During conventional Faster R-CNN training, the negative examples for the second stage are sampled from the regions proposed by the RPN in the target image. However, in many images there are only few relevant negative examples. In order to maximize the discriminative power of the re-detection head, we need to train it on hard nega-



Figure 3: Hard negative mining examples. The top left image shows the reference object and all other images show the hard negative examples retrieved from other videos.

tive examples. Mining hard examples for detection has been explored in previous works (*e.g.* [26, 79]). However, rather than finding general hard examples for detection, we find hard examples for re-detection conditioned on the reference object by retrieving objects from other videos.

Embedding Network. A straightforward approach to selecting relevant videos from which to get hard negative examples for the current video, is taking videos in which an object has the same class as the current object [118]. However, object class labels are not always available, and some objects of the same class could be easy to distinguish, while some objects of different classes could also be potentially hard negatives. Hence, we propose to use an embedding network, inspired by person re-identification, which extracts an embedding vector for every ground truth bounding box which represents the appearance of that object. We use the network from PReMVOS [60], which is trained with batch-hard triplet loss [36] to separate classes on COCO before being trained on YouTube-VOS to disambiguate between individual object instances. *E.g.*, two distinct persons should be far away in the embedding space, while two crops of the same person in different frames should be close.

Index Structure. We next create an efficient indexing structure for approximate nearest neighbor queries and use

Algorithm 1 Update tracklets for one time-step t

```

1: Inputs ff_gt_feats, tracklets, image $_t$ , dets $_{t-1}$ 
2: backbone_feats  $\leftarrow$  backbone(image $_t$ )
3: RoIs  $\leftarrow$  RPN(backbone_feats)  $\cup$  dets $_{t-1}$ 
4: dets $_t$   $\leftarrow$  redetection_head(RoIs, ff_gt_feats)
5:  $\triangleright$  scores are set to  $-\infty$  if spatial distance is  $> \gamma$ 
6: scores  $\leftarrow$  score_pairwise_redetection(dets $_t$ , dets $_{t-1}$ ,  $\gamma$ )
7: for  $d_t \in \text{dets}_t$  do
8:    $s_1 \leftarrow \max_{d_{t-1} \in \text{dets}_{t-1}} \text{scores}[d_t, d_{t-1}]$ 
9:    $\hat{d}_{t-1} \leftarrow \operatorname{argmax}_{d_{t-1} \in \text{dets}_{t-1}} \text{scores}[d_t, d_{t-1}]$ 
10:   $\triangleright$  Max score of all other current detections
11:   $s_2 \leftarrow \max_{\tilde{d}_t \in \text{dets}_t \setminus \{d_t\}} \text{scores}[\tilde{d}_t, \hat{d}_{t-1}]$ 
12:   $\triangleright$  Max score of all other previous detections
13:   $s_3 \leftarrow \max_{d_{t-1} \in \text{dets}_{t-1} \setminus \{\hat{d}_{t-1}\}} \text{scores}[d_t, d_{t-1}]$ 
14:  if  $s_1 > \alpha \wedge s_2 \leq s_1 - \beta \wedge s_3 \leq s_1 - \beta$  then
15:     $\triangleright$  Extend tracklet
16:    tracklet( $\hat{d}_{t-1}$ ).append( $d_t$ )
17:  else  $\triangleright$  Start new tracklet
18:    tracklets  $\leftarrow$  tracklets  $\cup \{\{d_t\}\}$ 
19:  end if
20: end for

```

it to find nearest neighbors of the tracked object in the embedding space. Fig. 3 shows examples of the retrieved hard negative examples. As can be seen, most of the negative examples are very relevant and hard.

Training Procedure. Evaluating the backbone on-the-fly on other videos to retrieve hard negative examples for the current video frame would be very costly. Instead, we pre-compute the ROI-aligned features for every ground truth box of the training data. For each training step, as usual, a random video and object in this video is selected and then a random reference and a random target frame. Afterwards, we use the indexing structure to retrieve for the reference box the 10,000 nearest neighbor bounding boxes from other videos and sample 100 of them as additional negative training examples. More details about video hard example mining can be found in the supplemental material.

3.3. Tracklet Dynamic Programming Algorithm

Our Tracklet Dynamic Programming Algorithm (TDPA) implicitly tracks both the object of interest and potential distractors, so that distractor objects can be consistently suppressed. To this end, TDPA maintains a set of tracklets, *i.e.*, short sequences of detections, for which it is almost certain that they belong to the same object. It then uses a dynamic programming based scoring algorithm to select the most likely sequence of tracklets for the template object between the first and current frame. Each detection is defined by a bounding box, a re-detection score, and its ROI-aligned features. Furthermore, each detection is part of exactly one tracklet. A tracklet has a start and an end time and is defined by a set of detections, one for each time step from the start to the end time, *i.e.*, there are no gaps allowed in a tracklet.

Tracklet Building. We extract the ROI aligned features

for the first-frame ground truth bounding box (ff_gt_feats) and initialize a tracklet consisting of just this box. For each new frame, we update the set of tracklets as follows (*c.f.* Algorithm 1): We extract backbone features of the current frame and evaluate the region proposal network (RPN) to get regions of interest (RoIs, lines 2–3). To compensate for potential RPN false negatives, the set of RoIs is extended by the bounding box outputs from the previous frame. We run the re-detection head (including bounding box regression) on these RoIs to produce a set of re-detections of the first-frame template (line 4). Afterwards, we re-run the classification part of the re-detection head (line 6) on the current detections dets $_t$, but this time with the detections dets $_{t-1}$ from the previous frame as reference instead of the first-frame ground truth box, to calculate similarity scores (scores) between each pair of detections.

To measure the spatial distance of two detections, we represent their bounding boxes by their center coordinates x and y , and their width w and height h , of which x and w are normalized with the image width, and y and h are normalized with the image height, so that all values are between 0 and 1. The spatial distance between two bounding boxes (x_1, y_1, w_1, h_1) and (x_2, y_2, w_2, h_2) is then given by the L_∞ norm, *i.e.*, $\max(|x_1 - x_2|, |y_1 - y_2|, |w_1 - w_2|, |h_1 - h_2|)$. In order to save computation and to avoid false matches, we calculate the pairwise similarity scores only for pairs of detections where this spatial distance is less than γ and set the similarity score to $-\infty$ otherwise.

We extend the tracklets from the previous frame by the current frame detections (lines 7–20) when the similarity score to a new detection is high ($> \alpha$) and there is no ambiguity, *i.e.*, there is no other detection which has an almost as high similarity (less than β margin) with that tracklet, and there is no other tracklet which has an almost as high similarity (less than β margin) with that detection. Whenever there is any ambiguity, we start a new tracklet which initially consists of a single detection. The ambiguities will then be resolved in the tracklet scoring step.

Scoring. A track $A = (a_1, \dots, a_N)$ is a sequence of N non-overlapping tracklets, *i.e.*, $\text{end}(a_i) < \text{start}(a_{i+1}) \forall i \in \{1, \dots, N-1\}$, where start and end denote the start and end times of a tracklet, respectively. The total score of a track consists of a unary score measuring the quality of the individual tracklets, and of a location score which penalizes spatial jumps between tracklets, *i.e.*

$$\text{score}(A) = \sum_{i=1}^N \text{unary}(a_i) + \sum_{i=1}^{N-1} w_{\text{loc}} \text{loc_score}(a_i, a_{i+1}). \quad (1)$$

$$\begin{aligned} \text{unary}(a_i) = & \sum_{t=\text{start}(a_i)}^{\text{end}(a_i)} w_{\text{ff}} \text{ff_score}(a_{i,t}) \\ & + (1 - w_{\text{ff}}) \text{tracklet_score}(a_{i,t}), \end{aligned} \quad (2)$$

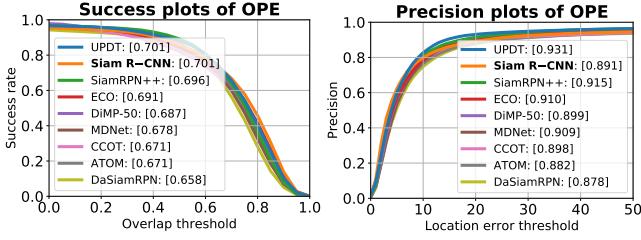


Figure 4: Results on OTB2015 [99].

where ff_score denotes the re-detection confidence for the detection $a_{i,t}$ of tracklet a_i at time t from the re-detection head using the first-frame ground truth bounding box as reference. There is always one tracklet which contains the first-frame ground truth bounding box, which we denote as the first-frame tracklet a_{ff} . All detections in a tracklet have a very high chance of being a correct continuation of the initial detection of this tracklet, because in cases of ambiguities tracklets are terminated. Hence, the most recent detection of the first-frame tracklet is also the most recent observation that is almost certain to be the correct object. Thus, we also use this most recent detection of the first-frame tracklet as an additional reference for re-detection. This score is denoted by ff_tracklet_score and linearly combined with the ff_score .

The location score between two tracklets a_i and a_j is given by the negative L_1 norm of the difference between the bounding box (x, y, w, h) of the last detection of a_i and the bounding box of the first detection of a_j , *i.e.*

$$\text{loc_score}(a_i, a_j) = -|\text{end_bbox}(a_i) - \text{start_bbox}(a_j)|_1.$$

Online Dynamic Programming. We efficiently find the sequence of tracklets with the maximum total score (Eq. 1) by maintaining an array θ which for each tracklet a stores the total score $\theta[a]$ of the optimal sequence of tracklets which starts with the first-frame tracklet and ends with a .

Once a tracklet is not extended, it is terminated. Thus, for each new frame only the scores for tracklets which have been extended or newly created need to be newly computed. For a new time-step, first we set $\theta[a_{\text{ff}}] = 0$ for the first-frame tracklet a_{ff} , since all tracks have to start with that tracklet. Afterwards, for every tracklet a which has been updated or newly created, $\theta[a]$ is calculated as

$$\theta[a] = \text{unary}(a) + \max_{\tilde{a}: \text{end}(\tilde{a}) < \text{start}(a)} \theta[\tilde{a}] + w_{\text{loc}} \text{loc_score}(\tilde{a}, a)$$

To retain efficiency for very long sequences, we allow a maximum temporal gap between two tracklets of 1500 frames, which is long enough for most applications.

After updating θ for the current frame, we select the tracklet \hat{a} with the highest dynamic programming score, *i.e.* $\hat{a} = \arg \max_a \theta[a]$. If the selected tracklet does not contain a detection in the current frame, then our algorithm has indicated that the object is not present. For benchmarks that require a prediction in every frame we use the most recent box from the selected tracklet, and assign it a score of 0.

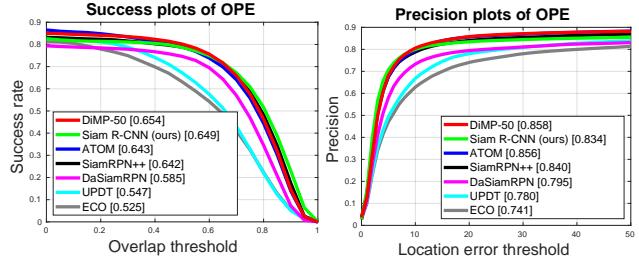


Figure 5: Results on UAV123 [65].

3.4. Box2Seg

To produce segmentation masks for the VOS task, we use an off-the-shelf bounding-box-to-segmentation (Box2Seg) network from PReMVOS [60]. Box2Seg is a fully convolutional DeepLabV3+ [11] network with an Xception-65 [16] backbone. It has been trained on Mapillary [68] and COCO [56] to output the mask for a bounding box crop. Box2Seg is fast, running it after tracking only requires 0.025 seconds per object per frame. We combine overlapping masks such that masks with less pixels are on top.

3.5. Training Details

Siam R-CNN is built upon the Faster R-CNN [74] implementation and the pre-trained weights from [96], with a ResNet-101-FPN backbone [34, 55], group normalization [97] and cascade [9]. It has been pre-trained from scratch [32] on COCO [56]. Except where specified otherwise, we train Siam R-CNN on the training sets of multiple tracking datasets simultaneously: ImageNet VID [75] (4000 videos), YouTube-VOS 2018 [101] (3471 videos), GOT-10k [38] (9335 videos) and LaSOT [23] (1120 videos). We train with motion blur and grayscale augmentations [118], as well as gamma and scale augmentations.

4. Experiments

We evaluate Siam R-CNN for standard visual object tracking, for long-term tracking, and on VOS benchmarks. We tune a single set of hyper-parameters for our Tracklet Dynamic Programming Algorithm (*c.f.* Section 3.3) on the DAVIS 2017 training set, as this is a training set that we did not use to train our re-detector. We present results using these hyper-parameters on all benchmarks, rather than tuning the parameters separately for each one.

4.1. Short-Term Visual Object Tracking Evaluation

For short-term VOT, we use the following benchmarks: OTB2015 [99], VOT2018 [45], UAV123 [65], TrackingNet [66], and GOT-10k [38]. In the supplemental material we also evaluate on VOT2015 [46], VOT2016 [44], TC128 [54], OTB2013 [98] and OTB50 [99].

OTB2015. We evaluate on OTB2015 (100 videos, 590 frames average length), calculating the success and precision over varying overlap thresholds. Methods are ranked

	Huang <i>et al.</i> [39]	UPDT [6]	ATOM [19]	Tripathi <i>et al.</i> [85]	DiMP-50 [5]	Siam R-CNN
Success	51.5	53.7	58.4	60.5	62.0	63.9

Table 1: Results on NfS [5].

	DaSiamRPN [118]	UPDT [6]	ATOM [19]	SiamRPN++ [48]	DiMP-50 [5]	Siam R-CNN
Precision	59.1	55.7	64.8	69.4	68.7	80.0
Norm. Prec.	73.3	70.2	77.1	80.0	80.1	85.4
Success	63.8	61.1	70.3	73.3	74.0	81.2

Table 2: Results on TrackingNet [66].

	LADCF [103]	ATOM [19]	SiamRPN++ [48]	THOR [5]	DiMP-50 [78]	Ours (short-t.)
EAO	0.389	0.401	0.414	0.416	0.440	0.140
Accuracy	0.503	0.590	0.600	0.582	0.597	0.624
Robustn.	0.159	0.204	0.234	0.234	0.153	0.039
AO	0.421	-	0.498	-	-	0.476

Table 3: Results on VOT2018 [45].

by the area under the success curve (AUC). Fig. 4 compares our results to eight state-of-the-art (SOTA) trackers [6, 48, 18, 5, 67, 20, 19, 118]. Siam R-CNN achieves 70.1% AUC, which equals the previous best result by UPDT [6].

UAV123. Fig. 5 shows our results on UAV123 [65] (123 videos, 915 frames average length) on the same metrics as OTB2015 compared to six SOTA approaches [5, 19, 48, 118, 6, 18]. We achieve an AUC of 64.9%, which is close to the previous best result of DiMP-50 [5] with 65.4%.

NfS. Tab. 1 shows our results on the NfS dataset [43] (30FPS, 100 videos, 479 frames average length) compared to five SOTA approaches. Siam R-CNN achieves a success score of 63.9%, which is 1.9 percentage points higher than the previous best result by DiMP-50 [5].

TrackingNet. Tab. 2 shows our results on the TrackingNet test set [66] (511 videos, 442 frames average length), compared to five SOTA approaches. Siam R-CNN achieves a success score of 81.2%, which is 7.2 percentage points higher than the previous best result of DiMP-50 [5]. In terms of precision the gap is more than 10 percentage points.

GOT-10k. Fig. 6 shows our results on the GOT-10k [38] test set (180 videos, 127 frames average length) compared to six SOTA approaches [5, 116, 19, 90, 48, 18]. On this benchmark, methods are only allowed to use the GOT-10k training set as video data for training. Therefore we train a new model starting from COCO pre-training, and train only on GOT-10k. We achieve a success rate of 64.9% which is 3.8 percentage points higher than the previous best result from DiMP-50 [5]. This shows that Siam R-CNN’s advantage over all previous methods is not just due to different training data, but from the tracking approach itself.

VOT2018. Table 3 shows our results on VOT2018 [45] (60 videos, 356 frames average length). As is the standard

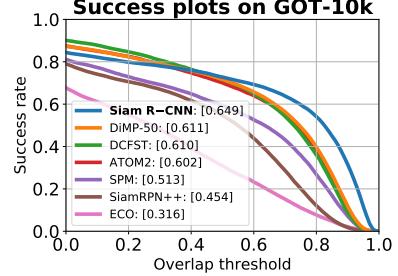


Figure 6: Results on GOT-10k [38].

for VOT2018, a reset-based evaluation is used, where once the object is lost, the tracker is restarted with the ground truth box five frames later and receives a penalty. The main evaluation criterion is the Expected Average Overlap (EAO) [46]. This extreme short-term tracking scenario is not what Siam R-CNN with the TDPA was designed for. It often triggers resets, which normally (without reset-based evaluation) Siam R-CNN can automatically recover from, resulting in an EAO of 0.140. Since VOT2018 is an important tracking benchmark, we created a simple short-term version of Siam R-CNN which assumes the previous-frame prediction is correct (it must be, or else it would reset) and then averages the predictions of re-detecting the first-frame reference and re-detecting the previous prediction and combines them with a strong spatial prior (more details in the supplemental material). Our short-term Siam R-CNN achieves an EAO score of 0.408 which is competitive with many SOTA approaches. Notably, both versions of Siam R-CNN achieve the highest accuracy scores. The last row shows the average overlap (AO), when using the normal (non-reset) evaluation. In this setup, Siam R-CNN achieves 0.476 AO, close to the best result of 0.498 of SiamRPN++.

4.2. Long-Term Visual Object Tracking Evaluation

To evaluate the ability of Siam R-CNN to perform long-term tracking, we evaluate on three benchmarks, LTB35 [62], LaSOT [23] and OxUvA [86]. In the supplemental material we also evaluate on UAV20L [65]. In long-term tracking, sequences are much longer, and objects may disappear and reappear again (LTB35 has on average 12.4 disappearances per video, each one on average 40.6 frames long). Siam R-CNN significantly outperforms all previous methods on all of these benchmarks, indicating the strength of our tracking by re-detection approach. By searching globally over the whole image rather than within a local window of a previous prediction, our method is more resistant to drift, and can easily re-detect a target after disappearance.

LTB35. Fig. 7 shows the results of our method on the LTB35 benchmark (also known at VOT18-LT) [62] (35 videos, 4200 frames average length) compared to eight other SOTA approaches. Trackers are required to output a confidence of the target being present for the prediction

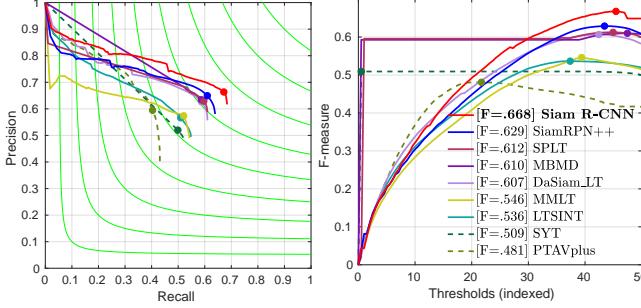


Figure 7: Results on LTB35 [62] (VOT18-LongTerm).

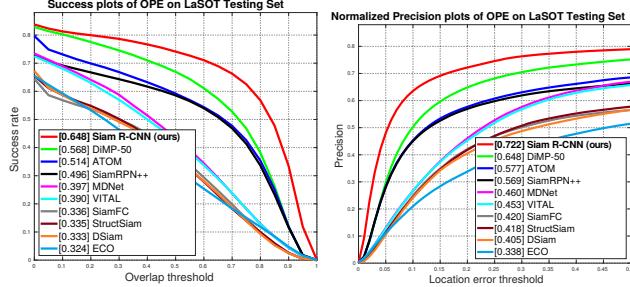


Figure 8: Results on LaSOT [23].

	DaSiam_LT	TLD	SiamFC+R	MBMD	SPLT	Siam
	[45]	[42]	[86]	[111]	[104]	R-CNN
MaxGM	41.5	43.1	45.4	54.4	62.2	72.3
TPR	68.9	20.8	42.7	60.9	49.8	70.1
TNR	0	89.5	48.1	48.5	77.6	74.5

Table 4: Results on OxUvA [86].

in each frame. Precision (Pr) and Recall (Re) are evaluated for a range of confidence thresholds, and the F-score is calculated as $F = \frac{2PrRe}{Pr+Re}$. Trackers are ranked by the maximum F-score over all thresholds. We compare to the 6 best-performing methods in the 2018 VOT-LT challenge [45] and to SiamRPN++ [48] and SPLT [104]. Siam R-CNN outperforms all previous methods with an F -score of 66.8%, 3.9 percentage points higher than the previous best result.

LaSOT. Fig. 8 shows results on the LaSOT test set [23] (280 videos, 2448 frames average length) compared to nine SOTA methods [5, 19, 48, 67, 80, 4, 112, 29, 18]. Siam R-CNN achieves an unprecedented result with a success rate of 64.8% and 72.2% normalized precision. This is 8 percentage points higher in success and 7.4 points higher in normalized precision than the previous best method.

OxUvA. Tab. 4 shows results on the OxUvA test set [86] (166 videos, 3293 frames average length) compared to five SOTA methods. Trackers must make a hard decision each frame whether the object is present. We do this by comparing the detector confidence to a threshold tuned on the dev set. Methods are ranked by the maximum geometric mean (MaxGM) of the true positive rate (TPR) and the true negative rate (TNR). Siam R-CNN achieves a MaxGM more than 10 percentage points higher than all previous methods.

Init	Method	FT	M	$\mathcal{J} \& \mathcal{F}$	\mathcal{J}	\mathcal{F}	\mathcal{J}_{box}	t(s)
bbox	Siam R-CNN (ours)	x	x	70.6	66.1	75.0	78.3	0.32
	Siam R-CNN (fastest)	x	x	70.5	66.4	74.6	76.9	0.12
	SiamMask [94]	x	x	55.8	54.3	58.5	64.3	0.06 [†]
	SiamMask [94] (Box2Seg)	x	x	63.3	59.5	67.3	64.3	0.11
	SiamRPN++ [48] (Box2Seg)	x	x	61.6	56.8	66.3	64.0	0.11
	DiMP-50 [5] (Box2Seg)	x	x	63.7	60.1	67.3	65.6	0.10
mask	STM-VOS [69]	x	✓	81.8	79.2	84.3	—	0.32 [†]
	FEELVOS [87]	x	✓	71.5	69.1	74.0	71.4	0.51
	RGMP [100]	x	✓	66.7	64.8	68.6	66.5	0.28 [†]
	VideoMatch [37]	x	✓	62.4	56.5	68.2	—	0.35
mask+ft	PReMVOS [60]	✓	✓	77.8	73.9	81.7	81.4	37.6
	Ours (Fine-tun. Box2Seg)	✓	✓	74.8	69.3	80.2	78.3	1.0
	DyeNet [52]	✓	✓	74.1	—	—	—	9.32 [†]
	OSVOS-S [64]	✓	✓	68.0	64.7	71.3	68.4	9 [†]
	OnAVOS [89]	✓	✓	63.6	61.0	66.1	66.3	26
	GT boxes (Box2Seg)	x	x	82.6	79.3	85.8	100.0	—
	GT boxes (Fine-t. Box2Seg)	✓	✓	86.2	81.8	90.5	100.0	—

Table 5: Results on the DAVIS 2017 validation set. FT: fine-tuning, M: using the first-frame masks, t(s): time per frame in seconds. [†]: timing extrapolated from DAVIS 2016 assuming linear scaling in the number of objects. An extended table can be found in the supplemental material. Siam R-CNN (fastest) denotes Siam R-CNN with ResNet-50 backbone, half input resolution, and 100 RoIs from the RPN.

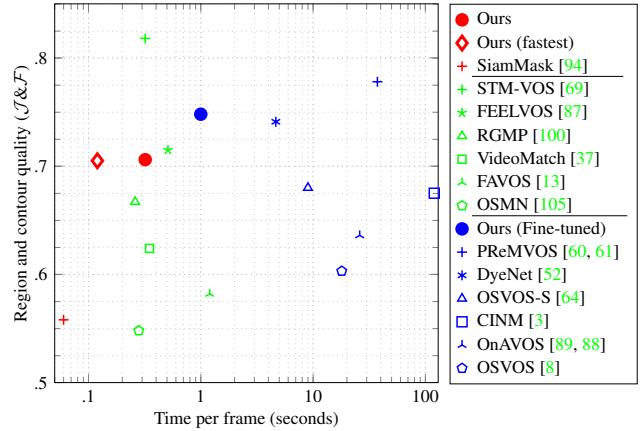


Figure 9: Quality versus timing on the DAVIS 2017 validation set. Only SiamMask [94] and our method (red) are able to work without the ground truth mask of the first frame and require just the bounding box. Methods shown in blue fine-tune on the first-frame mask. Ours (fastest) denotes Siam R-CNN with ResNet-50 backbone, half input resolution, and 100 RoIs from the RPN, see Section 4.5

4.3. Video Object Segmentation (VOS) Evaluation

We further evaluate our method on the VOS benchmarks DAVIS17 [72] and YouTube-VOS [101]. This evaluates two aspects beyond other tracking evaluations. Firstly, the ability to produce accurate segmentation tracking results (for which we use Box2Seg after running Siam R-CNN for tracking). Secondly the requirement to track multiple objects simultaneously. In VOS, trackers are evaluated using the \mathcal{J} metric (average intersection over union (IoU) of pre-

Init	Method	FT	M	\mathcal{O}	\mathcal{J}_{seen}	\mathcal{J}_{unseen}	t(s)
bbox	Siam R-CNN (ours)	X	X	68.3	69.9	61.4	0.32
	Siam R-CNN (fastest)	X	X	66.2	69.2	57.7	0.12
	SiamMask [94]	X	X	52.8	60.2	45.1	0.06
mask	STM-VOS [69]	X	✓	79.4	79.7	72.8	0.30 [†]
	RGMP [100]	X	✓	53.8	59.5	45.2	0.26 [†]
	Ours (Fi.-tu. Box2Seg)	✓	✓	73.2	73.5	66.2	0.65
mask+ft	PReMVOS [60, 59]	✓	✓	66.9	71.4	56.5	6
	OnAVOS [89]	✓	✓	55.2	60.1	46.6	24.5
	OSVOS [8]	✓	✓	58.8	59.8	54.2	17 [†]
	Ours (Fine-tuned)	✓	✓	73.2	73.5	66.2	0.65

Table 6: Results on the YouTube-VOS 2018 [101] validation set. The notation is explained in the caption of Tab. 5.

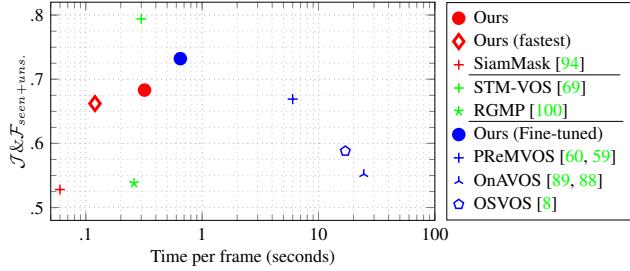


Figure 10: Quality versus timing on the YouTube-VOS 2018 [101] validation set. Only SiamMask [94] and our method (red) are able to work without the ground truth mask of the first frame and require just the bounding box. Methods shown in blue fine-tune on the first-frame mask. Ours (fastest) denotes Siam R-CNN with ResNet-50 backbone, half input resolution, and 100 RoIs from the RPN, see Section 4.5

dicted and ground truth mask) and the \mathcal{F} metric (boundary similarity). We also introduce a new metric \mathcal{J}_{box} which is the \mathcal{J} metric but calculated using the bounding box around a predicted and ground truth mask, which can be used to directly measure tracking quality without requiring masks. We divide VOS methods into those that do not use the first-frame mask (like ours), those that use it directly, and those that fine-tune upon it. As well as our standard Siam R-CNN with Box2Seg, we also present results where we fine-tune Box2Seg for 300 steps on the first-frame masks (Siam R-CNN is never fine-tuned and only uses the first-frame box).

DAVIS 2017. Table 5 and Figure 9 show results on the DAVIS 2017 validation set [72] (30 videos, 67.4 frames average length, 2.03 objects per video on average) compared to SOTA methods. Methods are ranked by the mean of \mathcal{J} and \mathcal{F} . Siam R-CNN significantly outperforms the previous best method that only uses the first-frame bounding boxes, SiamMask [94], by 14.8 percentage points. To evaluate how much of this improvement comes from Box2Seg and how much from our tracking, we applied Box2Seg to the output of SiamMask. This does improve the results while still being 7.3 percentage points worse than our method. We also run SiamRPN++ [48] and DiMP-50 [5] with Box2Seg for comparison. As a reference for the achievable performance for our tracker, we ran Box2Seg on the ground truth boxes

Dataset	Speed	OTB2015	LaSOT	LTB35
Eval measure	FPS	AUC	AUC	F
Siam R-CNN	4.7	70.1	64.8	66.8
No hard ex.	4.7	68.4	63.2	66.5
Argmax	4.9	63.8	62.9	65.5
Short-term	4.6	67.2	55.7	57.2
ResNet-50	5.1	68.0	62.3	64.4
$\frac{1}{2}$ res.	5.7	70.2	62.9	65.6
100 RoIs	8.7	68.7	64.1	67.3
ResNet-50 + $\frac{1}{2}$ res.	6.1	68.6	61.2	64
ResNet-50 + 100 RoIs	10.3	66.9	61.5	65.9
$\frac{1}{2}$ res. + 100 RoIs	13.6	69.1	63.2	66.0
ResNet-50 + $\frac{1}{2}$ res. + 100 RoIs	15.2	67.7	61.1	63.7

Table 7: Ablation and timing analysis of different components of Siam R-CNN. A V100 GPU is used for timing.

which resulted in a score of 82.6%.

Even without using the first-frame mask, Siam R-CNN outperforms many methods that use the mask such as RGMP [100] and VideoMatch [37], and even methods that perform slow first-frame fine-tuning such as OSVOS-S [64] and OnAVOS [89]. Our method is also more practical, as it is far more tedious to create a perfect first-frame segmentation mask by hand than a bounding box initialization. If the first-frame mask is available, then we are able to fine-tune Box2Seg on this, improving results by 4.2 percentage points at the cost of speed. We report on further evaluations on the DAVIS17 test-dev benchmark and on DAVIS16 [71] in the supplemental material.

YouTube-VOS. Table 6 and Figure 10 show the results of our method on the YouTube-VOS 2018 benchmark [101] (474 videos, 26.6 frames average length, 1.89 objects per video on average) compared to six SOTA methods. The evaluation distinguishes between classes which are in the training set (*seen*) and those that are not (*unseen*). Methods are ranked by the overall score \mathcal{O} , which is the mean of the \mathcal{J} and \mathcal{F} metric over both seen and unseen classes. Siam R-CNN again outperforms all methods which do not use the first-frame mask (by 15.5 percentage points), and also outperforms PReMVOS [60, 59] and all other previous methods except for STM-VOS [69]. The version of our method using the fine-tuned Box2Seg improves the results by 4.9 percentage points.

4.4. Ablation and Timing Analysis

Table 7 shows a number of ablations of Siam R-CNN on three datasets together with their speed (using a V100 GPU). Siam R-CNN runs at 4.7 frames per second (FPS) using a ResNet-101 backbone, 1000 RPN proposals per frame, and TDPA.

The row “No hard ex.” shows the results without hard example mining (c.f. Sec. 3.2). Hard example mining improves results on all datasets, by up to 1.7 percentage points.

We compare TDPA to using just the highest scoring

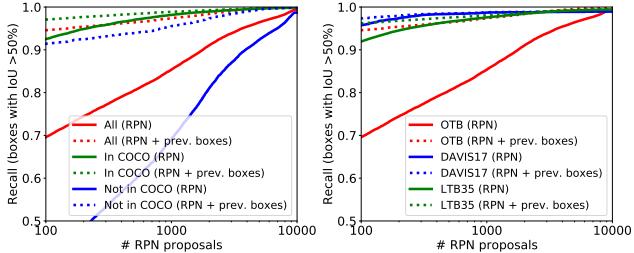


Figure 11: Recall of our RPN with varying number of proposals. Dotted lines have up to 100 re-detections from the previous frame added. Left: comparison on a split of OTB2015 for object classes present and not present in COCO. Right: comparison over three datasets.

re-detection in each frame (“Argmax”) and the short-term algorithm we used for the reset-based VOT18 evaluation (“Short-term”). TDPA outperforms both of these in all datasets. Argmax is slightly faster, as it does not re-detect using previous predictions as reference. For the long-term datasets, Argmax significantly outperforms both the short-term variant and even all previous methods, highlighting the strength of tracking by re-detection for long-term tracking.

4.5. Making Siam R-CNN Even Faster

Table 7 also shows the result of three improvements aimed at increasing the speed of Siam R-CNN (smaller backbone, smaller input resolution, and less RoI proposals).

When evaluating with a ResNet-50 backbone, Siam R-CNN performs slightly faster and still achieves SOTA results (62.3 on LaSOT, compared to 56.8 for DiMP-50 with the same backbone). This shows that the results are not only due to a larger backbone, but due to our tracking by re-detection approach.

We also evaluate reducing the image input size to a smaller image edge length of 400 pixels instead of 800 (row “ $\frac{1}{2}$ res”). This also results in only a slight decrease in performance in two benchmarks, and a slight increase in performance on OTB2015.

The row “100 RoIs” shows the results of using 100 RoIs from the RPN, instead of 1000. This almost doubles the speed as most compute occurs in the re-detection head. This results in only a small score decrease on two benchmarks, while improving results on LTB35. This shows that Siam R-CNN can run very quickly, even though it is based on a two-stage detection architecture, as very few RoIs are required.

The fastest setup with all three of these speed improvements (ResNet-50 + 100 RoIs + $\frac{1}{2}$ res.) achieves 15.2 frames per second with a V100 GPU, but still achieves strong results, especially for long-term tracking. The same setup with a ResNet-101 backbone instead of ResNet-50 (100 RoIs + $\frac{1}{2}$ res.) runs at 13.6 frames per second and achieves excellent results and loses at most 1.6 percentage points over these three datasets compared to the standard Siam R-CNN, while running almost three times as fast.

4.6. Generic Object Tracking Analysis

Siam R-CNN should be able to track any generic object. However, its backbone and RPN have been trained only on 80 object classes in COCO and have then been frozen when training for re-detection. In Fig. 11, we investigate the recall of our RPN on the 44% of OTB2015 sequences that contain objects not in COCO, compared to the rest. With the default of 1000 proposals, the RPN achieves only 69.1% recall for unknown objects, compared to 98.2% for known ones. One solution is to increase the number of proposals used. When using 10,000 proposals the RPN achieves 98.7% recall for unknown objects but causes Siam R-CNN to run much slower (around 1 FPS). Our solution is to instead include the previous-frame re-detections (up to 100) as additional proposals. This increases the recall to 95.5% for unknown objects when using 1000 RPN proposals and even results in adequate unknown object recall of 91.4% when using only 100 RPN proposals. This shows why Siam R-CNN is able to outperform all previous methods on OTB2015, even though almost half of the objects are not from COCO classes. We also run a recall analysis on the DAVIS17 and LTB35 datasets where nearly all objects belong to COCO classes and we achieve excellent recall (see Fig. 11 right).

5. Conclusion

We introduce Siam R-CNN as a Siamese two-stage full-image re-detection architecture with a Tracklet Dynamic Programming Algorithm. Siam R-CNN outperforms all previous methods on ten tracking benchmarks, with especially strong results for long-term tracking. We hope that our work will inspire future work on using two-stage architectures and full-image re-detection for tracking.

Acknowledgements: For partial funding of this project, PV, JL and BL would like to acknowledge the ERC Consolidator Grant DeeViSe (ERC-2017-COG-773161) and a Google Faculty Research Award. PHST would like to acknowledge CCAV project Streetwise and EPSRC/MURI grant EP/N019474/1. The authors would like to thank Sourabh Swain, Yuxin Wu, Goutam Bhat, and Bo Li for helpful discussions.

References

- [1] S. Avidan. Support vector tracking. *PAMI*, 2004. 1, 2
- [2] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. Robust object tracking with online multiple instance learning. *PAMI*, 2011. 1, 2
- [3] Linchao Bao, Baoyuan Wu, and Wei Liu. CNN in MRF: video object segmentation via inference in a cnn-based higher-order spatio-temporal MRF. In *CVPR*, 2018. 2, 7, 16, 17
- [4] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr. Fully-convolutional siamese networks for object tracking. In *ECCVW*, 2016. 1, 7, 16
- [5] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, 2019. 2, 6, 7, 8, 16, 18
- [6] Goutam Bhat, Joakim Johnander, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Unveiling the power of deep tracking. In *ECCV*, 2018. 6, 16, 18, 19
- [7] David S. Bolme, J. Ross Beveridge, Bruce A. Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, 2010. 2
- [8] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*, 2017. 2, 7, 8, 16, 17
- [9] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018. 3, 5
- [10] Boyu Chen, Dong Wang, Peixia Li, Shuang Wang, and Huchuan Lu. Real-time 'actor-critic' tracking. In *ECCV*, 2018. 15, 16, 18
- [11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 5
- [12] Y. Chen, J. Pont-Tuset, A. Montes, and L. Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *CVPR*, 2018. 2, 17
- [13] J. Cheng, Y-H. Tsai, W-C. Hung, S. Wang, and M-H. Yang. Fast and accurate online video object segmentation via tracking parts. In *CVPR*, 2018. 2, 7, 16, 17
- [14] Jongwon Choi, Hyung Jin Chang, Tobias Fischer, Sangdoo Yun, Kyuewang Lee, Jiyeoup Jeong, Yiannis Demiris, and Jin Young Choi. Context-aware deep feature compression for high-speed visual tracking. In *CVPR*, 2018. 18
- [15] Janghoon Choi, Junseok Kwon, and Kyoung Mu Lee. Deep meta learning for real-time target-aware visual tracking. In *ICCV*, 2019. 18
- [16] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017. 5
- [17] Kenan Dai, Dong Wang, Huchuan Lu, Chong Sun, and Jianhua Li. Visual tracking via adaptive spatially-regularized correlation filters. In *CVPR*, 2019. 2, 16, 18
- [18] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ECO: Efficient convolution operators for tracking. In *CVPR*, 2017. 2, 6, 7, 16
- [19] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: accurate tracking by overlap maximization. In *CVPR*, 2019. 2, 6, 7, 18
- [20] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *ECCV*, 2016. 2, 6
- [21] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *ECCV*, 2018. 18
- [22] Xingping Dong, Jianbing Shen, Wenguan Wang, Yu Liu, Ling Shao, and Fatih Porikli. Hyperparameter optimization for tracking with continuous deep q-learning. In *CVPR*, 2018. 18
- [23] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Si-jia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. LaSOT: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 2019. 2, 5, 6, 7
- [24] Heng Fan and Haibin Ling. Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. In *ICCV*, 2017. 16
- [25] Heng Fan and Haibin Ling. Siamese cascaded region proposal networks for real-time visual tracking. In *CVPR*, 2019. 2, 18
- [26] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 3
- [27] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Graph convolutional tracking. In *CVPR*, 2019. 18
- [28] Helmut Grabner, Michael Grabner, and Horst Bischof. Real-time tracking via on-line boosting. In *BMVC*, 2006. 1, 2
- [29] Qing Guo, Wei Feng, Ce Zhou, Rui Huang, Liang Wan, and Song Wang. Learning dynamic siamese network for visual object tracking. In *ICCV*, 2017. 7
- [30] Sam Hare, Stuart Golland, Amir Saffari, Vibhav Vineet, Ming-Ming Cheng, Stephen L Hicks, and Philip H.S. Torr. Struck: Structured output tracking with kernels. *PAMI*, 2015. 1, 2
- [31] Anfeng He, Chong Luo, Xinmei Tian, and Wenjun Zeng. A twofold siamese network for real-time object tracking. In *CVPR*, 2018. 16, 18
- [32] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *ICCV*, 2019. 5
- [33] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 3
- [34] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5
- [35] Joo F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *PAMI*, 2015. 2
- [36] Alexander Hermans*, Lucas Beyer*, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*, 2017. 3
- [37] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G. Schwing. Videomatch: Matching based video object segmentation. In *ECCV*, 2018. 2, 7, 8, 16, 17

- [38] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *arXiv preprint arXiv:1810.11981*, 2018. 2, 5, 6, 15
- [39] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Bridging the gap between detection and tracking: A unified approach. In *ICCV*, 2019. 2, 6, 18
- [40] Ziyuan Huang, Changhong Fu, Yiming Li, Fuling Lin, and Peng Lu. Learning aberrance repressed correlation filters for real-time uav tracking. In *ICCV*, 2019. 18
- [41] Ilchae Jung, Jeany Son, Mooyeon Baek, and Bohyung Han. Real-time mdnet. In *ECCV*, 2018. 18
- [42] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *PAMI*, 2012. 1, 2, 7
- [43] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *ICCV*, 2017. 2, 6, 15
- [44] Matej Kristan, Aleš Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Čehovin Zajc, Tomas Vojir, Gustav Häger, Alan Lukežič, and Gustavo Fernandez. The visual object tracking vot2016 challenge results. In *ECCVW*, 2016. 5, 15, 16
- [45] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pfugfelder, Luka Cehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Elde索key, Gustavo Fernandez, and et al. The sixth visual object tracking VOT2018 challenge results. In *ECCVW*, 2018. 2, 5, 6, 7, 14, 15
- [46] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Čehovin, G. Fernández, T. Vojir, G. Häger, G. Nebehay, R. Pfugfelder, A. Gupta, and et al. The visual object tracking VOT2015 challenge results. In *ICCVW*, 2015. 2, 5, 6, 15, 16
- [47] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pfugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *PAMI*, 2016. 2
- [48] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 2019. 1, 2, 6, 7, 8, 16, 18, 19
- [49] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, 2018. 1, 2, 16, 18
- [50] Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, and Ming-Hsuan Yang. Learning spatial-temporal regularized correlation filters for visual tracking. In *CVPR*, 2018. 16, 18
- [51] Peixia Li, Boyu Chen, Wanli Ouyang, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Gradnet: Gradient-guided network for visual object tracking. In *ICCV*, 2019. 18
- [52] Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *ECCV*, 2018. 2, 7, 16, 17
- [53] Xin Li, Chao Ma, Baoyuan Wu, Zhenyu He, and Ming-Hsuan Yang. Target-aware deep tracking. In *CVPR*, 2019. 18
- [54] Pengpeng Liang, Erik Blasch, and Haibin Ling. Encoding color information for visual tracking: Algorithms and benchmark. *Trans. Image Proc.*, 2015. 5, 15, 16
- [55] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 5
- [56] T-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3, 5
- [57] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 2
- [58] Xiankai Lu, Chao Ma, Bingbing Ni, Xiaokang Yang, Ian Reid, and Ming-Hsuan Yang. Deep regression tracking with shrinkage loss. In *ECCV*, 2018. 18
- [59] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. PReMVOS: Proposal-generation, refinement and merging for the YouTube-VOS challenge on video object segmentation 2018. *ECCVW*, 2018. 8
- [60] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. PReMVOS: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, 2018. 2, 3, 5, 7, 8, 16, 17
- [61] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. PReMVOS: Proposal-generation, refinement and merging for video object segmentation. *The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2018. 7
- [62] Alan Lukežič, Luka Čehovin Zajc, Tomáš Vojíř, Jiří Matas, and Matej Kristan. Now you see me: evaluating performance in long-term visual tracking. *arXiv preprint arXiv:1804.07056*, 2018. 2, 6, 7, 19
- [63] Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. Hierarchical convolutional features for visual tracking. In *ICCV*, 2015. 2
- [64] K.-K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal Taixé, and L. Van Gool. Video object segmentation without temporal information. *PAMI*, 2018. 2, 7, 8, 16, 17
- [65] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *ECCV*, 2016. 2, 5, 6, 15, 16
- [66] Matthias Müller, Adel Bibi, Silvio Giancola, Salman Al-Subaihi, and Bernard Ghanem. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, 2018. 2, 5, 6, 15
- [67] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, 2016. 2, 6, 7
- [68] Gerhard Neuhold, Tobias Ollmann, S Rota Bulo, and Peter Kortscheder. The Mapillary Vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 5
- [69] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 2, 7, 8, 16, 17
- [70] Eunbyung Park and Alexander C. Berg. Meta-tracker: Fast and robust online adaptation for visual object trackers. In *ECCV*, 2018. 18

- [71] 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 *CVPR*, 2016. 2, 8, 16
- [72] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2, 7, 8, 16, 17, 19
- [73] Liangliang Ren, Xin Yuan, Jiwen Lu, Ming Yang, and Jie Zhou. Deep reinforcement learning with iterative shift for visual tracking. In *ECCV*, 2018. 18
- [74] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 2, 5
- [75] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *IJCV*, 2015. 5
- [76] Amir Saffari, Martin Godec, Thomas Pock, Christian Leistner, and Horst Bischof. Online multi-class lpboost. In *CVPR*, 2010. 1, 2
- [77] Amir Saffari, Christian Leistner, Jakob Santner, Martin Godec, and Horst Bischof. On-line random forests. In *ICCVW*, 2009. 1, 2
- [78] Axel Sauer, Elie Aljalbout, and Sami Haddadin. Tracking holistic object representations. In *BMVC*, 2019. 2, 6
- [79] Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 3
- [80] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson Lau, and Ming-Hsuan Yang. VITAL: Visual tracking via adversarial learning. In *CVPR*, 2018. 7, 18
- [81] Chong Sun, Dong Wang, Huchuan Lu, and Ming-Hsuan Yang. Correlation tracking via joint discrimination and reliability learning. In *CVPR*, 2018. 16, 18
- [82] Chong Sun, Dong Wang, Huchuan Lu, and Ming-Hsuan Yang. Learning spatial-aware regressions for visual tracking. In *CVPR*, 2018. 18
- [83] Yuxuan Sun, Chong Sun, Dong Wang, You He, and Huchuan Lu. Roi pooled correlation filters for visual tracking. In *CVPR*, 2019. 16, 18
- [84] Ming Tang, Bin Yu, Fan Zhang, and Jinqiao Wang. High-speed tracking with multi-kernel correlation filters. In *CVPR*, 2018. 18
- [85] Ardhendu Shekhar Tripathi, Martin Danelljan, Luc Van Gool, and Radu Timofte. Tracking the known and the unknown by leveraging semantic information. In *BMVC*, 2019. 2, 6
- [86] Jack Valmadre, Luca Bertinetto, João F. Henriques, Ran Tao, Andrea Vedaldi, Arnold W. M. Smeulders, Philip H. S. Torr, and Efstratios Gavves. Long-term tracking in the wild: A benchmark. In *ECCV*, 2018. 2, 6, 7
- [87] Paul Voigtlaender, Yunling Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. FEELVOS: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, 2019. 2, 7, 16, 17
- [88] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for the 2017 DAVIS challenge on video object segmentation. *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2017. 7, 8
- [89] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *BMVC*, 2017. 2, 7, 8, 16, 17
- [90] Guangting Wang, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. Spm-tracker: Series-parallel matching for real-time visual object tracking. In *CVPR*, 2019. 6, 16, 18
- [91] Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, and Houqiang Li. Unsupervised deep tracking. In *CVPR*, 2019. 18
- [92] Ning Wang, Wengang Zhou, Qi Tian, Richang Hong, Meng Wang, and Houqiang Li. Multi-cue correlation filters for robust visual tracking. In *CVPR*, 2018. 16, 18
- [93] Qiang Wang, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, and Stephen Maybank. Learning attentions: Residual attentional siamese network for high performance online visual tracking. In *CVPR*, 2018. 18
- [94] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019. 2, 7, 8, 16, 17, 18, 19
- [95] Xiao Wang, Chenglong Li, Bin Luo, and Jin Tang. Sint++: Robust visual tracking via adversarial positive instance generation. In *CVPR*, 2018. 16, 18
- [96] Yuxin Wu et al. Tensorpack. <https://github.com/tensorpack/>, 2016. 5
- [97] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 5
- [98] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, 2013. 2, 5, 15, 16
- [99] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *PAMI*, 2015. 2, 5, 15, 16, 19
- [100] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR*, 2018. 2, 7, 8, 16, 17
- [101] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. YouTube-VOS: Sequence-to-sequence video object segmentation. In *ECCV*, 2018. 2, 5, 7, 8
- [102] Tianyang Xu, Zhen-Hua Feng, Xiao-Jun Wu, and Josef Kittler. Joint group feature selection and discriminative filter learning for robust visual object tracking. In *ICCV*, 2019. 15, 16, 18
- [103] Tianyang Xu, Zhen-Hua Feng, Xiao-Jun Wu, and Josef Kittler. Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. *Trans. Image Proc.*, 2019. 6
- [104] Bin Yan, Haojie Zhao, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 'Skimming-Perusal' Tracking: A framework for real-time and robust long-term tracking. In *ICCV*, 2019. 7, 18

- [105] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K. Katsaggelos. Efficient video object segmentation via network modulation. In *CVPR*, 2018. 2, 7, 16, 17
- [106] Tianyu Yang and Antoni B. Chan. Learning dynamic memory networks for object tracking. In *ECCV*, 2018. 18
- [107] Yingjie Yao, Xiaohe Wu, Lei Zhang, Shiguang Shan, and Wangmeng Zuo. Joint representation and truncated inference learning for correlation filter based tracking. In *ECCV*, 2018. 16, 18
- [108] Donghun Yeo, Jeany Son, Bohyung Han, and Joon Hee Han. Superpixel-based tracking-by-segmentation using markov chains. In *CVPR*, 2017. 2
- [109] Lichao Zhang, Abel Gonzalez-Garcia, Joost van de Weijer, Martin Danelljan, and Fahad Shahbaz Khan. Learning the model update for siamese trackers. In *ICCV*, 2019. 16, 18
- [110] Mengdan Zhang, Qiang Wang, Junliang Xing, Jin Gao, Peixi Peng, Weiming Hu, and Steve Maybank. Visual tracking via spatially aligned correlation filters network. In *ECCV*, 2018. 16, 18
- [111] Yunhua Zhang, Dong Wang, Lijun Wang, Jinqing Qi, and Huchuan Lu. Learning regression and verification networks for long-term visual tracking. *arXiv preprint arXiv:1809.04320*, 2018. 7
- [112] Yunhua Zhang, Lijun Wang, Jinqing Qi, Dong Wang, Mengyang Feng, and Huchuan Lu. Structured siamese network for real-time visual tracking. In *ECCV*, 2018. 7
- [113] Yunhua Zhang, Lijun Wang, Jinqing Qi, Dong Wang, Mengyang Feng, and Huchuan Lu. Structured siamese network for real-time visual tracking. In *ECCV*, 2018. 18
- [114] Zhipeng Zhang, Houwen Peng, and Qiang Wang. Deeper and wider siamese networks for real-time visual tracking. In *CVPR*, 2019. 2, 16, 18
- [115] Linyu Zheng, Ming Tang, Yingying Chen, Jinqiao Wang, and Hanqing Lu. Fast-deepkcf without boundary effect. In *ICCV*, 2019. 18
- [116] Linyu Zheng, Ming Tang, Hanqing Lu, et al. Learning features with differentiable closed-form solver for tracking. *arXiv preprint arXiv:1906.10414*, 2019. 6
- [117] Pengfei Zhu, Longyin Wen, Xiao Bian, Ling Haibin, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018. 2
- [118] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *ECCV*, 2018. 2, 3, 5, 6, 16, 18
- [119] Zheng Zhu, Wei Wu, Wei Zou, and Junjie Yan. End-to-end flow correlation tracking with spatial-temporal attention. In *CVPR*, 2018. 16, 18

Supplemental Material for Siam R-CNN: Visual Tracking by Re-Detection

Abstract

We provide more details of Siam RCNN’s training procedure, a more detailed description of the video hard example mining procedure, and of the short-term tracking algorithm. Finally, we present additional quantitative and qualitative results.

A. Further Method Details

A.1. Training

We train Siam R-CNN with random image scale sampling by scaling the small edge of the image to a random value between 640 and 800 pixels, while keeping the aspect ratio. In cases where this resizing would result in a larger edge size of more than 1333 pixels, it is resized to a larger edge size of 1333 pixels instead. During test time we resize the image to a smaller edge length of 800 pixels, again keeping the longer edge size no larger than 1333 pixels. Note that these settings are the default of the Mask R-CNN implementation we used.

We train our network with two NVIDIA GTX 1080 Ti GPUs for 1 million steps (5.8 days) with a learning rate of 0.01, and afterwards for 120,000 steps and 80,000 steps with learning rates of 0.001 and 0.0001, respectively. The hard example training is done afterwards with a single GPU for 160,000 steps and a learning rate of 0.001. A batch size of one pair of images (reference and target) per GPU is used.

A.2. Video Hard Example Mining

Index Structure. For the indexing structure, we use the “Approximate Nearest Neighbors Oh Yeah” library for approximate nearest neighbor queries¹.

Feature Pre-computation. During normal training without hard negative examples, the RoIs given to the second stage are generated automatically by the RPN and are thus not always perfectly aligned to the object boundaries. In the three cascade stages, the ROI will then be successively refined by bounding box regression. However, when naively pre-computing the features for the ground truth bounding boxes, the network might overfit to these perfect boxes.

In order for the network to learn to handle imperfect bounding boxes, we add random Gaussian noise to the ground truth bounding boxes before pre-computing the features, and afterwards run these jittered RoIs through the cas-

cade stages and also pre-compute the re-aligned features after every cascade stage.

In particular, we take the ground truth bounding box (x_0, y_0, x_1, y_1) and independently add to each component noise sampled from a clipped Gaussian with mean 0 and standard deviation 0.25, *i.e.*, for $i \in \{0, 1, 2, 3\}$, we have

$$\tilde{r}_i \sim \mathcal{N}(0, 0.25), \quad (3)$$

$$r_i = \text{clip}(\tilde{r}_i, -0.25, 0.25). \quad (4)$$

The jittered box is then given by

$$(x_0 + r_0, y_0 + r_1, x_1 + r_2, y_1 + r_3). \quad (5)$$

Training Procedure. Having pre-computed the ROI-aligned features for each object and each cascade stage, the training procedure now works as follows. For each training step, as usual, a random video and object in this video is selected and then a random reference and a random target frame. Afterwards, we use the indexing structure to retrieve the 10,000 nearest neighbor bounding boxes from other videos (50,000 boxes for LaSOT because of the long sequences). Note that the nearest neighbors are searched for over all training datasets, regardless where the reference comes from. Since the nearest neighbors are found per frame, often a few videos will dominate the set of nearest neighbors. To get more diverse negative examples, we create a list of all videos (excluding the reference) in which nearest neighbor boxes were found and randomly select 100 of these videos. For each of the 100 videos, we then randomly select 1 of the boxes which were retrieved as nearest neighbors from this video and add them as negative examples for the re-detection head.

Adding only additional negative examples creates an imbalance in the training data. Hence, we also retrieve the features of the ground truth bounding boxes of 30 randomly selected frames of the current reference video as additional positive examples.

A.3. Short-term Tracking Algorithm

For the VOT2018 dataset [45], it is standard to use a reset-based evaluation, where once the object is lost (0 IoU between predicted and ground truth box), the tracker is restarted with the ground truth box five frames later and receives a penalty. This extreme short-term tracking scenario is not what Siam R-CNN with the TDPA was designed for. It often triggers resets, which normally (without reset-based evaluation) Siam R-CNN could have automatically recovered from.

¹<https://github.com/spotify/annoy>

Algorithm 2 Perform Short-term Tracking for time-step t

```

1: Inputs ff_gt_feats, image $_t$ , det $_{t-1}$ 
2: backbone_feats  $\leftarrow$  backbone(image $_t$ )
3: RoIs  $\leftarrow$  RPN(backbone_feats)
4:  $\triangleright$  Add shifted versions of det $_{t-1}$  to RoIs
5: for shift $_x \in \{-1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5\} do
6:   for shift $_y \in \{-1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5\} do
7:     RoIs  $\leftarrow$  RoIs  $\cup$  {shift(det $_{t-1}$ , shift $_x$ , shift $_y$ )}
8:   end for
9: end for
10: dets $_t$ , det_scores $_t$   $\leftarrow$  redetection_head(RoIs, ff_gt_feats)
11: prev_scores $_t$   $\leftarrow$  score_redetection(dets $_t$ , det $_{t-1}$ )
12: loc_scores $_t$   $\leftarrow$  |bbox(dets $_t$ ) - bbox(det $_{t-1}$ )| $_1$ 
13: scores $_t$   $\leftarrow$  det_scores $_t$  + prev_scores $_t$  +  $\delta \cdot$  loc_scores $_t$ 
14:  $\triangleright$  Filter out detections with too large spatial distance
15: for det $_t \in$  dets $_t$  do
16:   if ||det $_t$  - det $_{t-1}$ || $_\infty > \xi$  then
17:     scores[det $_t$ ]  $\leftarrow$   $-\infty$ 
18:   end if
19: end for
20: return arg max $_{\text{det}_t \in \text{dets}_t}$  scores $_t$ [det $_t$ ]$$ 
```

Since VOT2018 is an important tracking benchmark, we created a short-term version of the Siam R-CNN tracking algorithm (see Alg. 2). Given the ROI Aligned features ff_gt_feats of the first-frame bounding box and the previous-frame tracking result det $_{t-1}$, we first extract the backbone features of the current image and afterwards generate regions of interest (RoIs) using the region proposal network (RPN, lines 1–3).

Note, that we know for sure that the previous-frame predicted box det $_{t-1}$ has a positive IoU with the ground truth box, as otherwise a reset would have been triggered. Hence, the object to be tracked should be located close to the previous-frame predicted box. In order to exploit this, and to compensate for potential false negatives of the RPN, we add shifted versions of the previous-frame prediction as additional RoIs (lines 5–9). Here, the function shift(.) shifts the previous-frame box det $_{t-1}$ by factors of its width and height, *e.g.*, if shift $_x = 0.5$ and shift $_y = 1.0$, the box is shifted by half its width in x-direction and by its full height in y-direction.

Afterwards, we use the redetection_head to produce detections det $_t$ with detection scores det_scores $_t$ for the current frame t (line 10). We then additionally compute previous-frame scores prev_scores $_t$ by using the previous-frame box as a reference to score the current detections (line 11). To exploit spatial consistency, we also compute location scores (loc_scores $_t$, line 12), given by the L_1 -norm of the pairwise differences between the current detection boxes and the previous-frame predicted box. All three scores are then combined (line 13) by a linear combination, where the current-frame and previous-frame scores have equal weight.

Even when using a location score, it can happen, that

a distractor object appears far away from the object to be tracked and gets a high combined score because it looks very similar to that object. However, since we know that the previous-frame box has positive overlap with the ground truth box, a far-away detection cannot be the object to be tracked. Hence, we explicitly filter out detections which have a large spatial distance (measured by the L_∞ -norm) to the previous-frame predicted box (lines 15–19).

Finally, we report the detection with the highest combined score as the result for the current frame (line 20), or in case there is no valid detection, we repeat the previous-frame result as result for the current frame.

B. Further Experimental Results

In addition to the 11 benchmarks that we presented in the main paper, we present here results on eight further benchmarks, of which five are short-term tracking benchmarks, one is a long-term tracking benchmark and two are video object segmentation benchmarks. For all benchmarks (except for the three VOT benchmarks) we use exactly the same tracking parameters. This is in contrast to many other methods which have parameters explicitly tuned for each dataset. This shows the generalization ability of our tracker to many different scenarios. For the three VOT datasets we use the short-term variant of the tracking parameters (with the same parameters across these three benchmarks).

B.1. Further Short-Term Tracking Evaluation

In the main paper we presented short-term tracking results on OTB2015 [99], UAV123 [65], NFS [43], TrackingNet [66], VOT2018 (the same as VOT2017) [45], and GOT-10k [38]. Here in the supplemental material we present results on five further short-term tracking benchmarks. These further benchmarks are OTB-50 [99], OTB2013 [98], VOT2015 [46], VOT2016 [44], and TempleColor128 (TC128) [54].

OTB-50. We evaluate on the OTB-50 benchmark [99] (50 videos, 539 frames average length). This dataset is a subset of OTB2015 using exactly half of the sequences. It is evaluated with the same evaluation measures as OTB2015. Tab. 8 compares our results to five SOTA trackers. Siam R-CNN achieves 66.3 AUC, which outperforms the previous best published results by ACT [10] by 0.6 percentage points.

OTB2013. We evaluate on the OTB2013 benchmark [98] (51 videos, 578 frames average length). This dataset is a predecessor to OTB2015 and is evaluated with the same evaluation measures. Tab. 9 compares our results to five SOTA trackers. Siam R-CNN achieves 70.4 AUC, which is comparable to SOTA trackers while being 1.8 percentage points behind the best published results by GFS-DCF [102].

TC128. We evaluate on the TempleColor128 (TC128)

	SA-Siam [31]	SINT++ [95]	RTINet [107]	SPM [90]	ACT [10]	Siam R-CNN
Success AUC	61.0	62.4	63.7	65.3	65.7	66.3

Table 8: Results on OTB-50 [99].

	RPCF [83]	SACF [110]	MCCT [92]	DRT [81]	GFS-DCF [102]	Siam R-CNN
Success AUC	71.3	71.3	71.4	72.0	72.2	70.4

Table 9: Results on OTB2013 [98].

	MCCT [92]	STRCF [50]	RTINet [107]	ASRCF [17]	UPDT [6]	Siam R-CNN
Success AUC	59.6	60.1	60.2	60.3	62.2	60.1

Table 10: Results on TC128 [54].

	FlowTrack [119]	SACF [110]	SiamRPN [49]	SiamDW [114]	DaSiamRPN [118]	Ours (short-t.)
EAO	34.1	34.3	35.8	38.0	44.6	45.4

Table 11: Results on VOT2015 [46].

	DaSiamRPN [118]	SPM [90]	DRT [81]	SiamMask [94]	UpdateNet [109]	Ours (short-t.)
EAO	41.1	43.4	44.2	44.2	48.1	46.5

Table 12: Results on VOT2016 [44].

	SiamFC [4]	PTAV [24]	ECO [18]	SiamRPN [49]	DaSiamRPN [118]	Siam R-CNN
Success AUC	39.9	42.3	43.5	45.4	61.7	67.2

Table 13: Results on UAV20L [65].

benchmark [54] (128 videos, 429 frames average length). This dataset is also evaluated using the OTB2015 evaluation measures. Tab. 10 compares our results to five SOTA trackers. Siam R-CNN achieves 60.1 AUC, which is comparable to SOTA trackers while being slightly inferior to the best published results by UPDT [6] by 2.1 percentage points.

VOT2015. We evaluate on the VOT2015 benchmark [46] (60 videos, 358 frames average length). This is evaluated with the same evaluation measures as VOT2018. Tab. 11 compares our results to five SOTA trackers. The short-term version of Siam R-CNN achieves 45.4 EAO, which outperforms the previous best published results by DaSiamRPN [118] by 0.8 percentage points.

VOT2016. We evaluate on the VOT2016 benchmark [44] (60 videos, 358 frames average length). This is also evaluated with the same evaluation measures as VOT2018. Tab. 12 compares our results to five SOTA trackers. The short-term version of Siam R-CNN achieves 46.5 EAO, which outperforms all previous published results except those of UpdateNet [109] which outperforms our results by 1.6 percentage points.

Init	Method	FT	M	$\mathcal{J} \& \mathcal{F}$	\mathcal{J}	\mathcal{F}	\mathcal{J}_{box}	t(s)
bbox	Siam R-CNN (ours)	x	x	70.6	66.1	75.0	78.3	0.32
	Siam R-CNN (fastest)	x	x	70.5	66.4	74.6	76.9	0.12
	SiamMask [94]	x	x	55.8	54.3	58.5	64.3	0.06 [†]
	SiamMask [94] (Box2Seg)	x	x	63.3	59.5	67.3	64.3	0.11
	SiamRPN++ [48] (Box2Seg)	x	x	61.6	56.8	66.3	64.0	0.11
	DiMP-50 [5] (Box2Seg)	x	x	63.7	60.1	67.3	65.6	0.10
mask	STM-VOS [69]	x	✓	81.8	79.2	84.3	—	0.32 [†]
	FEELVOS [87]	x	✓	71.5	69.1	74.0	71.4	0.51
	RGMP [100]	x	✓	66.7	64.8	68.6	66.5	0.28 [†]
	VideoMatch [37]	x	✓	62.4	56.5	68.2	—	0.35
	FAVOS [13]	x	✓	58.2	54.6	61.8	68.0	1.2 [†]
	OSMN [105]	x	✓	54.8	52.5	57.1	60.1	0.28 [†]
mask+ft	PReMVOS [60]	✓	✓	77.8	73.9	81.7	81.4	37.6
	Ours (Fine-t. Box2Seg)	✓	✓	74.8	69.3	80.2	78.3	1.0
	DyeNet [52]	✓	✓	74.1	—	—	—	9.32 [†]
	OSVOS-S [64]	✓	✓	68.0	64.7	71.3	68.4	9 [†]
	CINM [3]	✓	✓	67.5	64.5	70.5	72.9	>120
	OnAVOS [89]	✓	✓	63.6	61.0	66.1	66.3	26
	OSVOS [8]	✓	✓	60.3	56.6	63.9	57.0	18 [†]
	GT boxes (Box2Seg)	x	x	82.6	79.3	85.8	100.0	—
	GT boxes (Fine-t. Box2Seg)	✓	✓	86.2	81.8	90.5	100.0	—

Table 14: Results on the DAVIS 2017 validation set. FT: fine-tuning, M: using the first-frame masks, t(s): time per frame in seconds. [†]: timing extrapolated from DAVIS 2016 assuming linear scaling in the number of objects. Siam R-CNN (fastest) denotes Siam R-CNN with ResNet-50 backbone, half input resolution, and 100 RoIs from the RPN, see Section 4.5.

B.2. Further Long-Term Tracking Evaluation

We evaluate on one further long-term tracking dataset, in addition to the three benchmarks presented in the main paper.

UAV20L. We evaluate on the UAV20L benchmark [65] (20 videos, 2934 frames average length). This dataset contains 20 of the 123 sequences of UAV123, however each of these sequences extends for many more frames than the equivalent sequence in the UAV123 version. It is also evaluated with the same evaluation measures as OTB2015. Tab. 13 compares our results to five SOTA trackers. Siam R-CNN achieves 67.2 AUC, which outperforms the previous best published results by DaSiamRPN [118] by 5.5 percentage points, which further highlights the ability of Siam R-CNN to perform long-term tracking.

B.3. Further Video Object Segmentation Evaluation

Table 14 is an extended version of the results on the DAVIS 2017 [72] validation set shown in the main paper.

Table 15 shows results on the DAVIS 2016 validation set [71] (20 videos, 68.8 frames average length, 1 object

	Init	Method	FT	M	$\mathcal{J} \& \mathcal{F}$	\mathcal{J}	\mathcal{F}	\mathcal{J}_{box}	t(s)
bbox	Siam R-CNN (ours)	\times	\times	78.6	76.8	80.4	86.6	0.24	
	Siam R-CNN (fastest)	\times	\times	79.0	77.4	80.6	85.0	0.08	
	SiamMask [94]	\times	\times	69.8	71.7	67.8	73.3	0.03	
	SiamMask [94] (Box2Seg)	\times	\times	75.9	75.6	76.3	73.3	0.06	
mask	STM-VOS [69]	\times	\checkmark	89.3	88.7	89.9	—	0.16	
	RGMP [100]	\times	\checkmark	81.8	81.5	82.0	79.3	0.14	
	FEELVOS [87]	\times	\checkmark	81.7	81.1	82.2	80.2	0.45	
	FAVOS [13]	\times	\checkmark	81.0	82.4	79.5	83.1	0.6	
	VideoMatch [37]	\times	\checkmark	80.9	81.0	80.8	—	0.32	
	PML [12]	\times	\checkmark	77.4	75.5	79.3	75.9	0.28	
	OSMN [105]	\times	\checkmark	73.5	74.0	72.9	71.8	0.14	
mask+ft	Ours (Fine-t. Box2Seg)	\checkmark	\checkmark	87.1	85.3	88.8	86.6	0.56	
	PReMVOS [60]	\checkmark	\checkmark	86.8	84.9	88.6	89.9	32.8	
	DyeNet [52]	\checkmark	\checkmark	—	86.2	—	—	4.66	
	OSVOS-S [64]	\checkmark	\checkmark	86.5	85.6	87.5	84.4	4.5	
	OnAVOS [89]	\checkmark	\checkmark	85.0	85.7	84.2	84.1	13	
	CINM [3]	\checkmark	\checkmark	84.2	83.4	85.0	83.6	> 120	
	OSVOS [8]	\checkmark	\checkmark	80.2	79.8	80.6	76.0	9	
	GT boxes (Box2Seg)	\times	\times	80.5	79.1	81.9	100.0	—	
	GT boxes (Fine-t. Box2Seg)	\checkmark	\checkmark	89.0	87.6	90.5	100.0	—	

Table 15: Quantitative results on the DAVIS 2016 validation set. FT denotes fine-tuning, M denotes using the first-frame mask, and t(s) denotes time per frame in seconds. Siam R-CNN (fastest) denotes Siam R-CNN with ResNet-50 backbone, half input resolution, and 100 RoIs from the RPN, see Section 4.5.

per video) compared to 14 state-of-the-art methods. Methods are ranked by the mean of \mathcal{J} and \mathcal{F} . Among methods which only use the first-frame bounding box (without the mask), Siam R-CNN achieves the strongest result with 78.6% $\mathcal{J} \& \mathcal{F}$, which is 8.8 percentage points higher than SiamMask [94]. When fine-tuning Box2Seg, our method achieves 87.1% $\mathcal{J} \& \mathcal{F}$, which is close to the best result on DAVIS 2016 by STM-VOS with 89.3%.

Table 16 shows results on the DAVIS 2017 [72] test-dev set (30 videos, 67.9 frames average length, 2.97 objects per video on average) compared to six state-of-the-art methods. Siam R-CNN achieves 53.3% $\mathcal{J} \& \mathcal{F}$, which is more than 10 percentage points higher than the result of SiamMask [94]. STM-VOS performs significantly better with 72.3%, however it relies on the first-frame mask which makes it less usable in practice.

B.4. Thorough Comparison to Previous Methods

Throughout the main paper and supplemental material we presented results on 11 short-term tracking benchmarks and four long-term tracking benchmarks, however these comparisons are spread throughout a number of tables and figures. We provide a unified and thorough comparison of our results to previous methods across all of these benchmarks in Table 17. We compare to the results of every paper

	Init	Method	FT	M	$\mathcal{J} \& \mathcal{F}$	\mathcal{J}	\mathcal{F}	t(s)
bbox	Siam R-CNN (ours)	\times	\times	53.3	48.1	58.6	0.44	
	Siam R-CNN (fastest)	\times	\times	51.6	46.3	56.8	0.16	
	SiamMask [94]	\times	\times	43.2	40.6	45.8	0.09 [†]	
mask	STM-VOS [69]	\times	\checkmark	72.3	69.3	75.2	0.48 [†]	
	RGMP [100]	\times	\checkmark	52.9	51.4	54.4	0.42 [†]	
	FEELVOS [87]	\times	\checkmark	57.8	55.2	60.5	0.54	
mask+ft	PReMVOS [60]	\checkmark	\checkmark	71.6	67.5	75.7	41.3	
	Ours (Fine-t. Box2Seg)	\checkmark	\checkmark	62.1	57.3	66.9	1.48	
	OnAVOS [89]	\checkmark	\checkmark	56.5	53.4	59.6	39	

Table 16: Quantitative results on the DAVIS 2017 test-dev set. FT denotes fine-tuning, M denotes using the first-frame mask, and t(s) denotes time per frame in seconds. [†]: timing extrapolated from DAVIS 2016 assuming linear scaling in the number of objects.

that presents comparable tracking results from major vision conferences in 2018 and 2019. As well as including all results from all of these papers we also present additional results from some methods that were either taken from later papers or that we obtained by evaluating open-source code. Sometimes these additional results were different to those presented in the original papers, in which case both results are shown. By evaluating on all of these datasets, and comparing to all methods from these two years, we are able to present a complete and holistic evaluation of our method compared to previous works.

Our method outperforms all previous methods on six out of the 11 evaluated short-term tracking benchmarks, sometimes by up to 7.2 percentage points. On the remaining five benchmarks we achieve close to the best results, with only a few previous methods obtaining better results, and by not too large a margin.

For long-term tracking, Siam R-CNN performs extremely well. Siam R-CNN outperforms all previous methods over all four benchmarks by between 3.9 and 10.1 percentage points.

B.5. Further Qualitative Evaluation

In Figure 12 we present further qualitative results of our method on the OTB2015, LTB35 and DAVIS 2017 benchmarks. We present results of our method compared to the best competing method. We show sequences for which Siam R-CNN has the best and worst relative performance compared to the competing method, as well as the sequence with the median relative performance.

Short-Term Tracking													Long-Term Tracking			
	TrackNet	GOT10k	Nfs	VOT15	OTB50	OTB15	UAV123	VOT16	OTB13	TC128	VOT17/18	OxUVA	LaSOT	UAV20L	LTB35	
	AUC	AUC	AUC	EAO	AUC	AUC	AUC	EAO	AUC	AUC	EAO	maxGM	AUC	AUC	F	
Ours	Δ to SOTA	+7.2	+3.8*	+1.9	+0.8	+0.6	+0.0	-0.5	-1.6	-1.8	-2.1	-3.2	+10.1	+7.9	+5.5	+3.9
	Siam R-CNN	81.2	64.9*	63.9	45.4†	66.3	70.1	64.9	46.5†	70.4	60.1	40.8†	72.3	64.8	67.2	66.8
	DiMP [5]	74.0	61.1	62.0			68.4	65.4			44.0		56.9 / 56.8§			
	UpdateNet [109]	67.7						48.1			39.3		47.5			
	GFSDCF [102]	60.9					69.3		72.2							
ICCV	SPLT [104]											62.2				61.6
2019	fdKCF [115]						67.5		34.7	70.5		26.5				
	Bridging [39]			51.5			64.7	58.6		65.6						
	GradNet [51]						63.9				55.6	24.7		36.5		
	MLT [15]						61.1			62.1				36.8		
	ARCF [40]						47.3									
	SiamRPN++ [48]	73.3	45.4§				69.6	61.3 / 64.2§			41.4		49.6		62.9	
	ATOM [19]	70.3		59 / 58.4§			67.1§	65 / 64.3§			40.1		51.5 / 51.4§			
	ASRCF [17]						69.2		39.1	60.3		32.8		35.9		
	SPM [90]		51.3§		65.3		68.7		43.4	69.3		33.8				
CVPR	SiamMask [94]							44.2				38.7				
2019	SiamDW [114]			38.0			67.3		37.0	66.6		30.1				
	RPCF [83]						69.6			71.3		31.6				
	C-RPN [25]	66.9					66.3		36.3	67.5		28.9		45.5		
	TADT [53]				32.7		66.0		29.9	68.0	56.2					
	GCT [27]						64.8	50.8		67.0		27.4				
	UDT [91]						63.2		30.1		54.1					
	UPDT [6]	61.1§		54.1 / 53.7§			70.1§	55 / 54.7§		62.2		37.8				
	DaSiamRPN [118]	63.8§			44.6		65.8§	58.6 / 58.5§	41.1		32.6		41.5§		61.7	60.7§
	ACT [10]				65.7		64.3		27.5	66.3						
	RTINet [107]					63.7	68.2		29.8	60.2						
	SACF [110]				34.3		69.3		38.0	71.3						
ECCV	DRL-IS [73]						67.1				59.0					
2018	DSLT [58]						66.0	53.0	33.2	68.3	58.7					
	Meta-Tracker [70]						66.2			31.7						
	RT-MDNet [41]						65.0	53.5			56.3					
	MemTrack [106]						62.6			64.2						
	StructSiam [113]						62.1		26.4	63.8			33.5§			
	SiamFC-tri [21]				53.5	59.2				62.9		21.3				
	DRT [81]					69.9		44.2	72.0							
	MCCT [92]					69.5		39.3	71.4	59.6						
	SiamRPN [49]			35.8		63.7		34.4					45.4§			
	STRCF [50]					68.3		31.3		60.1						
	VITAL [80]					68.2		32.3	71.0			39.0§				
CVPR	LSART [82]					67.2					32.3					
2018	FlowTrack [119]			34.1		65.5		33.4	68.9							
	RASNet [93]				32.7	64.1			67.0		28.1					
	SA-Siam [31]			31.0	61.0	65.7		29.1	67.7		23.6					
	TRACA [14]					60.3			65.2							
	MKCF [84]		45.5						64.1							
	HP [22]				55.4	60.1			62.9							
	SINT++ [95]				62.4	57.4										

Table 17: Comparison to all trackers published in CVPR, ICCV and ECCV in 2018 and 2019. Results from original papers, except when marked with § which are from later papers, or from running open-source code. Results in {Red, Green, Blue} are the {Best, Second, Third}, respectively. Benchmarks are ordered by performance relative to the best method other than ours (Δ to SOTA). Methods are ordered first by conference date, then by most ‘bests’, most ‘seconds’, most ‘thirds’ and finally by approximate ‘head-to-head’ performance. *On all benchmarks Siam R-CNN uses exactly the same network weights and tracking hyper-parameters*, except for those marked with † which use the ‘short-term’ tracking parameters, and those marked with * which use weights trained only on GOT-10k.

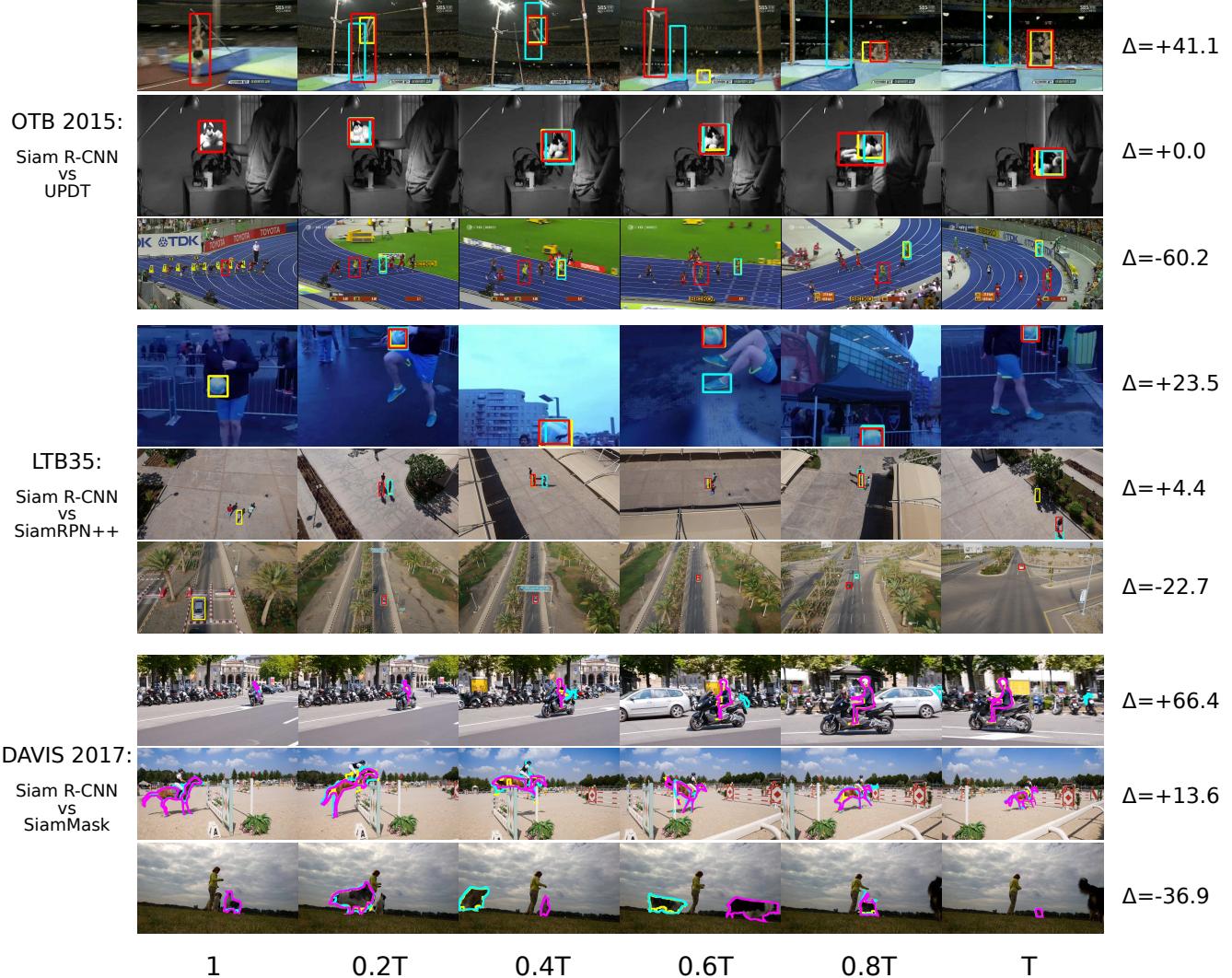


Figure 12: Qualitative results on OTB2015 [99], LTB35 [62], and DAVIS 2017 [72] (validation set). We compare the results of Siam R-CNN to the best competing methods, which are UPDT [6] for OTB 2015, SiamRPN++ [48] for LTB35, and SiamMask [94] for DAVIS 2017. Siam R-CNN’s result is shown in red (magenta for DAVIS 2017), the competing methods’ results are shown in blue, and the ground truth in yellow. For each benchmark, the sequences with the best, worst and median relative performance (Δ) between Siam R-CNN and the competing method are shown. Six frames spaced equally throughout each video are shown.