This ECCV 2018 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ECCV 2018
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/eccv

Distractor-aware Siamese Networks for Visual Object Tracking

Zheng Zhu$^{*1,2[0000-0002-4435-1692]}$, Qiang Wang^{*1,2}, Bo Li^{*3}, Wei Wu^{3},
Junjie Yan^{3}, and Weiming Hu^{1,2}

^{1}University of Chinese Academy of Sciences, Beijing, China
^{2}Institute of Automation, Chinese Academy of Sciences, Beijing, China
^{3}SenseTime Group Limited, Beijing, China

Abstract. Recently, Siamese networks have drawn great attention in visual tracking community because of their balanced accuracy and speed. However, features used in most Siamese tracking approaches can only discriminate foreground from the non-semantic backgrounds. The semantic backgrounds are always considered as distractors, which hinders the robustness of Siamese trackers. In this paper, we focus on learning distractor-aware Siamese networks for accurate and long-term tracking. To this end, features used in traditional Siamese trackers are analyzed at first. We observe that the imbalanced distribution of training data makes the learned features less discriminative. During the off-line training phase, an effective sampling strategy is introduced to control this distribution and make the model focus on the semantic distractors. During inference, a novel distractor-aware module is designed to perform incremental learning, which can effectively transfer the general embedding to the current video domain. In addition, we extend the proposed approach for long-term tracking by introducing a simple yet effective local-to-global search region strategy. Extensive experiments on benchmarks show that our approach significantly outperforms the state-of-the-arts, yielding 9.6% relative gain in VOT2016 dataset and 35.9% relative gain in UAV20L dataset. The proposed tracker can perform at 160 FPS on short-term benchmarks and 110 FPS on long-term benchmarks.

Keywords: Visual Tracking · Distractor-aware · Siamese Networks

1 Introduction

Visual object tracking, which locates a specified target in a changing video sequence automatically, is a fundamental problem in many computer vision topics such as visual analysis, automatic driving and pose estimation. A core problem of tracking is how to detect and locate the object accurately and efficiently in challenging scenarios with occlusions, out-of-view, deformation, background cluttering and other variations [38].

*The first three authors contributed equally to this work. This work is done when Zheng Zhu and Qiang Wang are interns at SenseTime Group Limited.

Recently, Siamese networks, which follow a tracking by similarity comparison strategy, have drawn great attention in visual tracking community because of favorable performance [31, 8, 2, 36, 33, 7, 37, 16]. SINT [31], GOTURN [8], SiamFC [2] and RASNet [36] learn a priori deep Siamese similarity function and use it in a run-time fixed way. CFNet [33] and DSiam [7] can online update the tracking model via a running average template and a fast transformation learning module, respectively. SiamRPN [16] introduces a region proposal network after the Siamese network, thus formulating the tracking as a one-shot local detection task.

Although these tracking approaches obtain balanced accuracy and speed, there are 3 problems that should be addressed: firstly, features used in most Siamese tracking approaches can only discriminate foreground from the non-semantic background. The semantic backgrounds are always considered as distractors, and the performance can not be guaranteed when the backgrounds are cluttered. Secondly, most Siamese trackers can not update the model [31, 8, 2, 36, 16]. Although their simplicity and fixed-model nature lead to high speed, these methods lose the ability to update the appearance model online which is often critical to account for drastic appearance changes in tracking scenarios. Thirdly, recent Siamese trackers employ a local search strategy, which can not handle the full occlusion and out-of-view challenges.

In this paper, we explore to learn Distractor-aware Siamese Region Proposal Networks (DaSiamRPN) for accurate and long-term tracking. SiamFC uses a weighted loss function to eliminate class imbalance of the positive and negative examples. However, it is inefficient as the training procedure is still dominated by easily classified background examples. In this paper, we identify that the imbalance of the *non-semantic* background and *semantic* distractor in the training data is the main obstacle for the representation learning. As shown in Fig. 1, the response maps on the SiamFC can not distinguish the people, even the athlete in the white dress can get a high similarity with the target person. High quality training data is crucial for the success of end-to-end learning tracker. We conclude that the quality of the representation network heavily depends on the distribution of training data. In addition to introducing positive pairs from existing large-scale detection datasets, we explicitly generate diverse semantic negative pairs in the training process. To further encourage discrimination, an effective data augmentation strategy customizing for visual tracking are developed.

After the offline training, the representation networks can generalize well to most categories of objects, which makes it possible to track general targets. During inference, classic Siamese trackers only use nearest neighbour search to match the positive templates, which might perform poorly when the target undergoes significant appearance changes and background clutters. Particularly, the presence of similar looking objects (distractors) in the context makes the tracking task more arduous. To address this problem, the surrounding contextual and temporal information can provide additional cues about the targets and help to maximize the discrimination abilities. In this paper, a novel distractor-aware module is designed, which can effectively transfer the general embedding to the current video domain and incrementally catch the target appearance variations during inference.

Besides, most recent trackers are tailored to short-term scenario, where the target object is always present. These works have focused exclusively on short sequences of a few tens of seconds, which is poorly representative of practitioners' needs. Except the challenging situations in short-term tracking, severe out-of-view and full occlusion introduce extra challenges in long-term tracking. Since conventional Siamese trackers lack discriminative features and adopt local search region, they are unable to handle these challenges. Benefiting from the learned distractor-aware features in DaSiamRPN, we extend the proposed approach for long-term tracking by introducing a simple yet effective local-to-global search region strategy. This significantly improves the performance of our tracker in out-of-view and full occlusion challenges.

We validate the effectiveness of proposed DaSiamRPN framework on extensive short-term and long-term tracking benchmarks: VOT2016 [14], VOT2017 [12], OTB2015 [38], UAV20L and UAV123 [22]. On short-term VOT2016 dataset, DaSiamRPN achieves a 9.6% relative gain in Expected Average Overlap compared to the top ranked method ECO [3]. On long-term UAV20L dataset, DaSiamRPN obtains 61.7% in Area Under Curve which outperforms the current best-performing tracker by relative 35.9%. Besides the favorable performance, our tracker can perform at far beyond real-time speed: 160 FPS on short-term datasets and 110 FPS on long-term datasets. All these consistent improvements demonstrate that the proposed approach establish a new state-of-the-art in visual tracking.

1.1 Contributions

The contributions of this paper can be summarized in three folds as follows:

1, The features used in conventional Siamese trackers are analyzed in detail. And we find that the imbalance of the *non-semantic* background and *semantic* distractor in the training data is the main obstacle for the learning.

2, We propose a novel Distractor-aware Siamese Region Proposal Networks (DaSiamRPN) framework to learn distractor-aware features in the off-line training, and explicitly suppress distractors during the inference of online tracking.

3, We extend the DaSiamRPN to perform long-term tracking by introducing a simple yet effective local-to-global search region strategy, which significantly improves the performance of our tracker in out-of-view and full occlusion challenges. In comprehensive experiments of short-term and long-term visual tracking benchmarks, the proposed DaSiamRPN framework obtains state-of-the-art accuracy while performing at far beyond real-time speed.

2 Related Work

Siamese Networks based Tracking. Siamese trackers follow a tracking by similarity comparison strategy. The pioneering work is SINT [31], which simply searches for the candidate most similar to the exemplar given in the starting frame, using a run-time fixed but learns a priori deep Siamese similarity function. As a follow-up work, Bertinetto et.al [2] propose a fully convolutional Siamese network (SiamFC) to estimate the feature similarity region-wise between two frames. RASNet [36] advances this similarity metric by learning the attention mechanism with a Residual Attentional Network. Different from SiamFC and RASNet, in GOTURN tracker [8], the motion between successive frames is predicted using a deep regression network. These threee trackers are able to perform at 86 FPS, 83FPS and 100 FPS respectively on GPU because no fine-tuning is performed online. CFNet [33] interprets the correlation filters as a differentiable layer in a Siamese tracking framework, thus achieving an end-to-end representation learning. But the performance improvement is limited compared with SiamFC. FlowTrack [40] exploits motion information in Siamese architecture to improve the feature representation and the tracking accuracy. It is worth noting that CFNet and FlowTrack can efficiently online update the tracking model. Recently, SiamRPN [16] formulates the tracking as a one-shot local detection task by introducing a region proposal network after a Siamese network, which is end-to-end trained off-line with large-scale image pairs.

Features for Tracking. Visual features play a significant role in computer vision tasks including visual tracking. Possegger et.al [26] propose a distractor-aware model term to suppress visually distracting regions, while the color histograms features used in their framework are less robust than the deep features. DLT [35] is the seminal deep learning tracker which uses a multi-layer autoencoder network. The feature is pretrained on part of the 80M Tiny Image dataset [32] in an unsupervised fashion. Wang et al. [34] learn a two-layer neural network on a video repository, where temporally slowness constraints are imposed for feature learning. DeepTrack [17] learns two-layer CNN classifiers from binary samples and does not require a pre-training procedure. UCT [39] formulates the features learning and tracking process into a unified framework, enabling learned features are tightly coupled to tracking process.

Long-term Tracking. Traditional long-term tracking frameworks can be divided into two groups: earlier methods regard tracking as local key point descriptors matching with a geometrical model [25, 24, 21], and recent approaches perform long-term tracking by combining a short-term tracker with a detector. The seminal work of latter categories is TLD [10], which proposes a memory-less flock of flows as a short-term tracker and a template-based detector run in parallel. Ma et al. [20]propose a combination of KCF tracker and a random ferns classifier as a detector that is used to correct the tracker. Similarly, MUSTer [9] is a long-term tracking framework that combines KCF tracker with a SIFT-based detector that is also used to detect occlusions. Fan and Ling [6] combines a DSST tracker [4] with a CNN detector [31] that verifies and potentially corrects proposals of the short-term tracker.

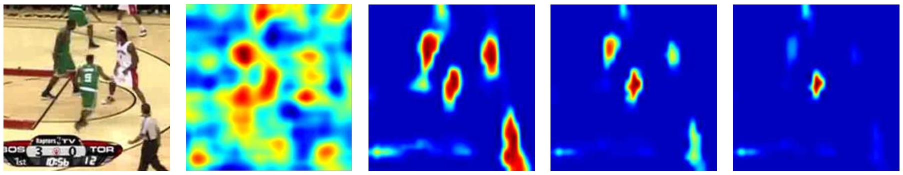

(a) ROI (b) SiamFC (c) SiamRPN (d) SiamRPN+ (e) Ours

Fig. 1: Visualization of the response heatmaps of Siamese network trackers. (a) shows the search images. (b-e) show the heatmaps that produced by SiamFC, SiamRPN, SiamRPN+ (trained with distractors) and the DaSiamRPN.

3 Distractor-aware Siamese Networks

3.1 Features and Drawbacks in Traditional Siamese Networks

Before the detailed discussion of our proposed framework, we first revisit the features of conventional Siamese network based tracking [2, 16]. Siamese trackers use metric learning at their core. The goal is to learn an embedding space that can maximize the interclass inertia between different objects and minimize the intraclass inertia for the same object. The key contribution leading to the popularity and success of Siamese trackers is their balanced accuracy and speed.

Fig. 1 visualizes of response maps of SiamFC and SiamRPN. It can be seen that for the targets, those with large differences in the background also achieve high scores, and even some extraneous objects get high scores. The representations obtained in SiamFC usually serve the discriminative learning of the categories in training data. In SiamFC and SiamRPN, pairs of training data come from different frames of the same video, and for each search area, the *non-semantic* background occupies the majority, while semantic entities and distractor occupy less. This imbalanced distribution makes the training model hard to learn instance-level representation, but tending to learn the differences between foreground and background.

During inference, nearest neighbor is used to search the most similar object in the search region, while the background information labelled in the first frame are omitted. The background information in the tracking sequences can be effectively utilized to increase the discriminative capability as shown in Fig. 1e.

To eliminate these issues, we propose to actively generate more semantics pairs in the offline training process and explicitly suppress the distractors in the online tracking.

3.2 Distractor-aware Training

High quality training data is crucial for the success of end-to-end representation learning in visual tracking. We introduce series of strategies to improve the generalization of the learned features and eliminate the imbalanced distribution of the training data.

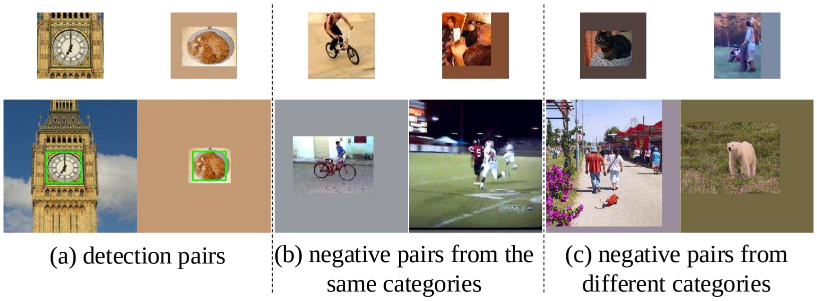
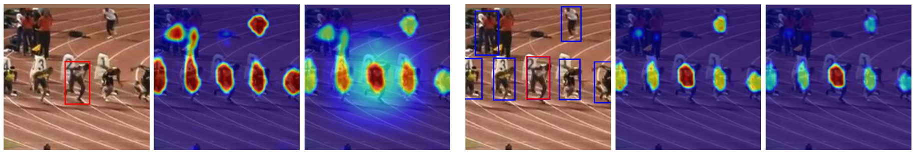

(a) detection pairs
(b) negative pairs from the same categories
(c) negative pairs from different categories

Fig. 2: (a) positive pairs generated from detection datasets through augmenting still images. (b) negative pairs from the same category. (c) negative pairs from different categories.

Diverse categories of positive pairs can promote the generalization ability. The original SiamFC is trained on the ILSVRC video detection datasets, which consists of only about 4,000 videos annotated frame-by-frame [28]. Recently, SiamRPN [16] explores to use sparsely labelled Youtube-BB [27] videos which consists of more than 200,000 videos annotated once in every 30 frames. In these two methods, target pairs of training data come from different frames in the same video. However, these video detection datasets only contain few categories (20 for VID [28], 30 for Youtube-BB [27]), which is not sufficient to train high-quality and generalized features for Siamese tracking. Besides, the bounding box regression branch in the SiamRPN may get inferior predictions when encountering new categories. Since labelling videos is time-consuming and expensive, in this paper, we greatly expand the categories of positive pairs by introducing large-scale ImageNet Detection [28] and COCO Detection [18] datasets. As shown in Fig. 2(a), through augmentation techniques (translation, resize, grayscale et.al), still images from detection datasets can be used to generate image pairs for training. The diversity of positive pairs is able to improve the tracker's discriminative ability and regression accuracy.

Semantic negative pairs can improve the discriminative ability. We attribute the less discriminative representation in SiamFC [2] and SiamRPN [16] to two level of imbalanced training data distribution. The first imbalance is the rare semantic negative pairs. Since the background occupies the majority in the training data of SiamFC and SiamRPN, most negative samples are non-semantic (not real object, just background), and they can be easily classified. That is to say, SiamFC and SiamRPN learn the differences between foreground and background, and the losses between semantic objects are overwhelmed by the vast number of easy negatives. Another imbalance comes from the intraclass distractors, which usually perform as hard negative samples in the tracking process. In this paper, semantic negative pairs are added into the training process. The constructed negative pairs consist of labelled targets both in the same categories and different categories. The negative pairs from different categories can help tracker to avoid drifting to arbitrary objects in challenges such as out-of-



(a) General Siamese tracker (b) Distractor-aware Siamese tracker

Fig. 3: Illustrations of our proposed Distractor-aware Siamese Region Proposal Networks (DaSiamRPN). The target and the background information are fully utilized in DaSiamRPN, which can suppress the influence of distractor during tracking.

view and full occlusion, while negative pairs from the same categories make the tracker focused on fine-grained representation. The negative examples are shown in Fig. 2(b) and Fig. 2(c).

Customizing effective data augmentation for visual tracking. To unleash the full potential of the Siamese network, we customize several data augmentation strategies for training. Except the common translation, scale variations and illumination changes, we observe that the motion pattern can be easily modeled by the shallow layers in the network. We explicitly introduce motion blur in the data augmentation.

3.3 Distractor-aware Incremental Learning

The training strategy in the last subsection can significantly improve the discrimination power on the offline training process. However, it is still hard to distinguish two objects with the similar attributes like Fig. 3a. SiamFC and SiamRPN use a cosine window to suppress the distractors. In this way, the performance is not guaranteed when the motion of objects are messy. Most existing Siamese network based approaches provide inferior performance when encountering with fast motion or background clutter. In summary, the potential flaw is mainly due to the misalignment of the general representation domain and the specifical target domains. In this section, we propose a distractor-aware module to effectively transfer the general representation to the video domain.

The Siamese tracker learns a similarity metric $f(z, x)$ to compare an exemplar image z to a candidate image x in the embedding space φ:

$$f(z, x) = \varphi(z) \star \varphi(x) + b \cdot \mathbb{1} \tag{1}$$

where $\star$ denotes cross correlation between two feature maps, $b \cdot \mathbb{1}$ denotes a bias which is equated in every location. The most similar object of the exemplar will be selected as the target.

To make full use of the label information, we integrate the hard negative samples (distractors) in context of the target into the similarity metric. In

DaSiamRPN, the Non Maximum Suppression (NMS) is adopted to select the potential distractors d_i in each frames, and then we collect a distractor set $\mathcal{D} := \{\forall d_i \in \mathcal{D}, f(z, d_i) > h \cap d_i \neq z_t\}$, where h is the predefined threshold, z_t is the selected target in frame t and the number of this set $|\mathcal{D}| = n$. Specifically, we get $17 * 17 * 5$ proposals in each frame at first, and then we use NMS to reduce redundant candidates. The proposal with highest score will be selected as the target z_t. For the remaining, the proposals with scores greater than a threshold are selected as distractors.

After that, we introduce a novel distractor-aware objective function to *re-rank* the proposals $\mathcal{P}$ which have *top-k* similarities with the exemplar. The final selected object is denoted as q:

$$q = \underset{p_k \in \mathcal{P}}{argmax} \;\; f(z, p_k) - \frac{\hat{\alpha} \sum_{i=1}^{n} \alpha_i f(d_i, p_k)}{\sum_{i=1}^{n} \alpha_i} \tag{2}$$

the weight factor $\hat{\alpha}$ control the influence of the distractor learning, the weight factor α_i is used to control the influence for each distractor d_i. It is worth noting that the computational complexity and memory usage increase n times by a direct calculation. Since cross correlation operation in the Equation (1) is a linear operator, we utilize this property to speed up the distractor-aware objective:

$$q = \underset{p_k \in \mathcal{P}}{argmax} \;\; (\varphi(z) - \frac{\hat{\alpha} \sum_{i=1}^{n} \alpha_i \varphi(d_i)}{\sum_{i=1}^{n} \alpha_i}) \star \varphi(p_k) \tag{3}$$

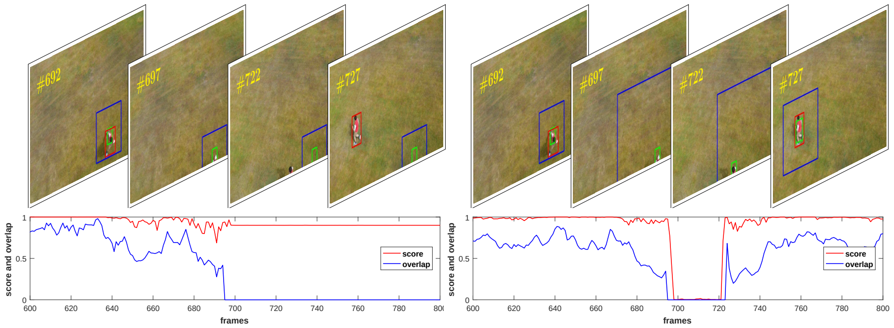
it enables the tracker run in the comparable speed in comparisons with SiamRPN. This associative law also inspires us to incrementally learn the target templates and distractor templates with a learning rate β_t:

$$q_{T+1} = \underset{p_k \in \mathcal{P}}{argmax} \;\; (\frac{\sum_{t=1}^{T} \beta_t \varphi(z_t)}{\sum_{t=1}^{T} \beta_t} - \frac{\sum_{t=1}^{T} \beta_t \hat{\alpha} \sum_{i=1}^{n} \alpha_i \varphi(d_{i,t})}{\sum_{t=1}^{T} \beta_t \sum_{i=1}^{n} \alpha_i}) \star \varphi(p_k) \tag{4}$$

This distractor-aware tracker can adapt the existing similarity metric (general) to a similarity metric for a new domain (specific). The weight factor α_i can be viewed as the dual variables with sparse regularization, and the exemplars and distractors can be viewed as positive and negative samples in correlation filters. Actually, an online classifier is modeled in our framework. So the adopted classifier is expected to perform better than these only use general similarity metric.

3.4 DaSiamRPN for Long-term Tracking

In this section, the DaSiamRPN framework is extended for long-term tracking. Besides the challenging situations in short-term tracking, severe out-of-view and full occlusion introduce extra challenges in long-term tracking, which are shown in Fig. 4. The search region in short-term tracking (SiamRPN) can not cover the target when it reappears, thus failing to track the following frames. We propose a simple yet effective switch method between short-term tracking phase and failure



(a) scores and overlaps in SiamRPN (b) scores and overlaps in DaSiamRPN

Fig. 4: The tracking results of video *person7* in out-of-view challenge. First row: tracking snapshots of SiamRPN and DaSiamRPN. Second row: detection scores and according overlaps of the two methods. The overlaps are defined as intersection-over-union (IOU) between tracking results and ground truth. Red: ground truth. Green: tracking box. Blue: Search region box.

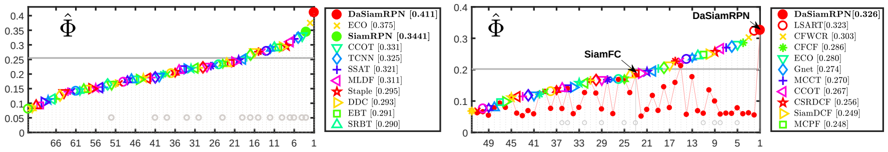
cases. In failure cases, an iterative local-to-global search strategy is designed to re-detect the target.

In order to perform switches, we need to identify the beginning and the end of failed tracking. Since the distractor-aware training and inference enable high-quality detection score, it can be adopted to indicate the quality of tracking results. Fig. 4 shows the detection scores and according tracking overlaps in SiamRPN and DaSiamRPN. The detection scores of SiamRPN are not indicative, which can be still high even in out-of-view and full occlusion. That is to say, SiamRPN tends to find an arbitrary objectness in these challenges which causes drift in tracking. In DaSiamRPN, detection scores successfully indicate status of the tracking phase.

During failure cases, we gradually increase the search region by local-to-global strategy. Specifically, the size of search region is iteratively growing with a constant step when failed tracking is indicated. As shown in Fig. 4, the local-to-global search region covers the target to recover the normal tracking. It is worth noting that our tracker employs bounding box regression to detect the target, so the time-consuming image pyramids strategy can be discarded. In experiments, the proposed DaSiamRPN can perform at 110 FPS on long-term tracking benchmark.

4 Experiments

Experiments are performed on extensive challenging tracking datasets, including VOT2015 [13], VOT2016 [14] and VOT2017 [12], each with 60 videos, UAV20L [22] with 20 long-term videos, UAV123 [22] with 123 videos and OTB2015 [38] with



(a) EAO on VOT2016 (b) EAO on VOT2017

Fig. 5: Expected average overlap plot for VOT2016 (a) and VOT2017 (b).

100 videos. All the tracking results are provided by official implementations to ensure a fair comparison.

4.1 Experimental Details

The modified AlexNet [15] pretrained using ImageNet [28] is used as described in SiamRPN [16]. The parameters of the first three convolution layers are fixed and only the last two convolution layers are fine-tuned. There are totally 50 epoches performed and the learning rate is decreased in log space from 10^{-2} to 10^{-4}. We extract image pairs from VID [28] and Youtube-BB [27] by choosing frames with interval less than 100 and performing crop procedure as described in Section 3.2. In ImageNet Detection [28] and COCO Detection [18] datasets, image pairs are generated for training by augmenting still images. To handle the gray videos in benchmarks, 25% of the pairs are converted to grayscale during training. The translation is randomly performed within 12 pixels, and the range of random resize varies from 0.85 to 1.15.

During inference phase, the distractor factor $\hat{\alpha}$ in Equation (2) is set to 0.5, α_i is set to 1 for each distractor, and the incremental learning factor β_t in Equation (4) is set to $\sum_{i=0}^{t-1}(\frac{\eta}{1-\eta})^i$, where $\eta = 0.01$. In the long-term tracking, we find that one step iteration of local-to-global is sufficient. Specifically, the sizes of the search region in short-term phase and defined failure cases are set to 255 and 767, respectively. The thresholds to enter and leave failure cases are set to 0.8 and 0.95. Our experiments are implemented using PyTorch on a PC with an Intel i7, 48G RAM, NVIDIA TITAN X. The proposed tracker can perform at 160 FPS on short-term benchmarks and 110 FPS on long-term benchmarks.

4.2 State-of-the-art Comparisons on VOT Datasets

In this section the latest version of the Visual Object Tracking toolkit (*vot2017-challenge*) is used. The toolkit applies a reset-based methodology. Whenever a failure (zero overlap with the ground truth) is detected, the tracker is re-initialized five frames after the failure. The performance is measured in terms of accuracy (A), robustness (R), and expected average overlap (EAO). In addition, VOT2017 also introduces a real-time experiment. We report all these metrics compared with a number of the latest state-of-the-art trackers on VOT2015, VOT2016 and VOT2017.

The EAO curve evaluated on VOT2016 is presented in Fig. 5a and 70 other state-of-the-art trackers are compared. The EAO of our baseline tracker SiamRPN on VOT2016 is 0.3441, which already outperforms most of state-of-the-arts. However, there is still a gap compared with the top-ranked tracker ECO (0.375), which improves continuous convolution operators on multi-level feature maps. Most remarkably, the proposed DaSiamRPN obtains a EAO of 0.411, outperforming state-of-the-arts by relative 9.6%. Furthermore, our tracker runs at state-of-the-art speed with 160FPS, which is 500× faster than C-COT and 20× faster than ECO.

For the evaluation on VOT2017, Fig. 5b reports the results of ours against 51 other state-of-the-art trackers with respect to the EAO score. DaSiamRPN ranks first with an EAO score of 0.326. Among the top 5 trackers, CFWCR, CFCF, ECO, and Gnet apply continuous convolution operator as the baseline approach. The top performer LSART [30] decomposes the target into patches and applies a weighted combination of patch-wise similarities into a kernelized ridge regression. While our method is conceptually much simpler, powerful and is also easy to follow.

Fig. 5b also reveals the EAO values in the real-time experiment denoted by red points. Our tracker obviously is the top-performer with a real-time EAO of 0.326 and outperforms the latest state-of-the-art real-time tracker CSRDCF++ by relative 53.8%.

Table 1 shows accuracy (A) and robustness (R), as well as expected average overlap (EAO) on VOT2015, VOT2016 and VOT2017. The baseline approach SiamRPN can process an astounding 200 frames per second while still getting an comparable performance with the state-of-the-arts. We find the performance gains of SiamRPN are mainly due to their accurate multi-anchors regression mechanism. We propose the distractor-aware module to improve the robustness, which can make our tracker much more harmonious. As a result, our approach, with the EAO of 0.446, 0.411 and 0.326 on three benchmarks, outperforms all the existing trackers by a large margin. We believe that the consistent improvements demonstrate that our approach makes real contributions by both the training process and online inference.

4.3 State-of-the-art Comparisons on UAV Datasets

The UAV [22] videos are captured from low-altitude unmanned aerial vehicles. The dataset contains a long-term evaluation subset UAV20L and a short-term evaluation subset UAV123. The evaluation is based on two metrics: precision plot and success plot.

Results on UAV20L UAV20L is a long-term tracking benchmark that contains 20 sequences with average sequence length 2934 frames. Besides the challenging situations in short-term tracking, severe out-of-view and full occlusion introduce extra challenges. In this experiment, the proposed method is compared against recent trackers in [22]. Besides, ECO [3] (state-of-the-art short-term

Table 1: Performance comparisons on public short-term benchmarks. OP: mean overlap precision at the threshold of 0.5; DP: mean distance precision of 20 pixels; EAO: expected average overlap, and mean speed (FPS). The **red bold** fonts and *blue italic* fonts indicate the best and the second best performance.

Trackers	OTB-2015		VOT2015			VOT2016			VOT2017			FPS
	OP	DP	A	R	EAO	A	R	EAO	A	R	EAO	
SiamFC	73.0	77.0	0.533	0.88	0.289	0.53	0.46	0.235	0.50	0.59	0.188	86
CFNet	69.9	74.7	-	-	-	-	-	-	-	-	-	75
Staple	70.9	78.4	0.57	1.39	0.300	0.54	0.38	0.295	*0.52*	0.69	0.169	80
CSRDCF	70.7	78.7	0.56	0.86	0.320	0.51	0.24	0.338	0.49	0.36	0.256	13
BACF	76.7	81.5	0.59	1.56	-	-	-	-	-	-	-	35
ECO-HC	78.4	85.6	-	-	-	0.54	0.30	0.322	0.49	0.44	0.238	60
CREST	77.5	83.7	-	-	-	0.51	0.25	0.283	-	-	-	1
MDNet	*85.4*	*90.9*	*0.60*	*0.69*	*0.378*	0.54	0.34	0.257	-	-	-	1
C-COT	82.0	89.8	0.54	0.82	0.303	0.54	0.24	0.331	0.49	*0.32*	0.267	0.3
ECO	84.9	**91.0**	-	-	-	0.55	**0.20**	*0.375*	0.48	**0.27**	*0.280*	8
SiamRPN	81.9	85.0	0.58	1.13	0.349	*0.56*	0.26	0.344	0.49	0.46	0.244	**200**
Ours	**86.5**	88.0	**0.63**	**0.66**	**0.446**	**0.61**	*0.22*	**0.411**	**0.56**	0.34	**0.326**	*160*

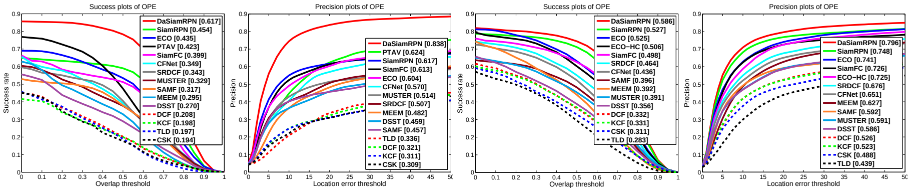


Fig. 6: Success and precision plots on UAV [22] dataset. First and second sub-figures are results of UAV20L, third and last sub-figures are results of UAV123.

tracker), PTAV [6] (state-of-the-art long-term tracker), SiamRPN [16] (the baseline), SiamFC [2] and CFNet [33] (representative Siamese trackers) are added for comparison.

The results including success plots and precision plots are illustrated in Fig. 6. It clearly illustrates that our algorithm, denoted by DaSiamRPN, outperforms the state-of-the-art trackers significantly in both measures. In the success plot, our approach obtains an AUC score of 0.617, significantly outperforming state-of-the-art short-term trackers SiamRPN [16] and ECO [3]. The improvement ranges are relative 35.9% and 41.8%, respectively. Compared with PTAV [6], MUSTer [9] and TLD [10] which are qualified to perform long-term tracking, the proposed DaSiamRPN outperforms these trackers by relative 45.8%, 87.5%

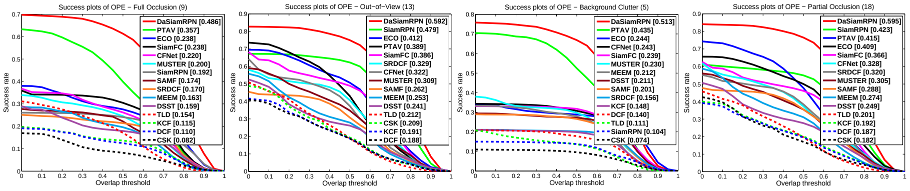


Fig. 7: Success plots with attributes on UAV20L. Best viewed on color display.

and 213.2%. In the precision plot, our approach obtains a score of 0.838, outperforming state-of-the-art long-term tracker (PTAV [6]) and short-term tracker (SiamRPN [16]) by relative 34.3% and 35.8%, respectively. The excellent performance of DaSiamRPN in this long-term tracking dataset can be attributed to the distractor-aware features and local-to-global search strategy.

For detailed performance analysis, we also report the results on various challenge attributes in UAV20L, i.e. full occlusion, out-of-view, background clutter and partial occlusion. Fig. 7 demonstrates that our tracker effectively handles these challenging situations while other trackers obtain lower scores. Specially, in full occlusion and background clutter attributes, the proposed DaSiamRPN outperforms SiamRPN [16] by relative 153.1% and 393.2%.

Results on UAV123 UAV123 dataset includes 123 sequences with average sequence length of 915 frames. Besides the recent trackers in [22], ECO [3], PTAV [6], SiamRPN [16], SiamFC [2], CFNet [33] are added for comparison. Fig. 6 illustrates the precision and success plots of the compared trackers. The proposed DaSiamRPN approach outperforms all the other trackers in terms of success and precision scores. Specifically, our method achieves a success score of 0.586, which outperforms the SiamRPN (0.527) and ECO (0.525) method with a large margin.

4.4 State-of-the-Art Comparisons on OTB Datasets

We evaluate the proposed algorithms with numerous fast and state-of-the-art trackers including SiamFC [2], CFNet [33], Staple [1], CSRDCF [19], BACF [11], ECO-HC [3], CREST [29], MDNet [23], CCOT [5], ECO [3], and the baseline tracker SiamRPN [16]. All the trackers are initialized with the ground-truth object state in the first frame. Mean overlap precision (OP) and mean distance precision (DP) are reported in Table 1.

Among the real-time trackers, SiamFC and CFNet are latest Siamese network based trackers while the accuracies is still left far behind the state-of-the-art BACF and ECO-HC with HOG features. The proposed DaSiamRPN tracker outperforms all these trackers by a large margin on both the accuracy and speed.

For state-of-the-art comparisons on OTB, MDNet, trained on visual tracking datasets, performs the best against the other trackers at a speed of 1 FPS. C-COT and ECO achieve state-of-the-art performance, but their tracking speeds

Table 2: Ablation analyses of our algorithm on VOT2016 [14] and UAV20L [22]

Component	SiamRPN	DaSiamRPN			
positive pairs in detection data?		✓	✓	✓	✓
semantic negative pairs?			✓	✓	✓
distractor-aware updating?				✓	✓
long-term tracking module?					✓
EAO in VOT2016	0.344	0.368	0.389	0.411	–
AUC in UAV20L(%)	45.4	47.2	48.6	49.8	61.7

are not fast enough for real-time applications. The baseline tracker SiamRPN obtains an OP score of 81.9%, which is slightly less accurate than CCOT. The bottleneck of SiamRPN is its inferior robust performance. Since the distractor-aware mechanisms in both training and inference focus on improving the robustness, the proposed DaSiamRPN tracker achieves 3.0% improvement on DP and performs best OP score of 86.5% on OTB2015.

4.5 Ablation Analyses

To verify the contributions of each component in our algorithm, we implement and evaluate four variations of our approach. Analyses results include EAO on VOT2016 [14] and AUC on UAV20L [22].

As shown in Table 2, SiamRPN is our baseline algorithm. In VOT2016, the EAO criterion increases to 0.368 from 0.344 when detection data is added in training. Similarly, when negative pairs and distractor-aware learning are adopted in training and inference, both the performance increases by near 2%. In UAV20L, detection data, negative pairs in training and distractor-aware inference gain the performance by 1%-2%. The AUC criterion increases to 61.7% from 49.8% when long-term tracking module is adopted.

5 Conclusions

In this paper, we propose a distractor-aware Siamese framework for accurate and long-term tracking. During offline training, a distractor-aware feature learning scheme is proposed, which can significantly boost the discriminative power of the networks. During inference, a novel distractor-aware module is designed, effectively transferring the general embedding to the current video domain. In addition, we extend the proposed approach for long-term tracking by introducing a simple yet effective local-to-global search strategy. The proposed tracker obtains state-of-the-art accuracy in comprehensive experiments of short-term and long-term visual tracking benchmarks, while the overall system speed is still far from being real-time.

References

1. Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.S.: Staple: Complementary learners for real-time tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2016)
2. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S.: Fully-convolutional siamese networks for object tracking. In: Proceedings of the European Conference on Computer Vision Workshop. pp. 850–865 (2016)
3. Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M.: Eco: Efficient convolution operators for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (July 2017)
4. Danelljan, M., Hger, G., Khan, F.S., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: Proceedings of the British Machine Vision Conference. pp. 65.1–65.11 (2014)
5. Danelljan, M., Robinson, A., Khan, F.S., Felsberg, M.: Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: Proceedings of the European Conference on Computer Vision. pp. 472–488 (2016)
6. Fan, H., Ling, H.: Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
7. Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L., Wang, S.: Learning dynamic siamese network for visual object tracking. In: Proceedings of the IEEE International Conference on Computer Vision (Oct 2017)
8. Held, D., Thrun, S., Savarese, S.: Learning to track at 100 fps with deep regression networks. In: Proceedings of the European Conference on Computer Vision. pp. 749–765 (2016)
9. Hong, Z., Chen, Z., Wang, C., Mei, X., Prokhorov, D., Tao, D.: Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 749–758 (2015)
10. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(7), 1409–1422 (2012)
11. Kiani Galoogahi, H., Fagg, A., Lucey, S.: Learning background-aware correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision (Oct 2017)
12. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Cehovin Zajc, L., Vojir, T., Hager, G., Lukezic, A., Eldesokey, A., Fernandez, G.: The visual object tracking vot2017 challenge results. In: Proceedings of the The IEEE International Conference on Computer Vision Workshop (Oct 2017)
13. Kristan, M., Matas, J., Leonardis, A., Felsberg, M.: The visual object tracking vot2015 challenge results. In: Proceedings of the IEEE International Conference on Computer Vision Workshop. pp. 564–586 (2015)
14. Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernandez, G., Vojir, T., Hager, G., Nebehay, G., Pflugfelder, R.: The visual object tracking vot2016 challenge results. In: Proceedings of the European Conference on Computer Vision Workshop. pp. 191–217. Springer International Publishing (2016)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 1097–1105 (2012)

16. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8971–8980 (2018)
17. Li, H., Li, Y., Porikli, F.: Deeptrack: Learning discriminative feature representations online for robust visual tracking. IEEE Transactions on Image Processing **25**(4), 1834–1848 (2016)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755. Springer (2014)
19. Lukezic, A., Vojir, T., Zajc, L.C., Matas, J., Kristan, M.: Discriminative correlation filter with channel and spatial reliability. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
20. Ma, C., Yang, X., Zhang, C., Yang, M.H.: Long-term correlation tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5388–5396 (2015)
21. Maresca, M.E., Petrosino, A.: Matrioska: A multi-level approach to fast tracking by learning. In: Proceedings of the International Conference on Image Analysis and Processing. pp. 419–428. Springer (2013)
22. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: Proceedings of the European Conference on Computer Vision. pp. 445–461. Springer (2016)
23. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2016)
24. Nebehay, G., Pflugfelder, R.: Clustering of static-adaptive correspondences for deformable object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2784–2791 (2015)
25. Pernici, F., Del Bimbo, A.: Object tracking by oversampling local features. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(12), 2538–2551 (2014)
26. Possegger, H., Mauthner, T., Bischof, H.: In defense of color-based model-free tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2113–2120 (2015)
27. Real, E., Shlens, J., Mazzocchi, S., Pan, X., Vanhoucke, V.: Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. arXiv preprint arXiv:1702.00824 (2017)
28. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015)
29. Song, Y., Ma, C., Gong, L., Zhang, J., Lau, R., Yang, M.H.: Crest: Convolutional residual learning for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
30. Sun, C., Lu, H., Yang, M.H.: Learning spatial-aware regressions for visual tracking. arXiv preprint arXiv:1706.07457 (2017)
31. Tao, R., Gavves, E., Smeulders, A.W.M.: Siamese instance search for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1420–1429 (2016)
32. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(11), 1958–1970 (2008)

33. Valmadre, J., Bertinetto, L., Henriques, J.F., Vedaldi, A., Torr, P.H.S.: End-to-end representation learning for correlation filter based tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
34. Wang, L., Liu, T., Wang, G., Chan, K.L., Yang, Q.: Video tracking using learned hierarchical features. IEEE Transactions on Image Processing **24**(4), 1424–1435 (2015)
35. Wang, N., Yeung, D.Y.: Learning a deep compact image representation for visual tracking. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 809–817 (2013)
36. Wang, Q., Teng, Z., Xing, J., Gao, J., Hu, W., Maybank, S.: Learning attentions: Residual attentional siamese network for high performance online visual tracking. In: The IEEE Conference on Computer Vision and Pattern Recognition (June 2018)
37. Wang, Q., Zhang, M., Xing, J., Gao, J., Hu, W., Maybank, S.: Do not lose the details: reinforced representation learning for high performance visual tracking. In: 27th International Joint Conference on Artificial Intelligence
38. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(9), 1834–1848 (2015)
39. Zhu, Z., Huang, G., Zou, W., Du, D., Huang, C.: Uct: Learning unified convolutional networks for real-time visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (Oct 2017)
40. Zhu, Z., Wu, W., Zou, W., Yan, J.: End-to-end flow correlation tracking with spatial-temporal attention. arXiv preprint arXiv:1711.01124 (2017)