Discriminative and Robust Online Learning for Siamese Visual Tracking

Jinghao Zhou, Peng Wang, Haoyang Sun

School of Computer Science and School of Automation, Northwestern Polytechnical University, China
National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, China
jensen.zhoujh@gmail.com, peng.wang@nwpu.edu.cn, sunhaoyang@mail.nwpu.edu.cn

Abstract

The problem of visual object tracking has traditionally been handled by variant tracking paradigms, either learning a model of the object's appearance exclusively online or matching the object with the target in an offline-trained embedding space. Despite the recent success, each method agonizes over its intrinsic constraint. The online-only approaches suffer from a lack of generalization of the model they learn thus are inferior in target regression, while the offline-only approaches (e.g., convolutional siamese trackers) lack the target-specific context information thus are not discriminative enough to handle distractors, and robust enough to deformation. Therefore, we propose an online module with an attention mechanism for offline siamese networks to extract target-specific features under $L2$ error. We further propose a filter update strategy adaptive to treacherous background noises for discriminative learning, and a template update strategy to handle large target deformations for robust learning. Effectiveness can be validated in the consistent improvement over three siamese baselines: *SiamFC*, *SiamRPN++*, and *SiamMask*. Beyond that, our model based on *SiamRPN++* obtains the best results over six popular tracking benchmarks and can operate beyond real-time.

1 Introduction

Visual object tracking is a fundamental topic in various computer vision tasks, such as vehicle navigation, automatic driving, visual surveillance, and video analytics. Briefly speaking, given the position of an arbitrary target of interest in the first frame of a sequence, visual object tracking aims to estimate its location in all subsequent frames.

In the context of visual object tracking, almost all the state-of-the-art trackers can be categorized into two categories: discriminative and generative trackers. Discriminative trackers train a classifier to distinguish the target from the background, while generative trackers find the object that can best match the target through computing the joint probability density between targets and search candidates.

Recently, the siamese framework (Bertinetto et al. 2016a) as a generative tracker has received surging attention in
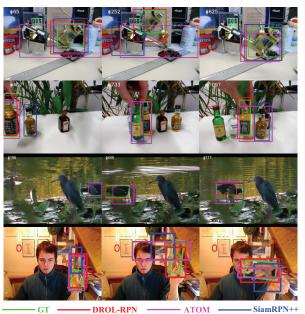
Figure 1: Qualitative demonstration and comparison. While accurate bounding box regression eludes *ATOM* and *SiamRPN++* struggles at situations like distractors and deformation, our method *DROL* outperforms *ATOM* with a siamese matching subnet, and SiamRPN++ with our proposed discriminative and robust online module.

that its performance has taken the lead in various benchmarks while running at real-time speed. Generally, these approaches obtain a similarity embedding space between subsequent frames and directly regress to the real state of the target (Li et al. 2018; Wang et al. 2019). Despite its recent success, since the siamese paradigm is exclusively trained offline, it suffers from severe limitations:

(I) Siamese approaches ignore the background information in the course of tracking, leading to inferior discriminative capability for target classification in the face of distractors.

(II) Siamese approaches solely utilize the first frame as

template or merely update it through averaging the subsequent frames, thus its performance degrades with huge deformation, rotation, and motion blur and results in poor robustness for target regression.

Probing into a unified solution to these issues arising from the parameter-fixed inherency of siamese tracking, we're motivated to introduce an online mechanism referring to the discriminative trackers which often feature a discriminative classifier and an efficient online update strategy. The unique strengths in online fashion, that is, the attainment of dynamic video-specific information, have long eluded siamese networks given its offline settings. However, the directly online updating siamese networks may be improper since any introduction of noises during tracking may deteriorate the generative ability of offline-trained embedding space.

Given the above intuitive analysis, a standalone online module that is systematically integrated but mutually independent with siamese networks is proposed as a complementary subnet. The online subnet is designed with an attention mechanism for extraction to most representative target features and is optimized efficiently to cater for time requirements. The response map of the online module is utilized for two earlier-mentioned issues separately. Specifically, limitation (I) is resolved by fusing the online response map with siamese classification score yielding an adaptive classification score for discriminative tracking, while limitation (II) is tackled by giving high-quality frames to siamese networks to update template for robust tracking.

To wrap up, we develop a highly effective visual tracking framework and establish a new state-of-the-art in terms of robustness and accuracy. The main contributions of this work are listed in fourfold:

- We propose an online module with attention mechanism optimized as a classification problem for siamese visual tracking, which can fully exploit background information to extract target-specific features.

- We propose discriminative learning using target-specific features via score fusion to help siamese networks handling distractors and background noises.

- We propose robust learning using target-specific features via template update to improve their robustness handling deformation, rotation, and illumination, etc.

- The proposed online module can be integrated with a variety range of siamese trackers without re-training them. Our method consistently set new state-of-the-art performance on 6 popular benchmarks of *OTB100*, *VOT2018*, *VOT2018-LT*, *UAV123*, *TrackingNet*, and *LaSOT*.

2 Related Work

Siamese visual tracking. Recently, the tracking methods based on siamese network (Bertinetto et al. 2016a) in the literature greatly outperformed other trackers due to heavy offline training, which largely enriches the depiction of the target. The following works include higher-fidelity object representation (Wang et al. 2019; Li et al. 2018), deeper backbone for feature extraction (Li et al. 2019a; Zhang and Peng 2019), and multi-stage or multi-branch network design

(Fan and Ling 2019; He et al. 2018). Beyond that, another popular optimization is template's update for robust tracking in the face of huge target's appearance change, either by using fused frames (Choi, Kwon, and Lee 2017a) or separately selected frames (Yang and Chan 2018). However, the lack of integrating background information degrades its performance on challenges like distractor, partial occlusions, and huge deformation. In fact, such rigidity is intrinsic given the fact that the object classification score from siamese framework turns out to be messy thus requires a centered Gaussian window to ensure stability. While many works intend to construct a target-specific space, by distractor-aware offline training (Zhu et al. 2018), residual attentive learning (Wang et al. 2018), gradient-based learning (Choi, Kwon, and Lee 2017b; Li et al. 2019b), this embedding space obtained through cross-correlation has yet to explicitly account for the distractors and appearance variation.

Online learning approach. Online learning is a dominant feature for discriminative trackers to precede, which is achieved by training an online component to distinguish the target object from the background. Particularly, the correlation-filter-based trackers (Henriques et al. 2014; Danelljan et al. 2017) and classifier-based trackers (Kalal, Mikolajczyk, and Matas 2011; Nam and Han 2016) are among the most representative and powerful methods. These approaches learn an online model of the object's appearance using hand-crafted features or deep features pre-trained for object classification. Given the recent prevalence of meta-learning framework, (Bhat et al. 2019; Park and Berg 2018) further learns to learn during tracking. Comparatively speaking, online learning for siamese-network-based trackers has had less attention. While previous approaches are to directly insert a correlation filter as a layer in the forward pass and update it online (Bertinetto et al. 2016b; Guo et al. 2017), our work focuses on the siamese's combination with an online classifier in a parallel manner with no need to retrain them.

Target regression and classification. Visual object tracking synchronously needs target regression and classification, which can be regarded as two different but related subtasks. While object classification has been successfully addressed by the discriminative methods, object regression is recently upgraded by the introduction of Regional Proposal Network (Li et al. 2018) in siamese network owing to the intrinsic strength for generative trackers to embed rich representation. Nevertheless, the majority of trackers in the gallery stand out usually with the strength within one subtask while degrading within the other. To design a high-performance tracker, lots of works have been focused on addressing these two needs in two separate parts. MBMD (Zhang et al. 2018) uses a regression network to propose candidates and classify them online. ATOM (Danelljan et al. 2019) combines a simple yet effective online discriminative classifier and high-efficiency optimization strategy with an IoUNet through overlap maximization, while the regression is achieved through a sparse sampling strategy compared with the dense one via cross-correlation in siamese networks.
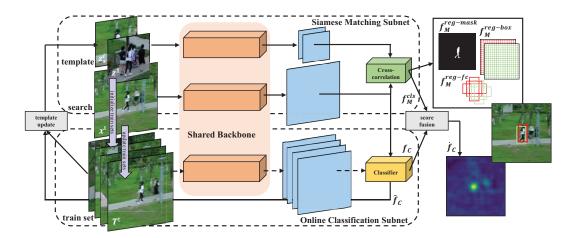
Figure 2: Overview of the network architecture for visual object tracking. It consists of siamese matching subnet M majorly accounting for bounding box regression and online classification subnet C generating classification score. Dashed line denotes reference scores passed from the classifier to the updater to select the short-term template z_s^t.

3 Proposed Method

The visual object tracking task can be formulated as a learning problem whose the primary task is to find the optimal target position by minimizing the following objective:

$$L(w) = \sum_{j=1}^{m} \gamma_j r(f(x_j; w), y_j) + \sum_{k} \lambda_k \|w_k\|^2, \quad (1)$$

where $r(f, w)$ computes the residual at each spatial location and $y_j \in \mathbb{R}^{W \times H}$ is the annotated label. The impact of each residual r is set by γ_j and regularization on w_k is set by λ_k.

3.1 Siamese Matching Subnet

Before moving to our approach, we firstly review the siamese network baseline. The siamese-network-based tracking algorithms take (1) as a template-matching task by formulating as a cross-correlation problem f_R and learn a embedding space $\phi(\cdot)$ that computes where in the search region can best match the template, shown as

$$f_M^{cls}(x, z) = \phi(x) * \phi(z) + b * \mathbb{1}, \quad (2)$$

where one branch learns the feature representation of the target z, and the other learns the search area x.

While the brute-force scale search to obtain target regression $f_M^{reg}(x, z; w)$ is insufficient in accuracy, siamese network extensions explicitly complement themselves with a Region Proposal Network (RPN) head or a mask head by separately regressing the bounding box and the mask of the target by encoding a subspace for box $[\cdot]^{reg-box}$ and mask $[\cdot]^{reg-mask}$. They each output the offsets for all prior anchors or a pixel-wise binary mask. These 3 variants of target regression can be formulated as:

$$
\begin{aligned}
f_M^{reg-fc}(x, z) &= \phi(x) * \phi(\hat{z}), \hat{z} \in \{z^i | i = 1, 2, 3\} \\
f_M^{reg-box}(x, z) &= [\phi(x)]^{reg-box} * [\phi(z)]^{reg-box} \quad (3) \\
f_M^{reg-mask}(x, z) &= [\phi(x)]^{reg-mask} * [\phi(z)]^{reg-mask}
\end{aligned}
$$

which is seperately adopted by our chosen baseline *SiamFC*, *SiamRPN++*, and *SiamMask*. $\hat{z}$ is the scaled template with maximal response value and $f_M^{reg-fc}(x, \hat{z})$ is a $1d$ vector, directly giving target center at pixel with maximal score. $f_M^{reg-box}(x, z)$ is a $4d$ vector which stands for the offsets of center point location and scale of the anchors to groudtruth. $f_M^{reg-mask}(x, z)$ is a $(63 \times 63)d$ vector encoding the spatial distribution of object mask.

The eventual predicted target state s is obtained using the same strategy as in (Bertinetto et al. 2016a; Li et al. 2018; 2019a; Wang et al. 2019). Additionally, the siamese network takes the residuals for classification commonly as cross-entropy loss, bounding box regression as smooth L_1 loss, and mask regression as logistic regression loss.

3.2 Target-specific Features

The features provided by siamese matching subnet which are trained offline and set fixed to ensure tracking stability and efficiency. However, they're not target-specific and adaptive to the target's domain. Inspired by discriminative trackers, in the initialization stage (the first frame), we propose to harness target-specific features on top of siamese features in a supervised manner by setting (1) as $L2$ loss yielding,

$$r_C(f, y_1) = \|f_C - y_1\|^2, \quad (4)$$

where the expected output f_C is the classification scores for each location belonging to the region of target and overall representing a spatial distribution of labels. y_1 is set to Gaussian according to the given target bounding box.

In light of related researches (Li et al. 2019b; Yang et al. 2019), the fitting of confidence scores for all negative background pixels dominates online learning by directly using the standard convolutions, as only a few convolutional filters play an important role in constructing each feature pattern or object category. The aforementioned data imbalance in both

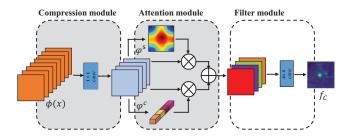**Compression module** · **Attention module** · **Filter module**

Figure 3: Network architecture of proposed online module, which consists of compression module to refine the feature map from backbone suitable for classification, attention module resolving data imbalance and filter module. φ^s denotes spatial attention consisting of 2 FC layers after global average pool (GAP) and φ^c channel-wise attention composed of a softmax after channel averaging.

spatial and channel-wise fashion degrades the model's discriminative capabilities.

To accommodate this issue, we introduce a dual attention mechanism to fully extract target-specific features. As shown in Figure 3, the compression module, and the attention module together form a target-specific feature extractor. Note that these 2 modules (gray area) are only fine-tuned with the first frame of given sequences and are kept fixed during tracking to ensure stability. The harnessed target-specific features are then leveraged to optimize the filter module (white area) in subsequent frames.

3.3 Discriminative Learning via Filter Update

Extensive study (Zhu et al. 2018) has proved siamese trackers can be easily distracted by a similar object during tracking even with distractor-aware (target-specific) features. A more profound reason for such inferiority is that there's no online weight update performed to suppress the treacherous background noises.

As a solution, by changing y_1 in (4) to y_i centered at predicted location, the update of filter module can be iteratively optimized as tracking proceeds. To further accelerate optimization, reformulating (1) into the squared norm of the residual vector $L(w) = \|r(w)\|^2$, where $r_j(w) = \sqrt{\gamma_j}(f(x_j) - y_j)$ and $r_{m+k}(w) = \sqrt{\lambda_k}w_k$, induces a positive definite quadratic problem. Instead of using the standard stochastic gradient descent (SGD) as the optimization strategy, following (Danelljan et al. 2019), we used Conjugate Gradient Descent better suited for solving quadratic problem in terms of convergence speed. This algorithm allows an adaptive search direction p and learning rate α during backpropagation in an iterative form.

The online classification score from the filter module is resized using cubic interpolation to the same spatial size as the siamese classification score and then fused with it through weighted sum yielding an adaptive classification score, which can be formulated as:

$$\dot{f}_C(x; w) = \lambda f_C(x; w) + (1 - \lambda)f_M^{cls}(x, z; w), \quad (5)$$

where λ is the impact of online confidence score.

Algorithm 1: Tracking algorithm

Input : Video frames $f^1, ..., f^t$ of length L;
Initial target state s^1;
Output: Target state s^t in subsequent frames;

Setting train set T for online classification subnet C and update with augmented search $\{\tilde{x}^1\}$;
Setting template z, z_s for siamese matching subnet M;
for $t = 2$ *to* L **do**
  Obtain the search x^t based on the previous predicted target state s^{t-1};
  // Track
  Obtain $f_C(x^t) = C(x^t)$ via classification subnet;
  Obtain $f_M^{cls}(x^t), f_M^{reg}(x^t) = R(x^t)$ using (2);
  Obtain current target state s^t based on $\dot{f}_C(x^t)$ using (5) and $f_M^{reg}(x^t)$;
  if $\delta_c(s^t) \geq \epsilon$ **then**
    $T^t \leftarrow T^{t-1} \cup x^t$;
    Train C with T^t;
  end
  if $(t \bmod T) = 0$ **then**
    $z_s \leftarrow z_s^t$ using (6);
  end
end

3.4 Robust Learning via Template Update

A crucial section of siamese trackers is the selection of templates, as the variation of target appearances often renders poor target regression and even target drift. We thus design an extra template branch for siamese networks to reserve the target information on the most recent frame. The template for this branch is denoted as z_s, representing the template in short-term memory as opposed to retaining the first frame for long-term memory.

It's worth noting that the key for good performance rests largely on the quality of identified templates, thus the selection of possible template candidates need to be designed reasonably. We use the maximal score $\hat{f}_C(x; w)$ of the output from online classifier as a yardstick to select high-quality templates, which can be formulated by

$$z_s = \text{warp}(\arg\max_x \hat{f}_C(x; w) * \mathbb{1}_{\hat{f}_C > \tau_c}), \quad (6)$$

where $\text{warp}(\cdot)$ is the cropping operation to feature size of the template and τ_c is filtering threshold suppressing low-quality template facing occlusion, motion blur, and distractors.

By shifting to the short-term template, z in (2) can be modified to z' which is decided with

$$z' = \begin{cases} z_s, & \text{IoU}(\hat{f}_M^{reg}(z_s), \hat{f}_M^{reg}(z^1)) \geq \upsilon_r, \\ & \hat{f}_M^{cls}(z_s) - \hat{f}_M^{cls}(z^1) \geq \upsilon_c, \\ z^1, & \text{otherwise}, \end{cases} \quad (7)$$

where $\hat{f}_M^{cls}$ denotes location with maximal score and $\hat{f}_M^{reg}$ its corresponding bounding box. This is to avoid bounding box drift and inconsistency using the short-term template. The full pipeline of our method is detailed in Algorithm 1.
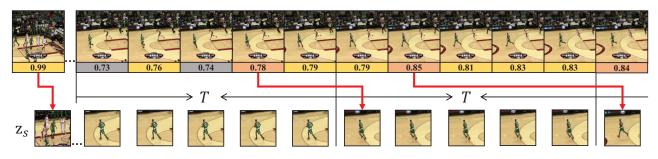
Figure 4: Visualized process of proposed template update strategy on sequence *Basketball* of OTB2015. The frame with the highest score in the historical frames will be selected to update the short-term template z_s.

4   Experiments

4.1   Implementation Details

Our method is implemented in Python with PyTorch, and the complete code and video demo will be made available at https://github.com/shallowtoil/DROL.

Architecture. Corresponding to the regression variants as described in (3), we apply our method in three siamese baselines *SiamFC*, *SiamRPN++*, and *SiamMask* yielding *DROL-FC*, *DROL-RPN*, *DROL-Mask* respectively. In *DROL-FC*, we use the modified AlexNet as the backbone and up-channel correlation. In *DROL-RPN*, we use the modified AlexNet and layer $2, 3, 4$ of ResNet50 and depth-wise correlation. While in *DROL-Mask*, we use layer 2 of ResNet50 and depth-wise correlation. For the classification subnet, the first layer is a 1×1 convolutional layer with ReLU activation, which reduces the feature dimensionality to 64. The last layer employs a 4×4 kernel with a single output channel.

Training phase. For offline training, since we directly use the off-the-shelf siamese models, no extra offline training stage for the whole framework is required. The training data for our models thus is completely the same as those variants' baselines. For online tuning, we use the region of size 255×255 of the first frame to pre-train the whole classifier. Similar to (Danelljan et al. 2019), we also perform data augmentation to the first frame of translation, rotation, and blurring, yielding total 30 initial training samples.

Discriminative learning. We add the coming frames to initial training samples with a maximal batch size of 250 by replacing the oldest one. Each sample is labeled with a Gaussian function centered at predicted target location. We discard the frames with the occurrence of distractors or target absence for filter update. The classifier is updated every 10 frame with a learning rate set to 0.01 and doubled once neighboured distractors are detected. To fuse classification scores, we set λ to 0.6 in *DROL-FC* and 0.8 in *DROL-RPN* and *DROL-Mask*.

Robust learning. For template update, to strike a balance between stability and dynamicity when tracking proceeds, we update the short-term template every $T = 5$ frames, while τ_c, υ_r, and υ_c are set to 0.75, 0.6, and 0.5 respectively. The above hyper-parameters are set using VOT2018 as the validation set and are further evaluated in Section 5.

Method	Backbone	OTB2015		VOT2018		FPS
		AUC	Pr	EAO	A	
ECO	VGGNet	0.694	0.910	0.280	0.484	8
UPDT	ResNet50	*0.702*	-	0.378	0.536	-
MDNet	VGGNet	0.678	0.909	-	-	1
ATOM	ResNet18	0.669	0.882	0.401	0.590	30
DiMP	ResNet50	0.684	0.894	*0.440*	0.597	40
SiamFC	AlexNet	0.582	0.771	0.206	0.527	120
DROL-FC	AlexNet	0.619	0.848	0.256	0.502	60
SiamRPN	AlexNet	0.637	0.851	0.326	0.569	**200**
DaSiamRPN	AlexNet	0.658	0.875	0.383	0.586	*160*
SiamRPN++	AlexNet	0.666	0.876	0.352	0.576	180
	ResNet50	0.696	*0.915*	0.414	0.600	35
DROL-RPN	AlexNet	0.689	0.914	0.376	0.583	80
	ResNet50	**0.715**	**0.937**	**0.481**	**0.616**	30
SiamMask	ResNet50	-	-	0.347	0.602	56
DROL-Mask	ResNet50	-	-	0.434	*0.614*	40

Table 1: state-of-the-art comparison on two popular tracking benchmarks OTB2015 and VOT2018 with their running speed. AUC: area under curve; Pr: precisoin; EAO: expected average overlap; A: accuracy; FPS: frame per second. The performance is evaluated using the best results over all the settings proposed in their original papers. The speed is tested on Nvidia GTX 1080Ti GPU.

4.2   Comparison with state-of-the-art

We first showcase the consistence improvement of DROL with each of its baseline on two popular benchmarks OTB2015 (Wu, Lim, and Yang 2015) and VOT2018 (Kristan et al. 2018), as shown in Table 1. We further verify our best model *DROL-RPN* with backbone ResNet50 on other four benchmarks: VOT2018-LT, UAV123 (Mueller, Smith, and Ghanem 2016), TrackingNet (Muller et al. 2018), and LaSOT (Fan et al. 2019).

OTB2015. We validate our proposed tracker on the OTB2015 dataset, which consists of 100 videos. Though the latest work SiamRPN++ unveiling the power of deep neural network, as shown in Figure 1, the best siamese trackers still suffer the agonizing pain from distractors, full occlusion and deformation in sequences like *Baseketball*, *Liquor* and *Skating1*. Comparatively speaking, our approach can robustly discriminate the aforementioned challenges thus obtain a desirable gain above the siamese baseline through on-
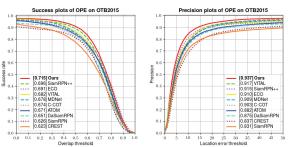
Figure 5: state-of-the-art comparison on the OTB2015 dataset in terms of success rate (left) and precision (right).

line learning. Figure 5 and Table 1 show that DROL-RPN achieves the top-ranked result in both AUC and Pr. Compared with previous best tracker UPDT, we improve 1.3% in overlap. While compared with ATOM and SiamRPN++, our tracker achieves a performance gain of 4.6%, 1.9% in overlap and 5.5%, 2.2% in precision respectively.

VOT2018. Compared with OTB2015, VOT2018 benchmark includes more challenging factors, thus may be deemed as a more holistic testbed on both accuracy and robustness. We validate our proposed tracker on the VOT2018 dataset, which consists of 60 sequences.

As Figure 6(a) shown, we can observe that our tracker DROL-RPN improves the ATOM and SiamRPN++ method by an absolute gain of 8.0% and 6.6% respectively in terms of EAO. Our approach outperformed the leading tracker DiMP by 4.8% in overlap and 1.9% in accuracy.

VOT2018-LT. VOT2018 Long-term dataset provides a challenging yardstick on robustness as the target may leave the field of view or encounters full occlusions for a long period. We report these metrics compared with the state-of-the-art method in Figure 6(b) and Table 2. We directly adopting the long-term strategy the same as SiamRPN++ (Li et al. 2019a) in our siamese subnet. When the target absence is detected and the search region is enlarged, we don't add any search at these moments into training sets to train the classifier. Compared with our baseline SiamRPN++, our approach achieves a maximum performance gain of 3.8% in precision, 3.8% in recall, and 3.8% in F-score respectively. Results show that even encountered huge deformation and long-term target absence, the online classification subnet can still perform desirably.

UAV123. We evaluate our proposed tracker on the UAV123 dataset. The dataset contains more than $110K$ frames and a total of 123 video sequences, each with on average 915 frames captured from low-altitude UAVs. Table 3 showcases the overlap and precision of the compared tracker. We compare our trackers with previous state-of-the-arts and results show that our approach outperforms the others by achieving an AUC score of 65.7%. Our tracker outperforms the SiamRPN++ baseline by a large margin of 4.4% in AUC score and 4.5% in precision.

TrackingNet. We validate our proposed tracker on the TrackingNet *test* dataset, which provides more than $30K$ videos with 14 million dense bounding box annotations. As illustrated in Table 4, our approach DROL-RPN outperforms
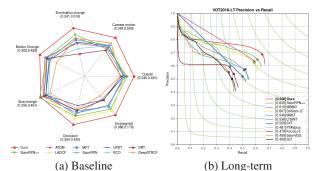

(a) Baseline   (b) Long-term

Figure 6: state-of-the-art comparison on VOT2018 Baseline (left) and Long-term (right) dataset. In right one, the maximal and minimal value for each attribute is shown in pair, and the rest are permuted propotionally. In left one, the average precision-recall curves are shown above, each with its maximal F-score pointed out at corresponding point.

	MMLT	DaSiam-LT	MBMD	SiamRPN++	**DROL-RPN**
Pr↑	0.574	0.627	0.642	*0.649*	**0.687**
Re↑	0.521	0.588	0.583	*0.612*	**0.650**

Table 2: state-of-the-art comparison on the VOT2018-LT dataset in terms precision (Pr) and Recall (Re).

	ECO	SiamRPN	SiamRPN++	ATOM	DiMP	**DROL-RPN**
AUC	0.525	0.527	0.613	0.631	*0.643*	**0.652**
Pr	0.741	0.748	0.807	0.843	*0.851*	**0.855**

Table 3: state-of-the-art comparison on the UAV123 dataset in terms of success (AUC), and precision (Pr).

	MDNet	ATOM	SPM	SiamRPN++	DiMP	**DROL-RPN**
AUC	0.614	0.703	0.712	0.733	*0.740*	**0.746**
Pr	0.555	0.648	0.661	*0.694*	0.687	**0.708**
NPr	0.710	0.771	0.778	0.800	*0.801*	**0.817**

Table 4: state-of-the-art comparison on the TrackingNet *test* dataset in terms of success (AUC), precision (Pr) and normalized precision (NPr).
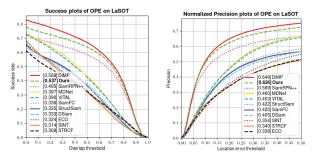

Figure 7: state-of-the-art comparison on LaSOT dataset in terms of success rate (left) and normalized precision (right).

all previous methods and improves the performance of state-of-the-art by 0.6% in overlap, 1.4% in precision, and 1.6% in normalized precision.

LaSOT. We validate our proposed tracker on the LaSOT dataset to further validate its discriminative and genera-

tive capability. Figure 7 reports the overall performance of DROL-RPN on LaSOT testing set. Specifically, our tracker achieves an AUC of 53.7% and an NPr of 62.4%, which outperformed ATOM and SiamRPN++ without bells and whistles.

4.3 Ablation Study

In this part, we perform ablation analysis to demonstrate the impact of each component in groups and illustrate the superiority of our method (only *DROL-RPN* with backbone ResNet50 is showcased for brevity). Since we search the hyper-parameters using VOT2018 (validation set), we report the performance with LaSOT and TrackingNet (test set).

Classification score fusion (G1). As data reveals, there's a synchronous need of scores from two subnets as two extremes with λ being 0 or 1 induce poor results while the optimal coefficient falls around 0.8. Theoretically, it can be partly explained by that, though discriminative enough, online classification score only locates a rough location, while more fine-grained details distinguishing the quality of dense bounding boxes around the peak region need to be fleshed out by the siamese classification score.

Target-specific feature learning (G2). With the best settings in G1, we further delve into the investigation of the proposed attention mechanism. As shown in Table 5, the existence of the attention module further improves trackers' performance. Qualitatively, in the comparison between Figure 8(b) and Figure 8(c), the online classification score is more discriminative to deal with distractors and occlusions. The attention module uplifts this discriminative capability given the fact the peak on ground-truth target is more focused as shown in Figure 8(c) and Figure 8(d).

Template update interval (G3). From Table 5, compared with directly updating the short-term sample once it obtains a qualified classification score by setting $T = 1$, we find that properly prolonging the update interval and selecting one with the highest classification confidence after sorting allows us a more trust-worthy template. To demonstrate the validity of the update strategy, we provide IoU curve of 2 sequences. As Figure 9(a) shown, short-term template improves the performance especially when putative box overlaps the ground truth not well, and Figure 9(b) shows such strategy can even avoid target drift in the long run.

Validity of update strategy (G4). To evaluate the effectiveness of our update strategy, G4 is conducted by only using the online classification score to guide template update without fusing it with siamese classification score. Compared with the baseline (see the first column in G1), the validity of updater can be more clearly observed.

5 Conclusion

In this work, we propose an online module with a dual attention mechanism for siamese visual tracking to extract video-specific features under a quadratic problem. The discriminative learning is performed by updating the filter module during tracking, which enhances siamese network's discriminative ability to deal with distractors. And the robust learning is achieved by an extra short-term template branch updating

group	λ	pipeline	T	LaSOT		TrackingNet	
				AUC	NPr	AUC	NPr
G1	0.0	*Cmp*	—	0.508	0.586	0.733	0.800
	0.2			0.514	0.596	0.737	0.808
	0.4			0.521	0.605	0.738	0.807
	0.6			0.528	0.615	0.739	0.810
	0.8			**0.532**	**0.619**	**0.743**	**0.814**
	0.9			0.525	0.612	0.740	0.811
	1.0			0.391	0.542	0.576	0.749
G2	0.8	*Att-Cmp*	—	0.528	0.610	0.741	0.812
		Cmp-Att		**0.534**	**0.621**	**0.745**	**0.817**
G3	0.8	*Cmp-Att*	1	0.526	0.614	0.744	0.815
			5	**0.537**	**0.624**	**0.746**	**0.817**
			10	0.533	0.620	0.745	0.817
G4	0.0	*Cmp*	5	0.513	0.592	0.738	0.805

Table 5: Ablation study of different modules and setups. λ is fusion weight of classification scores from 2 subnets. *Cmp*, *Att* represent *compresson module* and *attention module* respectively. T denotes the update interval for short-term, and − means no update is applied.

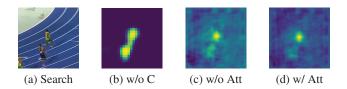(a) Search   (b) w/o C   (c) w/o Att   (d) w/ Att

Figure 8: Visualization on final classification output after the Classification subnet *C* and attention module is used on sequence *Bolt* of OTB2015.
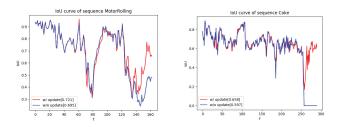

Figure 9: IoU curve and expected average IoU on sequence *MotorRolling* (left) and *Coke* (right) of OTB2015.

dynamically based on scores from the classifier to handle huge deformation. Without bells and whistles, our proposed online module can easily work with any siamese baselines without degenerating its speed too much. Extensive experiments on 6 major benchmarks are performed to demonstrate the validity and superiority of our approach.

References

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. 2016a. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, 850–865. Springer.

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.;

and Torr, P. H. 2016b. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, 850–865. Springer.

Bhat, G.; Danelljan, M.; Van Gool, L.; and Timofte, R. 2019. Learning discriminative model prediction for tracking. *arXiv preprint arXiv:1904.07220*.

Choi, J.; Kwon, J.; and Lee, K. M. 2017a. Deep meta learning for real-time visual tracking based on target-specific feature space. *arXiv preprint arXiv:1712.09153*.

Choi, J.; Kwon, J.; and Lee, K. M. 2017b. Deep meta learning for real-time visual tracking based on target-specific feature space. *arXiv preprint arXiv:1712.09153*.

Danelljan, M.; Bhat, G.; Shahbaz Khan, F.; and Felsberg, M. 2017. Eco: efficient convolution operators for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6638–6646.

Danelljan, M.; Bhat, G.; Khan, F. S.; and Felsberg, M. 2019. Atom: Accurate tracking by overlap maximization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fan, H., and Ling, H. 2019. Siamese cascaded region proposal networks for real-time visual tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; and Ling, H. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5374–5383.

Guo, Q.; Feng, W.; Zhou, C.; Huang, R.; Wan, L.; and Wang, S. 2017. Learning dynamic siamese network for visual object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, 1763–1771.

He, A.; Luo, C.; Tian, X.; and Zeng, W. 2018. Towards a better match in siamese network based visual object tracker. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 0–0.

Henriques, J. F.; Caseiro, R.; Martins, P.; and Batista, J. 2014. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence* 37(3):583–596.

Kalal, Z.; Mikolajczyk, K.; and Matas, J. 2011. Tracking-learning-detection. *IEEE transactions on pattern analysis and machine intelligence* 34(7):1409–1422.

Kristan, M.; Leonardis, A.; Matas, J.; Felsberg, M.; Pflugfelder, R.; Cehovin Zajc, L.; Vojir, T.; Bhat, G.; Lukezic, A.; Eldesokey, A.; et al. 2018. The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 0–0.

Li, B.; Wu, W.; Zhu, Z.; Yan, J.; and Hu, X. 2018. High performance visual tracking with siamese region proposal network. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019a. Siamrpn++: Evolution of siamese visual tracking

with very deep networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, X.; Ma, C.; Wu, B.; He, Z.; and Yang, M.-H. 2019b. Target-aware deep tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1369–1378.

Mueller, M.; Smith, N.; and Ghanem, B. 2016. A benchmark and simulator for uav tracking. In *European conference on computer vision*, 445–461. Springer.

Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; and Ghanem, B. 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 300–317.

Nam, H., and Han, B. 2016. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4293–4302.

Park, E., and Berg, A. C. 2018. Meta-tracker: Fast and robust online adaptation for visual object trackers. *Lecture Notes in Computer Science* 587–604.

Wang, Q.; Teng, Z.; Xing, J.; Gao, J.; Hu, W.; and Maybank, S. 2018. Learning attentions: residual attentional siamese network for high performance online visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4854–4863.

Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; and Torr, P. H. 2019. Fast online object tracking and segmentation: A unifying approach. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, Y.; Lim, J.; and Yang, M.-H. 2015. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(9):1834–1848.

Yang, T., and Chan, A. B. 2018. Learning dynamic memory networks for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 152–167.

Yang, K.; Song, H.; Zhang, K.; and Liu, Q. 2019. Hierarchical attentive siamese network for real-time visual tracking. *Neural Computing and Applications* 1–12.

Zhang, Z., and Peng, H. 2019. Deeper and wider siamese networks for real-time visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4591–4600.

Zhang, Y.; Wang, D.; Wang, L.; Qi, J.; and Lu, H. 2018. Learning regression and verification networks for long-term visual tracking. *arXiv preprint arXiv:1809.04320*.

Zhu, Z.; Wang, Q.; Li, B.; Wei, W.; Yan, J.; and Hu, W. 2018. Distractor-aware siamese networks for visual object tracking. *The European Conference on Computer Vision (ECCV)*.