Graphical Abstract (Optional)

To create your abstract, please type over the instructions in the template box below. Fonts or abstract dimensions should not be changed or altered.

Target-Cognisant Siamese Network for Robust Visual Object Tracking

Yingjie Jiang, Xiaoning Song, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Josef Kittler

Siamese trackers have become the mainstream framework for visual object tracking in recent years. However, the extraction of the template and search space features is disjoint for a Siamese tracker, resulting in a limited interaction between its classification and regression branches. This degrades the model capacity accurately to estimate the target, especially when it exhibits severe appearance variations. To address this problem, this paper presents a target-cognisant Siamese network for robust visual tracking. First, we introduce a new target-cognisant attention block that computes spatial cross-attention between the template and search branches to convey the relevant appearance information before correlation. Second, we advocate two mechanisms to promote the precision of obtained bounding boxes under complex tracking scenarios. Last, we propose a max filtering module to utilise the guidance of the regression branch to filter out potential interfering predictions in the classification map. The experimental results obtained on challenging benchmarks demonstrate the competitive performance of the proposed method.

Research Highlights (Required)

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- A new Target-Cognisant Siamese-based anchor-free tracker.
- The proposed method computes cross-spatial attention for refining the measurement of spatial similarity.
- Two tracking mechanisms are used to promote the precision of bounding box prediction.
- A max filtering module is proposed to filter out similar distractors.
- Our method achieves competitive performance on several tracking datasets.

Pattern Recognition Letters
journal homepage: www.elsevier.com

Target-Cognisant Siamese Network for Robust Visual Object Tracking

Yingjie Jiang[a], Xiaoning Song[a,**], Tianyang Xu[a,**], Zhenhua Feng[b,c], Xiaojun Wu[a], Josef Kittler[c]

[a]*School of Artificial Intelligence and Computer Science, Jiangnan University, 214122, Wuxi, China*
[b]*Department of Computer Science, University of Surrey, GU2 7XH, Guildford, UK*
[c]*Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH, Guildford, UK*

ABSTRACT

Siamese trackers have become the mainstream framework for visual object tracking in recent years. However, the extraction of the template and search space features is disjoint for a Siamese tracker, resulting in a limited interaction between its classification and regression branches. This degrades the model capacity accurately to estimate the target, especially when it exhibits severe appearance variations. To address this problem, this paper presents a target-cognisant Siamese network for robust visual tracking. First, we introduce a new target-cognisant attention block that computes spatial cross-attention between the template and search branches to convey the relevant appearance information before correlation. Second, we advocate two mechanisms to promote the precision of obtained bounding boxes under complex tracking scenarios. Last, we propose a max filtering module to utilise the guidance of the regression branch to filter out potential interfering predictions in the classification map. The experimental results obtained on challenging benchmarks demonstrate the competitive performance of the proposed method.

Keywords: Visual Object Tracking, Siamese Network, Anchor-free Regression

1. Introduction

Visual object tracking is a fundamental task in computer vision. It has many practical applications, *e.g.*, human-computer interaction [1], autonomous driving [2], video surveillance [3], etc. Visual object tracking aims to localise an arbitrary target in each frame of a video, given its state in the first frame. Despite the remarkable progress to date, it is still very challenging to achieve high-performance tracking in the presence of unpredictable changes of the target appearance and its surroundings, due to phenomena such as illumination changes, occlusion, scale variation, etc.

In recent years, different from online-learning Discriminative Correlation Filters (DCF) [4, 5, 6], Siamese-based trackers have attracted a wide attention due to their efficient pairwise matching formulation and the ability to extract robust features. The pioneering studies, SINT [7] and SiamFC [8], formulate the task as a matching problem, and use a Siamese network to extract the target and instance representations for measuring their similarity. Inspired by that, many Siamese trackers [9, 10, 11, 12, 13, 14] have been proposed, producing increasingly promising results. These methods can be divided into two main categories: anchor-based and anchor-free methods. Anchor-based methods usually perform better in terms of accuracy, while anchor-free trackers are more flexible, avoiding additional hyper-parameters associated with candidate boxes.

Siamese-based trackers are trained offline to produce robust feature embedding, and use fixed target models for inference. Hence the template cannot be updated online, leading to poor performance in the presence of target appearance deformation or scale changes. Moreover, the features of the template and search window content are computed independently, before cross-correlation, resulting in limited interaction. However, a close information interaction between the template and the candidates is very important for the accurate localisation of the target. Most existing studies [15, 16] focus on enhancing the template representation of the first frame, while only a few of them simultaneously improve the feature representation of

**Corresponding author
e-mail: x.song@jiangnan.edu.cn (Xiaoning Song), tianyang.xu@jiangnan.edu.cn (Tianyang Xu)

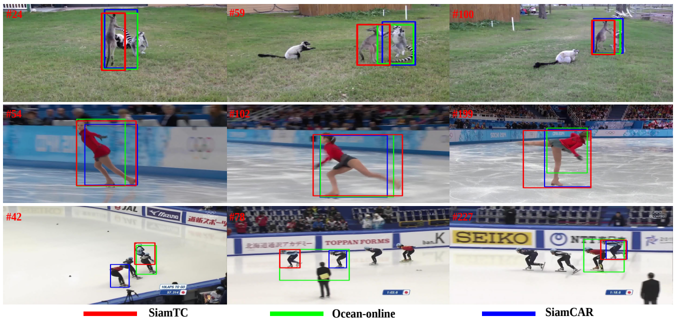


Fig. 1. A comparison of our SiamTC with Ocean-online [14] and SiamCAR [18] on 3 challenging videos. SiamTC predicts better bounding boxes than the others under the scenarios of object deformation, background clutter and scale change.

the template and the candidates [17]. However, the ability of anchor-free Siamese trackers accurately to estimate the target state cannot be guaranteed in complex tracking scenes.

To mitigate the above issues, we propose a new Target-Cognisant Siamese-based anchor-free tracker, namely SiamTC, to enable information propagation for robust and accurate tracking. In contrast to the existing cross-channel attention methods used in anchor-based trackers [17], we design a new target-cognisant attention block for an *anchor-free* tracker to gauge the spatial similarity between the template and search branches, achieving cross-spatial attention. The block implicitly updates the template to enhance the target representation, so as to compensate for dynamic appearance changes, and ultimately improve accuracy. It should be noted that the target-cognisant attention block is different with the attention-based feature fusion network used in the Transformer tracking method [19]. Transformer tracking mainly focuses on replacing the correlation operation in an attention-based manner, with additional position encoding to facilitate correspondence. Two simple but effective tracking mechanisms are further introduced to obtain more precise bounding boxes, without much computational overhead. Last, we present a max filtering module to discriminate between similar classification predictions based on a regression process. A comparison of our SiamTC with two state-of-the-art approaches presented in Fig.1, demonstrates the superiority of our method for high-precision bounding box prediction.

The main contributions of the proposed SiamTC method include:

- A novel Target-Cognisant Attention Block (TCAB) for refining the measurement of spatial similarity. We compute the cross-spatial attention between the template of a target and search windows in a video. TCAB is the key architectural element of the novel SiamTC method for robust object tracking.
- We adaptively aggregate multi-level correlation features and refine bounding boxes for accurate target state estimation, at the cost of only a small number of additional parameters. Further, a max filtering module to filter out similar distractors is presented.
- Experiments conducted on four challenging datasets, *i.e.* OTB100 [20], VOT2019 [21], GOT-10k [22] and LaSOT [23], demonstrate the merits and superiority of the proposed method.

2. Related Work

In recent years, Siamese-based trackers have become the mainstream methodology in visual object tracking, thanks to their end-to-end offline training capability and the high efficiency in online testing, achieving superior performance against the online-learning DCFs [25, 26]. SINT [7] is the first Siamese-based tracker in this category. It inspired the SiamFC [8] method, which uses a fully-convolutional Siamese network with cross-correlation for response calculation thus greatly improves the efficiency as compared to SINT. Many recent trackers are based on the SiamFC architecture, such as DSiam [15] and SA-siam [27]. However, all the above approaches use a multi-scale pyramid to address target scale changes, resulting in insufficient accuracy of the bounding box estimation.

To mitigate this problem, SiamRPN [10] proposed to use Region Proposal Network (RPN) [28] and achieved a much better performance in bounding box estimation accuracy. Based on SiamRPN, DaSiamRPN [29] focuses on improving the discriminative ability of the tracker through a distractor-aware feature learning scheme. Similarly, SPM [30] proposed a series-parallel matching framework to enhance the robustness and the capacity to discriminate the target from the background. SiamDW [12] and SiamRPN++ [11] suggested using deeper networks, such as ResNet [31], for performance boosting. On the other hand, SiamMask [32] uses fully-convolutional Siamese networks to produce class-agnostic binary segmentation masks for fine-grained object tracking. To facilitate the interaction of the tracking information conveyed by feature maps, SiamAttn [17] jointly advocated deformable self-attention and cross channel attention mechanisms to enhance the discriminative content of the target representation.

However, these anchor-based trackers are sensitive to the hyper-parameters of anchors. To eliminate the predefined set of anchors, many anchor-free trackers, *e.g.*, SiamBAN [13], SiamCAR [18] and Ocean [14], have been proposed. These methods directly classify each candidate target centre point and predict the distances of the four borders to the ground-truth. Despite their simplicity and the promising results achieved by offline anchor-free Siamese trackers, their accuracy and robustness are constrained by the limited supervision imposed in the online tracking stage. To address this issue, we propose a novel Target-Cognisant Attention Block (TCAB) to enhance the target information propagation. The robustness to appearance variations is also improved in our design by computing the cross spatial similarity. Recently, the Transformer tracking method [19] proposes a novel attention-based feature fusion module to replace the correlation operation in Siamese trackers, demonstrating the potential of attention mechanisms in visual object tracking.

More recently, FCOS [33] based anchor-free detectors have attracted considerable attention in general object detection. For example, GFL [34] and VFNet [35] use the generalised focal loss and varifocal loss to alleviate the inconsistencies between localisation quality and classification score. Moreover, GFLv2 [36] exploits the guidance from the statistics of bounding box distributions to facilitate reliable localisation estimation. Besides, E2ENet [37] introduces the concept of

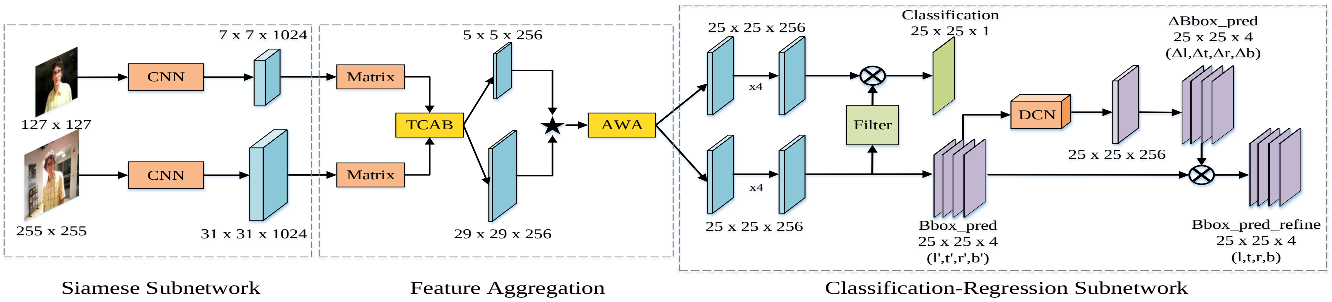


Fig. 2. Overview of the proposed tracking framework, consisting of the Siamese subnetwork, feature aggregation and classification regression subnetwork with a novel bounding box refinement module and max filtering module (MFM). TCAB and AWA denote the proposed Target-Cognisant Attention Block and Adaptive Weight Aggregation, respectively. Matrix represents the three parallel irregular dilation convolution layers utilised in the baseline method [14] and ⋆ denotes the depth-wise cross-correlation operation. The Matrix module outputs two branches of feature maps at three different scales, and then performs the TCAB and depth-wise cross-correlation operations to obtain three different correlation maps. These correlation maps are used by AWA for feature aggregation. For simplification, we only draw the process of the feature maps of a single scale in the feature aggregation stage. DCN stands for the deformable convolution layer [24].

prediction-aware one-to-one label assignment to enable end-to-end detection. Although the improved anchor-free approaches is popular in the detection field, they have not been sufficiently investigated in the tracking community. To bridge this gap, we propose an anchor-free Siamese tracker employing two new tracking mechanisms with a max filtering module to improve its accuracy and robustness.

3. The Proposed Method

We take the Ocean-offline [14] (without object-aware module) as our baseline method to establish an efficient anchor-free tracking framework. Based on this, we propose a novel SiamTC tracker, as shown in Fig. 2, which has three main components: 1) a Target-Cognisant Attention Block (TCAB), enhancing the feature robustness against appearance variations; 2) two accurate tracking mechanisms providing Adaptive Weight Aggregation (AWA) and a bounding box refinement; 3) a Max Filtering Module (MFM) that callbacks the ability of regression to improve the discrimination of the tracker in the adjacent regions.

3.1. The Target-Cognisant Attention Block

Siamese trackers rely on the correspondence of the target appearance conveyed by the template and search window, with reduced ability to discriminate the target from distractors, especially when faced with extreme background clutter. We attribute the above deficiency to the lack of communication between the template and search branches, resulting in inference uncertainty affecting the online tracking process. We argue that the extraction of mutual information for accurate and robust target identification is of paramount importance and this viewpoint motivates the innovations presented in this paper.

Different from previous channel attention or spatial self-attention [17], we propose a Target-Cognisant Attention Block (TCAB) using spatial cross-attention to enhance the consistency of appearance of the template and target candidate before correlation, as shown in Fig. 3(a). By design, the search branch learns class-agnostic target information, while the template branch provides discriminative feature representations, jointly improving the model capacity to discriminate between similar interferences. It takes a pair of convolutional features computed by the feature encoding network as inputs, and outputs a feature pair spatially enhanced by cross-attention.

Given the template feature map $\mathbf{Z} \in \mathbb{R}^{C\times h\times w}$ and the search window feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, we first apply two 1×1 convolution layers to $\mathbf{Z}$ and $\mathbf{X}$ to generate $\mathbf{\Phi} \in \mathbb{R}^{C'\times h\times w}$ and $\mathbf{\Theta} \in \mathbb{R}^{C'\times H\times W}$, where $C' = \frac{C}{2}$. These two feature maps are fed into two parallel branches to generate pixel-wise spatial cross-attention maps. Let's take the search subbranch as an example. $\mathbf{\Phi}$ and $\mathbf{\Theta}$ are reshaped to $\bar{\mathbf{\Phi}} \in \mathbb{R}^{C'\times n}$ and $\bar{\mathbf{\Theta}} \in \mathbb{R}^{C'\times N}$, where $n = h \times w$, $N = H \times W$. The cross-attention map $\mathbf{CA}^X$ of $\mathbf{X}$ is obtained by using the Global Max Pooling (GMP):

$$\mathbf{CA}^X = reshape(GMP(\bar{\mathbf{\Theta}}^T \times \bar{\mathbf{\Phi}})) \in \mathbb{R}^{H\times W}. \tag{1}$$

Finally, the corresponding search window feature $\mathbf{X}$ is multiplied by the cross-attention map $\mathbf{CA}^X$ to estimate a spatial similarity between the two frames, and the result added to get a spatially enhanced search feature $\mathbf{EA}^X$:

$$\mathbf{EA}^X = \mathbf{CA}^X \times \mathbf{X} + \mathbf{X} \in \mathbb{R}^{C\times H\times W}. \tag{2}$$

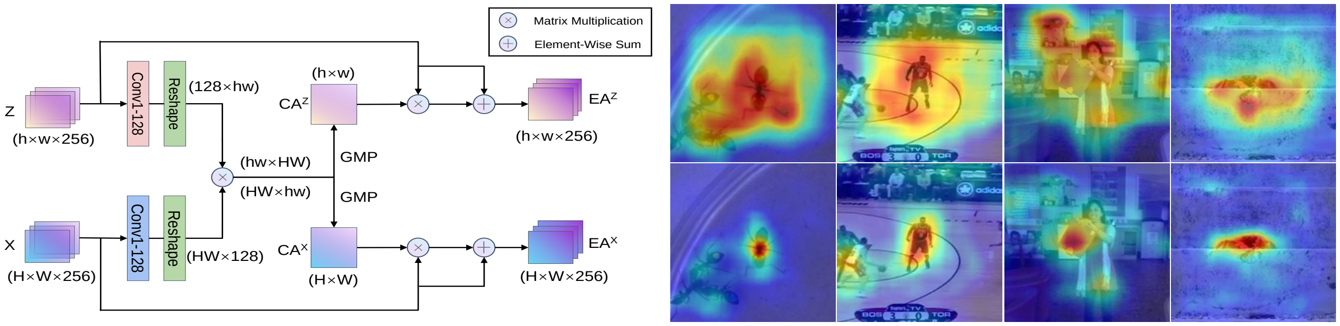
Similarly, an enhanced template feature $\mathbf{EA}^Z$ is obtained in a similar way:

$$\mathbf{CA}^Z = reshape(GMP(\bar{\mathbf{\Phi}}^T \times \bar{\mathbf{\Theta}})) \in \mathbb{R}^{h\times w}. \tag{3}$$

$$\mathbf{EA}^Z = \mathbf{CA}^Z \times \mathbf{Z} + \mathbf{Z} \in \mathbb{R}^{C\times h\times w}. \tag{4}$$

3.2. Accurate Tracking Mechanisms

As the geometric properties, such as scale and aspect ratio of the target, usually change in a video, an accurate prediction of the target state is crucial to achieving high-performance tracking. In contrast to the existing trackers, we use two simple but effective mechanisms to obtain more precise bounding boxes.



(a) Structure

(b) Visualisation

Fig. 3. The structure and visualisation of the Target-Cognisant Attention Block (TCAB), as detailed in Sec. 3.1. On the right hand side, the first row indicates the features produced by the baseline method, and the second row shows the attention feature furnished by TCAB.

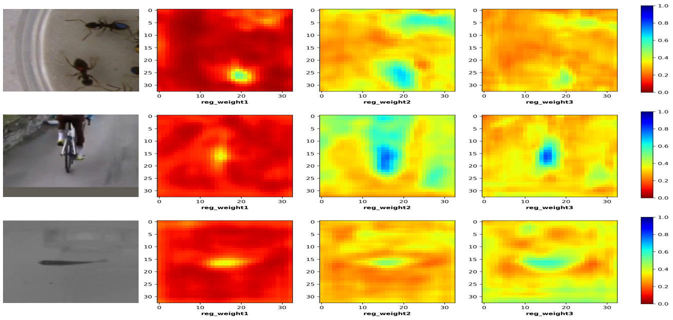


Fig. 4. Visualisation results of our proposed adaptive weight aggregation maps in three different video sequences on VOT2019. From top to bottom are ants, wheels and zebrafish.

3.2.1. *Adaptive Weight Aggregation*

In existing tracking methods [11, 13, 14], layer-wise aggregation is used to fuse multi-level deep network features. This is achieved by learning the layer-level weights corresponding to each correlation map, and optimising them together with the network. During the test stage, the learned layer-level weights are fixed. Take Ocean [14] as an example, denote the correlation map corresponding to the output of the i-th dilated convolution layer as $\mathcal{S}_i$. The layer-wise aggregation can be formulated as:

$$\mathcal{S}_{all} = \sum_{i=1}^{3} \alpha_i * \mathcal{S}_i \tag{5}$$

where α_1, α_2 and α_3 are 0.3071, 0.3368 and 0.3561 respectively for regression branch. In SiamRPN++ [11], the fusion weights corresponding to the features from last three residual blocks are 0.1764, 0.1656 and 0.6579, respectively. They both show that features from large dilation convolution layers contribute more to bounding box regression.

Although this plain fusion method boosts the performance, the fusion of the extracted deep features is not necessarily effective from the fine-grain perspective. Moreover, the learned layer-level weights are not adaptive, resulting in performance degradation for some frames. To resolve this issue, we propose an Adaptive Weight Aggregation (AWA) mechanism that assigns independent weights to each pixel of the correlation map, instead of sharing the same weight, to realise a fine-grained multi-layer feature fusion.

We apply different 1×1 convolutions and a Sigmoid function to each correlation map to generate the corresponding weight map. Once the pixel-level weight map is obtained, we use the weighted sum to fuse all the correlation maps. Compared with the existing methods, the corresponding weight map can be adaptively generated according to the correlation map in each frame. The above process can be formulated as:

$$\mathcal{S}_{all} = \sum_{i=1}^{3} \Omega_i * \mathcal{S}_i = \sum_{i=1}^{3} \sigma(\varphi_i(\mathcal{S}_i)) * \mathcal{S}_i \tag{6}$$

where Ω_i, φ_i and σ are the i-th weight map, 1×1 convolution layer, and the Sigmoid function. In addition, for the regression branch, we also visualise the weight maps Ω_i obtained by our proposed method for better comparison. As shown in Fig. 4, finer-grained weight maps can be obtained to adapt to the changes of tracking objects and scenarios. The same conclusion can be drawn that large dilation features contribute more, but their relative weight can be adjusted more flexibly.

3.2.2. *Bounding Box Refinement*

Siamese trackers usually use a single regression head with standard convolutional layers to obtain bounding boxes. However, to cope with complex tracking scenarios, the size of the region, subject to convolution fixed a priori during offline training, cannot always guarantee accurate bounding box regression. Therefore, we introduce a bounding box refinement module for performance boosting.

Different from previous methods [17, 30], our method does not rely on RoIPooling [28], but directly predicts the offset and distance scaling factors to obtain the refined bounding box, as shown in Fig. 2, which saves the cost and improves the accuracy of features. Specifically, for each position in the regression map, we first predict an initial bounding box $B'(l', t', r', b')$. With the regression map and predicted initial bounding box, we can compute the offset and obtain the aligned features using deformable convolution layer [24]. Then we predict four distance scaling factors $\Delta B(\Delta l, \Delta t, \Delta r, \Delta b)$ using the aligned features. By applying them to the initially predicted distance vectors, we can generate a refined bounding box as:

$$(l, t, r, b) = (\Delta l \times l', \Delta t \times t', \Delta r \times r', \Delta b \times b'). \tag{7}$$

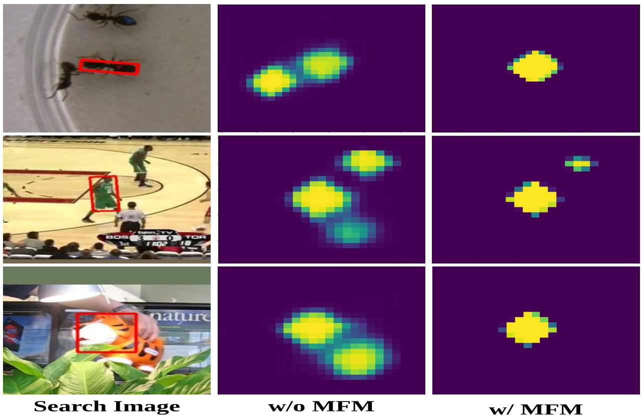


Fig. 5. Visualisation results of the classification maps without (w/o) and with(w/) our proposed MFM.

We use the regression loss and bounding box refinement loss as the supervision signals for the initial and refined bounding boxes, respectively, to avoid the interference of low-quality initial boxes. By assigning a larger weight to the latter one, the network training pays more attention to the final refined results, as detailed in Sec. 3.4. With the refinement module, the proposed method perceives and reacts to the target variation, thus predicting the bounding boxes more accurately.

3.3. *The Max Filtering Module*

In general, object tracking task requires a high-confidence classification ability to distinguish the target from distracting objects, while existing anchor-free trackers are not fully empowered to discriminate between similar predictions. Since the regression branch has the ability to guide object localisation quality estimation in general object detection, as demonstrated in [36], we exploit this potential ability to improve classification performance in tracking. In detail, a Max Filtering Module (MFM) is introduced to enable directional interaction and guidance of the classification branch by the regression branch.

MFM is applied to the final regression feature map, as shown in Fig. 2. We first apply 3×3 convolution and 2D max-pooling layers with the kernel size 3. The specific features are then added with the input final regression map and fed into the group normalisation and activation function. In this way, the maximum activation of the adjacent region is selected to filter out the interference of similar objects. A single channel 3x3 convolution is further applied to obtain the activation map. Finally, the activation map is multiplied with the original classification map to instil guidance from regression to classification.

$$\mathbf{F}_{cls}^{'} = \sigma(\varphi_2^{'}(\delta(MaxPool(\varphi_1^{'}(\mathbf{F}_r)) + \mathbf{F}_r))) \times \sigma(\mathbf{F}_{cls}) \qquad (8)$$

where $\mathbf{F}_r$ is the obtained feature map before bounding box prediction in the regression branch, $\mathbf{F}_{cls}$ and $\mathbf{F}_{cls}^{'}$ are the original classification map and final classification map obtained by MFM, respectively. $\varphi_1^{'}$, $\varphi_2^{'}$ and MaxPool denote the two different 3×3 convolution layers and the max pooling layer. δ and σ denote the GN+ReLU and Sigmoid functions. As shown in Fig. 5, with our MFM, the response of distractors in the classification map is suppressed, and the confidence of the target is higher. Please note that MFM only consists of simple differential operators with low computational overhead, affecting the tracking speed of our method in a minimal way, as described in Sec. 4.1.

Table 1. Ablation study on GOT-10k. The reported speed is evaluated on the same server with two RTX 2080ti GPUs.

	TCAB	ATM	MFM	$SR_{0.50}$	$SR_{0.75}$	AO	FPS	FLOPs
GOT-10k				0.694	0.465	0.590	72	23.6G
	✓			0.715	0.493	0.615	70	23.8G
	✓	✓		0.726	0.497	0.621	68	24.2G
	✓	✓	✓	0.743	0.515	0.635	66	24.5G

3.4. *Loss Functions*

We use a multi-task loss for the end-to-end training of our network:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{reg} + \lambda_2 \mathcal{L}_{refine}, \qquad (9)$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_{reg}$ and $\mathcal{L}_{refine}$ are the classification, regression and bounding box refinement loss functions. We set the balancing parameters λ_1 and λ_2 to 1.2 and 1.5 in this paper.

We use the Binary Cross-Entropy (BCE) loss [38], $\mathcal{L}_{cls} = -\sum_{i=1} p^* log(p) + (1 - p^*)log(1 - p)$, for classification, where i is the index of a training sample, p and p^* are the predicted classification score maps and ground truth labels.

The IoU loss [39], $\mathcal{L}_{reg} = -\sum_{i=1} ln(IoU(B_{reg}, B^*))$, is used for regression, where B_{reg} and B^* are the predicted initial bounding boxes and ground truth.

Last, the bounding box refinement loss with a higher weight is introduced to enhance the network training further, $\mathcal{L}_{refine} = -\sum_{i=1} ln(IoU(B_{refine}, B^*))$, where B_{refine} represents the refined bounding boxes.

4. Experimental Results

We use the pretrained ResNet50 [31] as our backbone. We then train our network on COCO [40], ImageNet DET [41], ImageNet VID [41], YouTube-BoundingBoxes [42] and GOT-10K [22] using the stochastic gradient descent optimiser [43], with the batch size of 64. The momentum and weight decay are set to 0.9 and 10^{-4}. For a fair comparison, we follow SiamRPN++ [11] and set the size of template and search window to 127×127 and 255×255. We train our network for 20 epochs with 6×10^5 training pairs per epoch. The learning rate warms up from 10^{-3} to 5×10^{-3} for the first 5 epochs, and then exponentially decays to 5×10^{-4}. The backbone is frozen for the first 10 epochs and fine-tuned with a learning rate 10 times smaller than the other parts of the proposed tracker.

We use the same offline tracking protocol as in Ocean [14] for testing. For every tracking sequence, we crop the target object in the first frame as the template and calculate its feature only once. In the subsequent frames, we crop the search patch based on the target location in the previous frame. For a more accurate and smoother prediction between adjacent frames, the size and cosine window penalty is applied to promote a smooth transformation of the target size and position. The proposed method is implemented using PyTorch on a server with 2 RTX 2080Ti GPUs and an Intel Core i9-9900X CPU.

Table 2. A comparison with the state-of-the-art methods on VOT2019 real-time benchmark. The best three results are highlighted in red, green and blue, the same below.

	SA_SIAM [21]	SPM [30]	SiamRPN++ [11]	SiamMask [32]	MAML [44]	DiMP [45]	SiamBAN [13]	Ocean-offline [14]	**Ours SiamTC**
EAO ↑	0.252	0.275	0.285	0.287	0.313	0.321	**0.327**	**0.327**	**0.345**
Accuracy ↑	0.563	0.577	**0.599**	**0.594**	0.570	0.582	**0.602**	0.590	**0.594**
Robustness ↓	0.507	0.507	0.482	0.461	**0.366**	**0.371**	0.396	0.376	**0.371**

Table 3. A comparison with the state-of-the-art methods on GOT-10k.

	SiamFC [8]	SPM [30]	SiamRPN++ [11]	ATOM [46]	SiamCAR [18]	Ocean-offline [14]	DiMP-50 [45]	Ocean-online [14]	**Ours SiamTC**
AO ↑	0.348	0.513	0.517	0.556	0.579	0.592	**0.611**	**0.611**	**0.635**
$SR_{0.50}$ ↑	0.353	0.593	0.615	0.634	0.677	0.695	**0.717**	**0.721**	**0.743**
$SR_{0.75}$ ↑	0.098	0.359	0.329	0.402	0.437	**0.473**	**0.492**	**0.473**	**0.515**

Table 4. The ablation experiment of Accurate Tracking Mechanisms on VOT2019. AWA and BBR indicate the proposed Adaptive Weight Aggregation and Bounding Box Refinement.

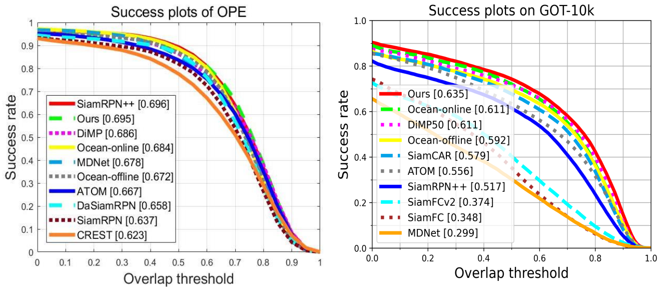
	AWA	BBR	Accuracy	Robustness	EAO
VOT2019			0.590	0.356	0.320
	✓		0.605	0.401	0.328
	✓	✓	0.601	0.356	0.334

4.1. *Ablation Study*

We first perform a detailed ablation study on GOT-10k to analyse the influence of each component of the proposed method. As mentioned in Sec. 3, our baseline method is the no object-aware version of Ocean-offline [14]. In Table 1, the AO score and speed of the baseline method are 0.590 and 72 FPS. The proposed Target-Cognisant Attention Block (TCAB) improves the AO score by 2.5%, demonstrating the importance of close communication between the template and search branches. By using the proposed Accurate Tracking Mechanisms (ATM), we can further improve the AO score to 0.621, which shows that ATM can predict more precise bounding boxes. The Max Filtering Module (MFM) also brings a considerable improvement due to its potential to filter out similar distractors, reaching the final performance boosting of 4.5%. This proves that the proposed innovations are complementary. In addition, the tracking speed is minimally affected, achieving a trade-off between accuracy and efficiency.

We also evaluated the proposed method in terms of FLOPs. As reported in Table 1, the FLOPs of the baseline method is 23.6G. Each component of the proposed method increases only 0.2-0.4G FLOPs. Therefore, our SiamTC has comparable computational complexity as compared with Ocean-offline [14] (24.5G vs 24.0G) while SiamTC achieves better results on four datasets, as presented in Sec. 4.2. Compared with SiamRPN++ [11](48.9G FLOPs), SiamTC greatly decreases the computational cost and improves the tracking performance.

To further demonstrate the superiority of TCAB, we visualise the target-cognisant attention in Fig. 3(b). We can see that the output feature of TCAB focuses more accurately on the tracked objects and filters out other distractors as well as background information, improving the tracker accuracy and robustness to appearance variations. In order to analyse the impact of each component of the proposed ATM, we further conducted a more detailed ablation experiment on VOT2019. As shown in Table 4, each brings 2.5% and 1.8% relative gain in EAO.



(a) OTB100

(b) GOT-10k

Fig. 6. Success plots on (a) OTB100 and (b) GOT-10k.

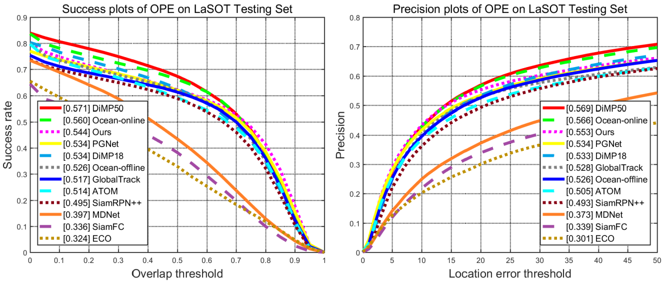


Fig. 7. Success and precision plots on LaSOT.

4.2. *Comparison with the State-of-the-art Approaches*

We compare our method with the state-of-the-art trackers, including SiamRPN++ [11], ATOM [46], DiMP [45], MAML [44], SiamBAN [13], SiamCAR [18], PGNet [47] and Ocean [14], etc., on four benchmarks: OTB100 [20], VOT2019 [21], GOT-10k [22] and LaSOT [23].

OTB100 has 100 videos with two evaluation metrics: precision score and success rate. A comparison with the state-of-the-art trackers is reported in Fig. 6(a) by using the one-pass evaluation (OPE). Our SiamTC achieves 0.695 in the success score, which is comparable to the anchor-based method (SiamRPN++ [11]) and outperforms the best anchor-free tracker (Ocean-offline [14]). It demonstrates that our method is effective and improves the tracking performance, thus bridging the gap between anchor-free and anchor-based trackers. Furthermore, SiamTC outperforms the online update methods (DiMP [45] and Ocean-online [14]), proving that our tracker has a good ability to distinguish the target from distracting objects.

VOT2019 is a very challenging benchmark that has 60 videos with rotated bounding boxes annotations. The performance measurement used here is the Expected Average Overlap (EAO). We evaluate our tracker on the VOT2019 real-time benchmark in Table 2. Compared with the anchor-base method SiamRPN++ [11], our tracker achieves a similar accuracy, with 23.0% lower failure rate and 21.1% higher EAO. In terms of the overall performance, our method surpasses the best two existing offline anchor-free trackers, SiamBAN [13] and Ocean-offline [14], achieving a relative gain of 5.5% in EAO. The improvement mainly comes from our carefully design of the proposed modules and accurate tracking mechanisms, which reduce the number of tracking failures.

LaSOT is a large-scale benchmark with 1,400 videos, with an average video length of 2,500+ frames. Success and precision in OPE are used as metrics. The results of ATOM [46], DiMP [45], PGNet [47] and Ocean [14] are provided by the original authors, while the rest results are provided by the official website of LaSOT. In Fig.7, our tracker achieves a success score of 0.544, outperforming SiamRPN++ [11] and ATOM [46] by a large margin. Compared with PGNet [47], our tracker improves the absolute scores of success and precision by 1.0% and 1.9%. Furthermore, it can be seen from the success plots that our tracker performs best when the overlap threshold exceeds 0.7, demonstrating that it can predict more accurate bounding boxes benefiting from accurate tracking mechanisms. The achievements of our SiamTC on such a large dataset illustrate its superiority in generalisation capability.

GOT-10k is a recently released database that offers unprecedentedly wide coverage of common moving objects in the wild. It has 10,000+ videos with more than 1.5 million high-precision bounding box annotations. Based on the protocol, the results of trackers must be evaluated on the official online server to avoid overfitting. The provided evaluation indicators include average overlap (AO), and success rate (SR) at overlap thresholds 0.5 and 0.75. As shown in Table 3, our tracker achieves an AO score of 0.635, outperforming all the state-of-the-art trackers by a large margin. In more detail, compared with the previous best method, it has improved the three indicators by 3.9%, 3.1%, and 4.7% respectively, setting a new SOTA record on this large dataset even without online training. Fig. 6(b) and Fig.1 provide the quantitative and qualitative comparison results, respectively, both of which demonstrate the superiority of the proposed tracker in predicting more accurate bounding boxes.

5. Conclusion

In this paper, we presented a novel Target-Cognisant anchor-free Siamese tracking framework, namely SiamTC, to achieve robust visual object tracking in unconstrained scenarios. We introduced the target-cognisant attention block and accurate tracking mechanisms to enhance the robustness and accuracy of the proposed tracker. Additionally, a max filtering module was developed to further improve the target discrimination capability. The comprehensive experimental results obtained on several benchmarking datasets validate the proposed tracking methodology, and demonstrate that the proposed network architecture is instrumental to the state-of-the-art performance achieved by our tracker.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (61876072, 61902153, 62106089, 62072243), Natural Science Foundation of Jiangsu Province (BK20221535) and the Fundamental Research Funds for the Central Universities (JUSRP122036).

References

[1] X. Lu, F. Tang, H. Huo, T. Fang, Learning channel-aware deep regression for object tracking, Pattern Recognition Letters 127 (2019) 103–109.

[2] K.-H. Lee, J.-N. Hwang, On-road pedestrian tracking across multiple driving recorders, IEEE Transactions on Multimedia 17 (9) (2015) 1429–1438.

[3] X. Lin, C.-T. Li, V. Sanchez, C. Maple, On the detection-to-track association for online multi-object tracking, Pattern Recognition Letters 146 (2021) 200–207.

[4] T. Xu, Z. Feng, X.-J. Wu, J. Kittler, Adaptive channel selection for robust visual object tracking with discriminative correlation filters, International Journal of Computer Vision 129 (5) (2021) 1359–1375.

[5] M. Danelljan, G. Bhat, F. Shahbaz Khan, M. Felsberg, ECO: Efficient convolution operators for tracking, in: CVPR, 2017, pp. 6931–6939.

[6] T. Xu, Z.-H. Feng, X.-J. Wu, J. Kittler, Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking, IEEE Transactions on Image Processing 28 (11) (2019) 5596–5609.

[7] R. Tao, E. Gavves, A. W. Smeulders, Siamese instance search for tracking, in: CVPR, 2016, pp. 1420–1429.

[8] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, P. H. Torr, Fully-convolutional siamese networks for object tracking, in: ECCV, 2016, pp. 850–865.

[9] L. Zheng, S. Zhao, Y. Zhang, L. Yu, Thermal infrared pedestrian tracking using joint siamese network and exemplar prediction model, Pattern Recognition Letters 140 (2020) 66–72.

[10] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High performance visual tracking with siamese region proposal network, in: CVPR, 2018, pp. 8971–8980.

[11] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, J. Yan, SiamRPN++: Evolution of siamese visual tracking with very deep networks, in: CVPR, 2019, pp. 4282–4291.

[12] Z. Zhang, H. Peng, Deeper and wider siamese networks for real-time visual tracking, in: CVPR, 2019, pp. 4591–4600.

[13] Z. Chen, B. Zhong, G. Li, S. Zhang, R. Ji, Siamese box adaptive network for visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6668–6677.

[14] Z. Zhang, H. Peng, J. Fu, B. Li, W. Hu, Ocean: Object-aware anchor-free tracking, in: European Conference on Computer Vision (ECCV), 2020.

[15] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, S. Wang, Learning dynamic siamese network for visual object tracking, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1763–1771.

[16] T. Yang, A. B. Chan, Learning dynamic memory networks for object tracking, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 152–167.

[17] Y. Yu, Y. Xiong, W. Huang, M. R. Scott, Deformable siamese attention networks for visual object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6728–6737.

[18] D. Guo, J. Wang, Y. Cui, Z. Wang, S. Chen, Siamcar: Siamese fully convolutional classification and regression for visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6269–6277.

[19] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, H. Lu, Transformer tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8126–8135.

[20] Y. Wu, Lim, et al., Object tracking benchmark, TPAMI 37 (9) (2015) 1834–1848.

[21] M. Kristan, Matas, et al., The seventh visual object tracking vot2019 challenge results, in: ICCVW, 2019, pp. 1–36.

[22] L. Huang, X. Zhao, K. Huang, Got-10k: A large high-diversity benchmark for generic object tracking in the wild, TPAMI 1 (1) (2019) 1–17.

[23] H. Fan, L. Lin, F. Yang, et al., LaSOT: A high-quality benchmark for large-scale single object tracking, in: CVPR, 2019, pp. 5374–5383.

[24] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 764–773.

[25] T. Xu, Z.-H. Feng, X.-J. Wu, J. Kittler, Learning low-rank and sparse discriminative correlation filters for coarse-to-fine visual object tracking, IEEE Transactions on Circuits and Systems for Video Technology 30 (10) (2019) 3727–3739.

[26] T. Xu, Z.-H. Feng, X.-J. Wu, J. Kittler, An accelerated correlation filter tracker, Pattern Recognition 102 (2020) 107172.

[27] A. He, C. Luo, X. Tian, W. Zeng, A twofold siamese network for real-time object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4834–4843.

[28] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, arXiv preprint arXiv:1506.01497 (2015).

[29] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, W. Hu, Distractor-aware siamese networks for visual object tracking, in: ECCV, 2018, pp. 101–117.

[30] G. Wang, et al., SPM-tracker: Series-parallel matching for real-time visual object tracking, in: CVPR, 2019, pp. 3643–3652.

[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[32] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, P. H. Torr, Fast online object tracking and segmentation: A unifying approach, in: CVPR, 2019, pp. 1328–1338.

[33] Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9627–9636.

[34] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, J. Yang, Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection, in: NeurIPS, 2020.

[35] H. Zhang, Y. Wang, F. Dayoub, N. Sünderhauf, Varifocalnet: An iou-aware dense object detector, in: CVPR, 2021.

[36] X. Li, W. Wang, X. Hu, J. Li, J. Tang, J. Yang, Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection, arXiv preprint arXiv:2011.12885 (2020).

[37] J. Wang, L. Song, Z. Li, H. Sun, J. Sun, N. Zheng, End-to-end object detection with fully convolutional network, arXiv preprint arXiv:2012.03544 (2020).

[38] P.-T. De Boer, D. P. Kroese, S. Mannor, R. Y. Rubinstein, A tutorial on the cross-entropy method, Annals of operations research 134 (1) (2005) 19–67.

[39] J. Yu, Y. Jiang, Z. Wang, Z. Cao, T. Huang, Unitbox: An advanced object detection network, in: ACM MM, 2016, pp. 516–520.

[40] T.-Y. Lin, M. Maire, S. Belongie, et al., Microsoft COCO: Common objects in context, in: ECCV, 2014, pp. 740–755.

[41] O. Russakovsky, Deng, et al., ImageNet large scale visual recognition challenge, IJCV 115 (3) (2015) 211–252.

[42] E. Real, J. Shlens, et al., Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video, in: CVPR, 2017, pp. 7464–7473.

[43] Y. LeCun, B. Boser, et al., Backpropagation applied to handwritten zip code recognition, Neural computation 1 (4) (1989) 541–551.

[44] G. Wang, C. Luo, X. Sun, Z. Xiong, W. Zeng, Tracking by instance detection: A meta-learning approach, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6288–6297.

[45] G. Bhat, M. Danelljan, L. V. Gool, R. Timofte, Learning discriminative model prediction for tracking, in: ICCV, 2019, pp. 6182–6191.

[46] M. Danelljan, G. Bhat, F. S. Khan, M. Felsberg, Atom: Accurate tracking by overlap maximization, in: CVPR, 2019, pp. 4660–4669.

[47] B. Liao, C. Wang, Y. Wang, Y. Wang, J. Yin, Pg-net: Pixel to global matching network for visual tracking, in: European Conference on Computer Vision, Springer, 2020, pp. 429–444.