CROSS-ATTENTION GUIDED NETWORK FOR VISUAL TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Most traditional Siamese trackers are used to process the feature information of the template branch and search branch respectively, and to correlate with each other to get the location of the max response which regarded as the target location. However, it is difficult for these traditional methods to reduce the influence of the similar object and background clutters because of ignoring the information between target template and search frame. Accordingly, a novel cross-attention guided network (SiamCAN), which involving cross-channel attention and self-spatial attention to learn discriminative object representations, is proposed to solve the above problem in this paper. Here, the channel attention of target template is introduced to guide the feature learning for search branch, and then the self-spatial attention is used to localize the informative part location after the correlation processing. Moreover, in order to get more accurate target estimation, an anchor-free mechanism and a distance-IoU (DIoU) loss are applied to minimize the distance between the center points of the ground-truth bounding boxes and predicted boundary boxes in the model training process. The proposed method achieves state-of-the-art performance on four visual tracking benchmarks including UAV123, OTB100, VOT2018 and VOT2019, outperforming the strong baseline, SiamRPN++, by $0.417 \rightarrow 0.445$ and $0.292 \rightarrow 0.323$ EAO on VOT2018 and VOT2019.

1 INTRODUCTION

Visual object tracking is the fundamental task of computer vision, aiming to track unknown object of which the information is given by the first frame. Although great progress has been achieved in recent years, a robust tracker is still in desperate demand due to tricky challenge such as scale variation, appearance deformation and similar object with complex background which can deteriorate tracking performance (Wu et al. (2013); Zhang et al. (2014)).

Recently, Siamese Network based trackers have taken a vital place in SOT field due to its accuracy and speed. Since (Tao et al. (2016)) and (Bertinetto et al. (2016)) introduced Siamese networks in visual tracking, Siamese structure has been adopted as baseline for researchers to design efficient trackers (Li et al. (2018); Zhu et al. (2018a); Zhang & Peng (2019); Li et al. (2019); Xu et al. (2020); Chen et al. (2020)). Moreover, region proposal network has become an essential part of tracker after SiamRPN (Li et al. (2018)) being proposed to obscure more accurate anchor boxes. However, the anchors these trackers use are manual-set which goes against the fact that the tracking target is unknown. Moreover, the design of anchors is not only tedious but also determines a lot on the success of tracking. Besides, the performance of the Siamese based trackers depends greatly on offline training by using massive frame pairs. Therefore, it highly increases the risk of tracking drift when facing significant deformations, similar object distractors, or complex background, due to the non-discriminative feature abstracted from the target when the category of the target is excluded from the training dataset.

In these years, the attention mechanism has become the spotlight in computer vision which inspires the relative works not only in detection task but also in visual tracking (He et al. (2018); Abdelpakey et al. (2018); Wang et al. (2018); Zhu et al. (2018b)). The attention mechanism includes channel attention and spatial attention, the former tends to generate a set of channel-weights for modeling interdependencies between channels while the latter focuses on finding the informative part by uti-

lizing the inter-spatial relationship of features. Considering these benefits, Siamese based trackers try to introduce attention module to distinguish target from complex background. Nevertheless, the performance of these trackers is not satisfactory for just calculating the attention of target template and the search frame separately not exploiting the deep relationship between each other.

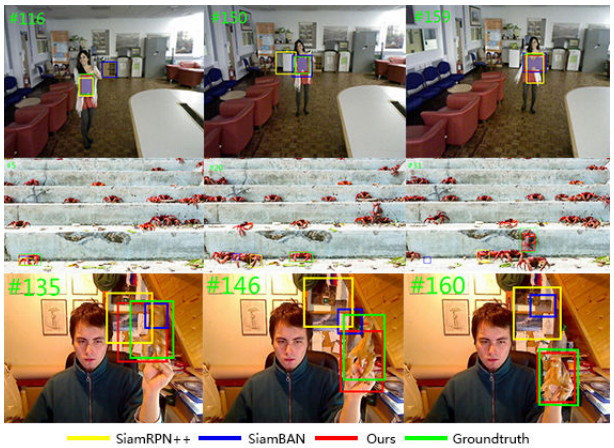

Figure 1: Tracking results on the challenges of three aspects: scale variation, similar distractors and fast motion. Compared with other two state-of-the-art trackers, our tracker (SiamCAN) performs better.

Based on the limitations discussed above, we design a Cross-attention Guided Siamese network (SiamCAN) based tracker with anchor-free strategy which performs better than the state-of-the-art trackers when facing the similar object challenge. Different from (He et al. (2018); Wang et al. (2018)), which use channel-wise attention feature or spatial-wise attention to have self-attention processing, SiamCAN takes template channel attention to guide the feature extraction of search image by which can strengthen the ability of tracker to overcome distractors and complex backgrounds, performing better than most of Siamese-based trackers, as shown in Figure 1. The main contributions of this work are:

- We formulate a novel cross-attention guided Siamese framework (SiamCAN) including cross-channel attention and self-spatial attention. The cross-channel attention builds an interactive bridge between the target template and search frame to share the identical channel weights. The self-spatial attention focuses on the discriminative part of the correlated feature map, which is complementary to the cross-channel attention.
- The proposed tracker is anchor-free, without numerous hyper-parameters setting. In order to get more accurate bounding box, we use DIoU loss to minimize the distance between the predicted bounding box and the ground-truth boundary box.
- The proposed SiamCAN achieves state-of-the-art results on four large tracking benchmarks, including OTB100 (Wu et al. (2013)), UAV123 (Mueller et al. (2016)), VOT2018 (Kristan et al. (2018)) and VOT2019 (Kristan et al. (2019)). The speed of tracker can also achieve 35 FPS.

2 Related Work

In this section, we briefly review the recent Siamese based trackers, the anchor-free approaches and attention mechanism in both tracking and detection filed.

2.1 Siamese network based tracker

The pioneering works, SINT (Tao et al. (2016)) and SiamFC (Bertinetto et al. (2016)), first introduce the siamese network in tracking filed. Due to its fast speed with light structure, Siamese network draws great attention from the visual tracking community. SiamFC tries to use siamese network to

learn the feature of both target template and search frame, and compare the similarity of them to find the most confident candidates. Although tracks fast, it cannot handle the scale variation problem by applying several scales of feature map. Inspired by Faster-RCNN (Ren et al. (2015)) from object detection, SiamRPN (Li et al. (2018)) draws on the region proposal network to get more various scale ratio bounding boxes. Since then, the RPN module has become an essential part of the tracker (Zhu et al. (2018a); Zhang & Peng (2019); Li et al. (2019); Dong & Shen (2018); Fan & Ling (2019)). However, the complexity of anchor design makes the performance of trackers depends greatly on the effect of anchor training.

2.2 Anchor-free Approaches

In recent time, Anchor-free approaches have developed fast. Different from anchor-based methods, anchor-free approach treats each pixel as an anchor which can directly predict the location of object. The achor-free work can be divided into two categories. The first one (Kong et al. (2019); Law & Deng (2018)) aims to estimate the keypoints of the objects. The other (Redmon et al. (2016); Tian et al. (2019)) predicts the bounding box for each pixel. Our anchor-free regression is inspired by the second category, but not the same with (Tian et al. (2019)) for the label design.

2.3 Attention Mechanism

Attention mechanism has been the focus of much attention in detection filed, on account to its powerful ability of enhancing deep CNNs. SE-Net (Hu et al. (2018)) firstly puts forward the mechanism to generate channel weights in return to direct the learning of channel attention. After that, CBAM (Woo et al. (2018)) utilizes both max-pooling and average-pooling to generate the merged attention, includes channel and spatial attention. Recently, ECA-Net (Wang et al. (2020b)) finds that avoiding dimensionality is of great importance for channel attention learning, and propose a cross-channel interaction strategy which performs better than SE-Net. In the tracking field, the recent trackers began to equip with the attention mechanism to get better performance. SA_Siam (He et al. (2018)) simply combines the SE-Net and SiamFC to get both discriminative and general features which boost the tracking performance. DensSiam (Abdelpakey et al. (2018)) uses attention mechanism to let tracker focus on the non-local feature by which can get robust appearance information. RASNet (Wang et al. (2018)) designs residual attention, general attention and channel attention to learn target feature better. However, these trackers only calculate alongside each branch and neglect the information flow between them, as a result, the ability of attention mechanism can not be fully utilized.

3 Our approach

As shown in Figure 2, the proposed framework mainly consists of two components, the one is the feature-extracted Siamese network, the other is the cross-attention network with anchor-free bounding box regression network and foreground classification network.

3.1 Feature extracted Siamese Network

Like the most Siamese based tracker, we adopt the fully convolution network without padding, which can guarantee the accurate location calculation. Feature extracted network composes of two parts, template branch and search branch. Both of them share the same parameters in backbone, by this mean, CNNs can abstract the familiar feature for them to calculate similarity in the subsequent operation. The template branch intends to encode the target template feature in the first frame while the other branch aims to encode the candidates feature which may involve target in the follow-up frames. Set the input in template branch as I_t, the subsequent frames in search branch as I_s. We feed the I_t and I_s into backbone, and get the feature $\phi_i(I_t)$ and $\phi_i(I_s)$ from different i-th backbone layers. Next, the given features are sent to the corresponding branch after having a convolution with a neck layer to reduce feature channel size to 256. Then, we use $\phi(I_t)$ to generate a set of channel weights vector $V(\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_n)$ to direct search branch channel learning, at meantime, applying a convolution layer of which kernel size 3×3 and stride 1 to get $\psi_t^{cls}(I_t)$ and $\psi_s^{cls}(I_s)$ for classification subnetwork, $\psi_t^{reg}(I_t)$ and $\psi_s^{reg}(I_s)$ for the regression subnetwork.

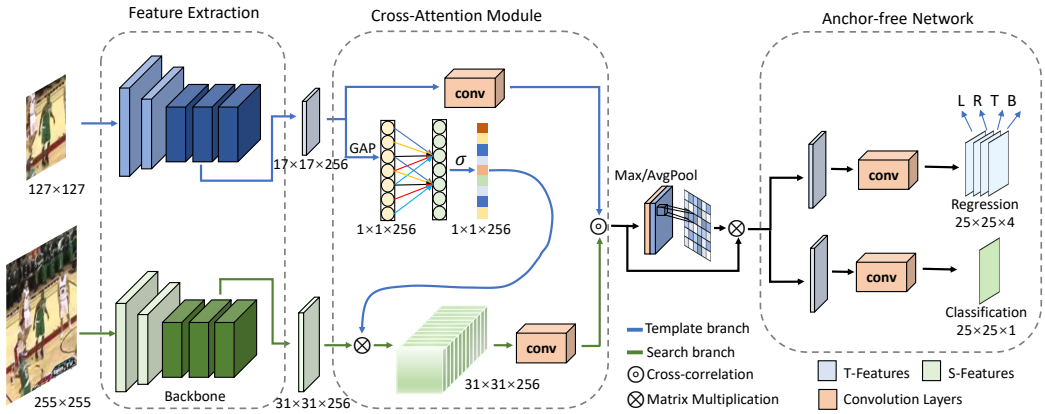

Figure 2: Illusion the proposed tracking framework, consisting of feature extraction, cross-attention module and anchor-free network. In the feature extraction part, the features of the third, fourth and fifth blocks are sent to the cross-attention module. The cross-attention module uses the channel attention from the template branch to combine with the feature from the search branch. The right side shows the classification and anchor-free regression network, which is taken to localize the target and predict the size of bounding box.

3.2 Cross-attention Network

Attention mechanism is created for the purpose that enforce CNNs to focus on the special parts which is of great importance, i.e., channels information and spatial information. The channel attention is designed to explore the interdependencies between channels by which CNNs distribute various weights to each channel according to the its materiality. The spatial attention tends to make CNNs pay more attention to the most critical areas of the feature, so that the foreground object can be distinguished from the background. Inspired by (Wang et al. (2020b)), we add channel attention and spatial attention into our network. Different from (He et al. (2018); Wang et al. (2018)), the channel attention is used between two branches not being applied as self-attention. In this paper, the target branch feature $\psi_t(I_t)$ is sent to global average pooling to get aggregated feature Y_t, i.e.,

$$Y_t = \frac{1}{WH}\sum_{i=0}^{W,H}\psi_t(I_t) \tag{1}$$

Given the aggregated feature, the channel weight is obtained by performing a 1D convolution of size k, i.e.,

$$V_i = \sigma(\sum_{j=1}^{k}\omega^j y_i^j), y_i^j \in \Omega_i^k \tag{2}$$

Where σ is a Sigmoid function, ω indicates the parameters of 1D convolution and Ω_i^k indicates the set of k adjacent channels of y_i. To let search branch learns the information from target template, we multiply the channel weights with the search feature, i.e.,

$$\widetilde{\psi_s}(I_s) = \psi_s(I_s) * V \tag{3}$$

3.3 Classification and anchor-free regression subnetwork

As shown in Figure 2, the correlation feature map is calculated by the depth-wise correlation operation between $\widetilde{\psi_s}(I_s)$ and $\psi_t(I_t)$, i.e.,

$$F_{w\times h\times c}^{cls} = \widetilde{\psi_s}(I_s) \star \psi_t(I_t) \tag{4}$$

$$F_{w\times h\times c}^{reg} = \widetilde{\psi_s}(I_s) \star \psi_t(I_t) \tag{5}$$

where $\star$ denotes depth-wise convolution operation. Then, we apply self-spatial attention to the feature map in order to focus on discriminative part automatically. After that, we use two convolution layers with kernel size 1×1 to reduce the number of channel from 256 to 2 and 4 respectively for

each branch and concatenate the feature maps from different layers of backbone by the trainable weights α, i.e.,

$$P^{cls}_{w\times h\times 2} = \sum_{i=1}^{N} \alpha_i * F^{cls}_{w\times h\times 2}(i) \tag{6}$$

$$P^{reg}_{w\times h\times 4} = \sum_{i=1}^{N} \alpha_i * F^{reg}_{w\times h\times 4}(i) \tag{7}$$

where N denotes the total number of the backbone layers we use. The classification feature map has two channels, the one represents the foreground and the points (i, j) on $\boldsymbol{P}^{cls}_{w\times h\times 2}(0, i, j)$ refers to the probability scores of target, the other represents the background and the points (i, j) on $\boldsymbol{P}^{cls}_{w\times h\times 2}(1, i, j)$ refer to the probability scores of background. The regression feature map has four channels, each of them represents the four directions distances from the points location in search branch input to the four sides of the bounding box respectively, that is to say, each point (i, j) in $\boldsymbol{P}^{reg}_{w\times h\times 2}(:, i, j)$ is a vector which can denoted as (l, r, t, b).

Classification label and regression label. Different from (Tian et al. (2019)), we use ellipse figure region to design label for points (i, j) in feature map instead of using triangle figure area, which is inspired by (Chen et al. (2020)). The ellipse E_1 center and axes length are set by groundtruth center (g_{x_c}, g_{y_c}) of groundtruth size $(\frac{g_w}{2}, \frac{g_h}{2})$, We also can get the ellipse E_2 with center (g_{x_c}, g_{y_c}) and $(\frac{g_w}{4}, \frac{g_h}{4})$ as axes length, i.e.,

$$\frac{(B(p_i) - g_{x_c})^2}{(\frac{g_w}{2})^2} + \frac{(B(p_j) - g_{y_c})^2}{(\frac{g_h}{2})^2} = 1 \tag{8}$$

$$\frac{(B(p_i) - g_{x_c})^2}{(\frac{g_w}{4})^2} + \frac{(B(p_j) - g_{y_c})^2}{(\frac{g_h}{4})^2} = 1 \tag{9}$$

where B denotes the calculation for the location of points (i, j) in feature map $P^{cls}_{w\times h\times 2}$ back to the search frame. If the point $B(p_i, p_j)$ falls within the E_2 region, it will be given a positive label, and if it falls outside the E_1 area, it will be given a negative label, i.e.,

$$label = \begin{cases} 1, \ if E_2(p(i, j)) < 1 \\ 0, \ if E_1(p(i, j)) > 1 \\ -1, \ otherwise \end{cases} \tag{10}$$

For regression branch, the regression targets can be defined by:

$$d^l_{(i,j)} = p_i - g_{x_0}, d^t_{(i,j)} = p_j - g_{y_0} \tag{11}$$

$$d^r_{(i,j)} = g_{x_1} - p_i, d^b_{(i,j)} = g_{y_1} - p_j \tag{12}$$

where (g_{x_0}, g_{y_0}), (g_{x_1}, g_{y_1}) denote the left-top and right-bottom coordinates location of the groundtruth.

Loss function. We employ cross entropy loss to train the classification network. To predict more accurate bounding box, we adopt the DIoU loss (Zheng et al. (2020)) to train the regression network, i.e.,

$$L_{reg} = 1 - IoU + \frac{\rho^2(p, p^{gt})}{c} \tag{13}$$

where $\rho(.)$ is the Euclidean distance, p and p_{gt} denote the central points of predicted box and groundtruth and c is the diagonal length of the smallest enclosing box covering the two boxes. For regression branch training, DIoU loss can optimize the bounding faster than GIoU loss (Rezatofighi et al. (2019)). The overall loss function is:

$$L = \lambda_1 L_{cls} + \lambda_2 L_{reg} \tag{14}$$

where constants λ_1 and λ_2 wights the classification loss and regression loss. During model training, we simply set $\lambda_1 = 1$, $\lambda_2 = 1$ without hyper-parameters searching.

Classification label and regression label. We train our model by using image pairs, a 127×127 pixels template patch and a 255×255 pixels search patch. The training datasets include ImageNet VID (Russakovsky et al. (2015)), COCO (Lin et al. (2014)), YouTube-Bounding Boxes (Real et al. (2017)), ImageNet Det (Russakovsky et al. (2015)) and GOT10k (Huang et al. (2019)). Due to the numbers of negative samples are more than the positive samples, we set the at most 16 positive samples and 48 negative samples from the search image. Besides, in order to get more accurate regression information, we adopt the DIoU loss to optimize the regression branch.

3.4 TRAINING AND INFERENCE

Training. We train our model by using image pairs, a 127×127 pixels template patch and a 255×255 pixels search patch. The training datasets include ImageNet VID (Russakovsky et al. (2015)), COCO (Lin et al. (2014)), YouTube-Bounding Boxes (Real et al. (2017)), ImageNet Det (Real et al. (2017)) and GOT10k (Huang et al. (2019)). Due to the numbers of negative samples are more than the positive samples, we set the at most 16 positive samples and 48 negative samples from the search image. Besides, in order to get more accurate regression information, we adopt the DIoU loss to optimize the regression branch.

Inference. We feed the cropped first frame and the subsequent frames to the feature abstract network as template image and search image. Next, the features are sent to the cross-attention module and pass the classification branch and regression branch. After that, we get the classification map. The highest score location represents the most probable center location of the tracking target. After getting the highest score, we use the scale change penalty and the cosine window as that introduced in (Li et al. (2018)) to guarantee the smooth movements of target. According to the location p of the final score, we can get the predicted boxes B, i.e.,

$$b_{x1} = p_i - d_l^{reg}, b_{y1} = p_j - d_t^{reg} \tag{15}$$

$$b_{x2} = p_i + d_r^{reg}, b_{y2} = p_j + d_b^{reg} \tag{16}$$

where $d_{l,r,t,b}^{reg}$ denotes the predicted values of the regression targets on the regression map, (b_{x1}, b_{y1}) and (b_{x2}, b_{y2}) are the top-left corner and bottom-left corner of the predicted box.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

Our approach is implemented in Python with Pytorch on 1 RTX2080Ti. The backbone is modified ResNet-50 as in (He et al. (2016)), and the weights are pre-trained on ImageNet (Russakovsky et al. (2015)). During the training phase, the model is optimized by the stochastic gradient descent (SGD), at the meantime, the total number of epochs is 20 and the batch size is set as 28. For the first 10 epochs, we frozen the parameters of the backbone and only train the heads structures, for the last 10 epochs, we unfrozen the last 3 blocks of backbone to be trained together. Besides, we warm up the training during the first 5 epoch with a warmup learning rate of 0.001 to 0.005, and in the last 15 epochs, the learning rate exponentially decayed from 0.005 to 0.00005.

4.2 RESULTS ON THREE BENCHMARKS

To recify the effect of our method, we evaluate our tracker performance with the recent trackers MAML (Wang et al. (2020a)), PriDiMP (Danelljan et al. (2020)), SiamBAN (Chen et al. (2020)), SiamCAR (Guo et al. (2020)), ROAM (Yang et al. (2020)), SiamFC++ (Xu et al. (2020)), STN (Tripathi et al. (2019)), ARTCS (Kristan et al. (2019)), SiamRPN++ (Li et al. (2019)), SiamDW_ST(Zhang & Peng (2019)), ATOM (Danelljan et al. (2019)), DIMP-18 (Bhat et al. (2019)), DaSiamRPN (Zhu et al. (2018a)), SPM (Wang et al. (2019))ECO (Danelljan et al. (2017)), CFNet (Valmadre et al. (2017)), SiamRPN (Li et al. (2018)), DeepSRDCF (Danelljan et al. (2015a)), SRDCF (Danelljan et al. (2015b)) on four benchmarks UAV123 (Mueller et al. (2016)), VOT2018 (Kristan et al. (2018)), VOT2019 (Kristan et al. (2019)) and OTB100 (Wu et al. (2013))(details in Appendix A.1).

4.2.1 RESULTS ON UAV123

UAV123 contains 123 challenging video sequences, which can be divided into 9 categories according to their attributes. The performance of the tracker is evaluated by two evaluation metrics, the precision scores and the AUC scores. The AUC scores reflect the overlap between the predict bounding box and ground-truth box, while the precision scores are relative with the distance between the center of the predict bounding box and ground-truth box. In Figure 3, our method achieves the best performance in precision scores, i.e., 0.850, and the second best success scores 0.678. As for the 9 categories of the challenge, SiamCAN ranks 1st or 2nd in Scale Variation, Similar Object, Fast

Table 1: Comparison with state-of-the-art trackers on VOT2018. Red, blue and green fonts indicate the top-3 trackers.

	ECO	SiamRPN	MAML	ATOM	SiamFC++	SiamRPN++	PrDiMP	Ours
EAO↑	0.280	0.244	0.392	0.401	0.400	0.417	0.442	0.445
A↑	0.484	0.490	0.635	0.590	0.556	0.604	0.618	0.591
R↓	0.276	0.460	0.220	0.204	0.183	0.234	0.165	0.173

Table 2: Comparison with state-of-the-art trackers on VOT2019. Red, blue and green fonts indicate the top-3 trackers.

	SPM	ARTCS	SiamRPN++	MAML	SiamDW_ST	ATOM	STN	Ours
EAO↑	0.275	0.287	0.292	0.295	0.299	0.301	0.314	0.323
A↑	0.577	0.602	0.580	0.637	0.600	0.603	0.589	0.591
R↓	0.507	0.482	0.446	0.421	0.467	0.411	0.349	0.381

Motion and Low Resolution, in Appendix A.2. The results demonstrate that our trackers can handle the similar object and scale change challenge, due to the learning of cross-attention subnetwork and anchor-free mechanism.

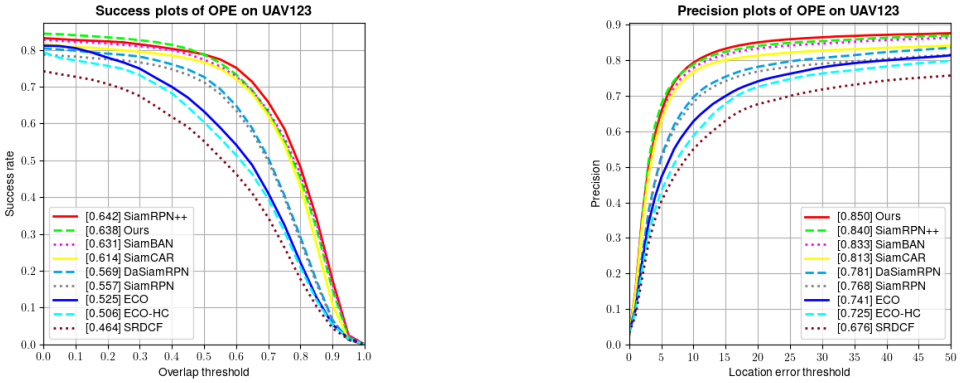

Figure 3: Success and precision plots on UAV123.

4.2.2 Results on VOT2018

VOT2018 consists of 60 challenging videos. The evaluation metrics to rank the performance of the tracker based on EAO (Expected Average Overlap) which depends on the accuracy and the robustness. As shown in Table 1, our tracker outperforms SiamRPN++ by 2.8 points by introducing the anchor-free regression network. Moreover, we achieve EAO improvements of 4.5 points compared with another anchor-free tracker, i.e., SiamFC++. Besides, the robustness scores of Siamese-based trackers are higher than our methods, which affect the EAO scores, in other words, SiamCAN is a robust Siamese-based tracker.

4.2.3 Results on VOT2019

VOT2019 video sequences are 20% different from VOT2018, adding more fast motion and similar object videos. Table 2 reflects the evaluation results on VOT2019 compared with the recent trackers. We can see that the recent proposed MAML gets the highest accuracy scores, while our SiamCAN surpassing MAML by 2.5 points in terms of EAO. Besides, our robustness scores also rank 2nd.

Table 3: Results of each Components of our tracker on VOT2018.

#Num	Components	Regression loss	EAO
①	baseline	GIoU Loss	0.351
②	baseline	DIoU Loss	0.378
③	+ cross-attention	DIoU Loss	0.421
④	+ spatial-attention	DIoU Loss	0.445

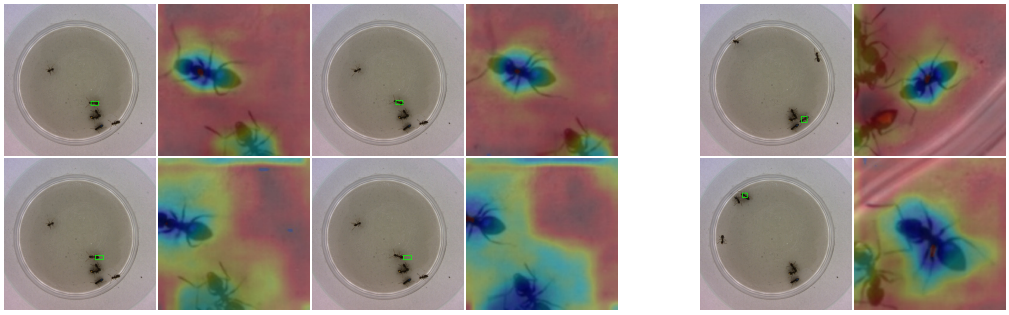

Figure 4: Visualization of confidence map. The 1st line: the left is the confidence map with cross-attention module, the right is not. The 2nd line: subsequent frames, and the 3rd line: the corresponding confidence map with cross-attention module.

4.3 Analysis of the Proposed Method

Discussion on components of tracker. To verify the role of each components of our tracker, we use the results of components on VOT2018 to analysis the details. As shown in Table 3, the baseline model can get an EAO of 0.351, which consists of a regular backbone, i.e., ResNet50, a classification and an anchor-free regression network. Exchanging DIoU Loss for GIoU Loss during training can get higher scores, due to the more accurate predicted bounding box (② vs ①). Adding cross-attention module can obtain a large improvement, i.e., 4.3 points on EAO (③ vs ②). This demonstrates the information interaction between the template branch and the search branch is of great significant. Finally, the tracker utilizes the self-spatial attention can gain another improvement of 2.4 points on EAO (④ vs ③).

Feature visualization. We visualize the features extracted from both the tracker equipped with cross-attention module and the tracker without cross-attention module, in Figure 4. On the left side of Figure 4, the first line and second line show the feature of the tracker equipped with cross-attention module and baseline which not has cross-attention module for two consecutive frames. What we can get from the picture is that the baseline is easier to lose target and focus on the wrong object when appear similar object, and the visualized feature demonstrates that cross-attention module can enable tracker tell target from similar distactors. On the right side of Figure 4, the pictures depict the performance of our tracker which shows the effectiveness of the cross-attention module.

5 Conclusion

In this paper, we propose a Siamese-based tracker called SiamCAN, which combined with cross-attention and anchor-free mechanism. The cross-attention module utilizes the target template channel attention to guide the feature learning of search frame, bridging the information flow between each branch. The anchor-free regression discards the fussy design of the anchors, and adjusts the scale ratio of the bounding box automatically. Additionally, using the proper loss function, i.e., DIou Loss, to maximize the anchor-free mechanism effect. Extensive experiments are conducted on four benchmarks, which demonstrates our trackers achieves the state-of-the-art performance, especially in scale variation, background cluster, deformations and similar distractors challenges.

REFERENCES

Mohamed H Abdelpakey, Mohamed S Shehata, and Mostafa M Mohamed. Denssiam: End-to-end densely-siamese network with self-attention model for object tracking. In *International Symposium on Visual Computing*, pp. 463–473. Springer, 2018.

Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pp. 850–865. Springer, 2016.

Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6182–6191, 2019.

Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6668–6677, 2020.

Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Convolutional features for correlation filter based visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 58–66, 2015a.

Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Learning spatially regularized correlation filters for visual tracking. In *Proceedings of the IEEE international conference on computer vision*, pp. 4310–4318, 2015b.

Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6638–6646, 2017.

Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4660–4669, 2019.

Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7183–7192, 2020.

Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 459–474, 2018.

Heng Fan and Haibin Ling. Siamese cascaded region proposal networks for real-time visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7952–7961, 2019.

Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. Siamcar: Siamese fully convolutional classification and regression for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6269–6277, 2020.

Anfeng He, Chong Luo, Xinmei Tian, and Wenjun Zeng. A twofold siamese network for real-time object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4834–4843, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks, 7132–7141. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT*, 2018.

Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, and Jianbo Shi. Foveabox: Beyond anchor-based object detector. *arXiv preprint arXiv:1904.03797*, 2019.

Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Cehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, et al. The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0, 2018.

Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka Cehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, et al. The seventh visual object tracking vot2019 challenge results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 734–750, 2018.

Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. 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.

Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4282–4291, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *European conference on computer vision*, pp. 445–461. Springer, 2016.

Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5296–5305, 2017.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 658–666, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Ran Tao, Efstratios Gavves, and Arnold WM Smeulders. Siamese instance search for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1420–1429, 2016.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 9627–9636, 2019.

Ardhendu Shekhar Tripathi, Martin Danelljan, Luc Van Gool, and Radu Timofte. Tracking the known and the unknown by leveraging semantic information. In *BMVC*, volume 2, pp. 6, 2019.

Jack Valmadre, Luca Bertinetto, Joao Henriques, Andrea Vedaldi, and Philip HS Torr. End-to-end representation learning for correlation filter based tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2805–2813, 2017.

Guangting Wang, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. Spm-tracker: Series-parallel matching for real-time visual object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3643–3652, 2019.

Guangting Wang, Chong Luo, Xiaoyan Sun, Zhiwei Xiong, and Wenjun Zeng. Tracking by instance detection: A meta-learning approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6288–6297, 2020a.

Qiang Wang, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, and Stephen Maybank. Learning attentions: residual attentional siamese network for high performance online visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4854–4863, 2018.

Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11534–11542, 2020b.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2411–2418, 2013.

Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *AAAI*, pp. 12549–12556, 2020.

Tianyu Yang, Pengfei Xu, Runbo Hu, Hua Chai, and Antoni B Chan. Roam: Recurrently optimizing tracking model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6718–6727, 2020.

Kaihua Zhang, Lei Zhang, and Ming-Hsuan Yang. Fast compressive tracking. *IEEE transactions on pattern analysis and machine intelligence*, 36(10):2002–2015, 2014.

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

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *AAAI*, pp. 12993–13000, 2020.

Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 101–117, 2018a.

Zheng Zhu, Wei Wu, Wei Zou, and Junjie Yan. End-to-end flow correlation tracking with spatial-temporal attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 548–557, 2018b.

A APPENDIX

A.1 EXPERIMENT RESULTS ON OTB100

OTB100 contains 100 challenging video sequences, and the evaluation metrics are same with UAV123. In Figure 5, our method achieves the best performance in precision scores, i.e., 0.913, and the second best success scores 0.684. As for the 9 categories of the challenge, SiamCAN ranks 1st or 2nd in Deformation, Background Clutters, Scale Varation and Out-of-Plane Rotation. The results demonstrate that our trackers can handle the deformation, background clutters, scale variation and out-of-plane rotation, in Figure 6.

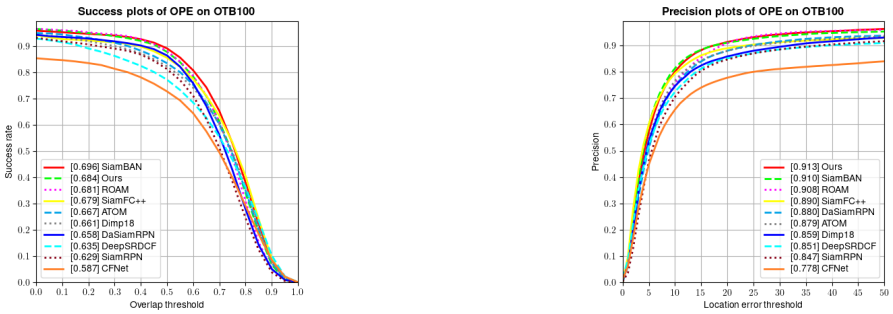

Figure 5: Success and precision plots on OTB100.

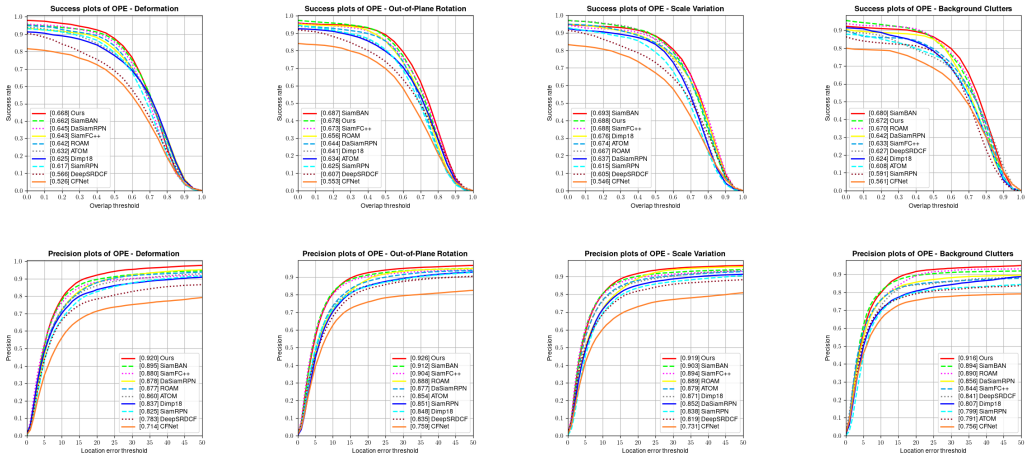

Figure 6: Comparisons on OTB-100 with challenges: Deformation, Background Clusters, Scale Variation and Out-of-Plane Rotation. Our SiamCAN performance ranks in top two.

A.2 More experiment results on UAV123

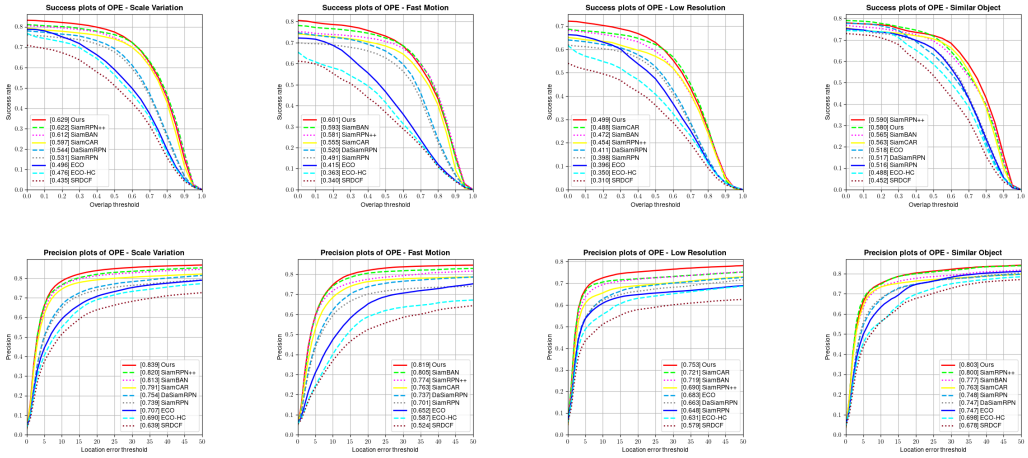

Figure 7: Comparisons on UAV123 with challenges: Scale Variation, Similar Object, Fast Motion and Low Resolution. Our SiamCAN performance ranks in top two.