Siamese Visual Tracking with Dual-Pipeline Correlated Fusion Network

Ying KANG††, Cong LIU†††, Ning WANG††††, Dianxi SHI†††††, Ning ZHOU††, Nonmembers, Mengmeng LI††††, Student Member, and Yunlong WU††††, Nonmember

SUMMARY  Siamese visual tracking, viewed as a problem of max-similarity matching to the target template, has absorbed increasing attention in computer vision. However, it is a challenge for current Siamese trackers that the demands of balance between accuracy in real-time tracking and robustness in long-time tracking are hard to meet. This work proposes a new Siamese based tracker with a dual-pipeline correlated fusion network (named as ADF-SiamRPN), which consists of one initial template for robust correlation, and the other transient template with the ability of adaptive feature optimal selection for accurate correlation. By the promotion from the learnable-correlation-response fusion network afterwards, we are in pursuit of the syntethical improvement of tracking performance. To compare the performance of ADF-SiamRPN with state-of-the-art trackers, we conduct lots of experiments on benchmarks like OTB100, UAV123, VOT2016, VOT2018, GOT-10k, LaSOT and TrackingNet. The experimental results of tracking demonstrate that ADF-SiamRPN outperforms all the compared trackers and achieves the best balance between accuracy and robustness.

key words: visual tracking, siamese network, dual-pipeline correlation, transient template, fusion network, adaptive selection

1. Introduction

Visual tracking, defined as an instance detection in the literature [1], is a compound task of class-agnostic classification and regression. With the designated target by a bounding box in the first frame, the goal of tracking is to ascertain the location and extent of the target in the following video sequences automatically. Influenced by the target movement and viewpoint change of camera, it is inevitable that the scenarios with occlusion, deformation, scale variation, out of view, background cluttering and illumination variation often occur as the video frames scroll. The key role for trackers is how to trace the target accurately and efficiently, as well as in a high discriminative manner between the target and non-target foreground like distractors or semantic background.

There are two main reasearch interests in visual tracking area: correlation filter (CF) based methods and deep learning based approaches. Due to the transformation in the Fourier domain, the CF based trackers [2], [3] have the ability of tracking target in real time and update the parameters of filters online. Meanwhile, benefited from the application of deep learning in the computer vision field, the performance of visual tracking has been promoted vastly. One striking model is the Siamese tracking network [4], which formulates the tracking process as a similarity comparison between the target template and the search region.

Currently, it is still a challenge for the Siamese based trackers that the demands of balance between accuracy in real-time tracking and robustness in long-time tracking are unable to meet. For the computational efficiency, the similarity function in early Siamese based trackers [5], [6] is learnt offline from training phase while remains fixed during reference phase, which is at the cost of accuracy. Although some trackers [7], [8] with fine-tuning strategies emerge, there exists a large risk of tracking drift owing to focusing on the variation of target appearance and ignoring the feature cluttered by semantic background or missed by occlusion.

Actually, the initial template is the most robust for the tracking process, but lack of the instantaneous capture of feature. If updating the template blindly frame by frame, it is easy to blend clutter knowledge or miss feature in reference phase, resulting in some terrible tracking loss. We are intended to make use of the respective advantages of initial template and transient varied feature for compensation. Therefore, remaining the innate cross-correlation pipeline with the initial template for robust tracking, we add another cross-correlation operation between the search region and a novel transient template, constructing the dual-pipeline cross correlation and then exploring a learnable fusion network to optimize the eventual tracking results by syncratic enhancement. To avoid the tracking drift with the maximum, we design an adaptive optimal selection scheme to generate the transient template, and adopt the alternative atrous convolution [9] to extend the receptive fields for robustness boosting. In summary, this is a double Y-shaped Siamese tracking network appended by a dual-pipeline correlated fusion subnetwork, which named as ADF-SiamRPN for short.

2. Related Works

Visual tracking has witnessed a rapid boost in the last decade, especially with the successful utilization of deep CNN in various computer vision tasks.
The CF based tracking methods prevailed in earlier years [10], [11], while remained popular for skill improvement. For example, MMCT [12] constructed multiple experts on CF by aggregating the diverse representations of target appearance. TRACa [13], built on multiple expert auto-encoders, realized the deep feature compression by a context-aware scheme for the high computational speed. ASRCF [14] was an efficient and robust CF-based tracker with the capacity of adaptive spatial regularization and efficient scale estimation. In effect, the most advantage of CF-based tracker is its high efficiency suitable for deployment on board.

Due to the powerful presentation of CNN, the deep-learning based tracking methods have attracted more and more attention, especially the Siamese tracking networks. SiamFC [4] is the first to use Siamese network for searching the maximum correlation between the target template and the search region to realize tracking. Afterwards, many improvement and extension works [5], [15] sprung up. Inspired by the region proposal network (RPN), the anchor-based Siamese networks for tracking were initially proposed by SiamRPN [6] and revised by the following methods DaSiamRPN [8] and SiamRPN++ [16] etc. C-RPN [17] designed a multi-stage Siamese tracking network, where multiple stages represented for cascading a sequence of RPNs from deep high-level layers to shallow low-level layers. SiamAttn [18] introduced a new Siamese attention mechanism of computing deformable self-attention and cross-attention jointly to improve the target representation. In addition, some works allowed for the limitation of presupposed bounding boxes and thought of applying the anchor-free skills in detection [19] to tracking. SiamBAN [20] took the anchor-free measure to address the problems of scale and aspect ratio estimation more flexibly in the tracking process.

3. Method

3.1 Overview

The key point for visual tracking is how to accurately grasp the change of target appearance with the video frames scrolling, while avoid the confusion from distractors or clutter from semantic background or information missing from occlusion. In accordance with the definition of tracking, we keep the robust tracking pipeline of correlation with the initial template firstly. Meanwhile, we introduce the other correlation with a novel transient template to seize the appearance variation for tracking accuracy. The transient template is originated from an adaptive optimal selection from the initial template, the last-frame transient template and the last-frame tracking result. Guided by the dual templates, a learnable response fusion network (named as FNet) is embedded for the sake of the associated enhancement for tracking results. The whole framework of our proposed algorithm ADF-SiamRPN is shown in Fig. 1, which contains two sub-networks of ATTONet and FNet on the right side. The depicted in the right-up corner of Fig. 1 is ATTONet, the output of which is the transient template at frame $T_t$. We will describe the realization of ADF-SiamRPN in detail.

3.2 Siamese Tracker with Double Y-Shaped Correlation

Firstly, let’s review the formulation of Siamese based tracker in work [4]. Given the target template $z$, the goal of visual tracking is to search the most similar patch from the search region $x$ in the following video sequence. The Siamese tracking network is composed of a pair of CNNs with sharing parameters $\phi$, which extract the respective features of the target template $z$ and the search region $x$ within a common embedding space. Then the most matching patch to $z$ can be measured by a similarity metric $f$:

$$f(x, z) = \phi(x) \ast \phi(z)$$  (1)

where $\ast$ denotes the operation of cross correlation in the feature space.

In general, the Siamese tracking network only contains one correlation operation between the template which is fixed at the first frame of video sequence or updated with the frame scrolling and the search region. Here we assume this correlation operation as a unified expression in Eq. (1). As mentioned above, the appearances of target like shape, size and brightness are bound to change. If the template is fixed as the initialized, when the target feature varies greatly, it will lead to tracking lost. But if updating the template timely with the video sequence, once being in the scenes like occlusion, blur or out of view, it is easy to blend miss or inferential information to incur the unexpected tracking drift. The same to the phenomena of cluttered background or distractors appearing. Therefore, why not to utilize the separate advantages of initial template and updated template to construct a better tracker by compensation?

We are intended to introduce a dual-pipeline correlation module into the Siamese tracking network, which is shown as a double Y-shaped neural network in the green boxes of Fig. 1. Specifically, keeping up the robust cross-correlation pipeline between the initial template $z_o$ and the search region $x$, we add another cross-correlation pipeline between a novel transient template $z_t$ and the search region $x$, for the aim of synchronically reinforcing the tracking performance afterwards. The transient template (referred to Sect. 3.4) is updated by an adaptive optimal selection schema (referred to Sect. 3.5), which aims at outputting the most accurate updated template for tracking. Similar to Eq. (1), we can obtain both cross-correlation responses from:

$$f_o(x, z_o) = \phi(x) \ast \phi(z_o)$$  \hspace{0.5cm} (2)

$$f_t(x, z_t) = \phi(x) \ast \phi(z_t)$$  \hspace{0.5cm} (3)

where $z_o$ and $z_t$ separately denote the initial template and the transient template. Actually, there is no need to extract the feature map for the transient template due to ATTONet.
3.3 Learnable Correlated Fusion Network

Built on the two outputs of dual-pipeline cross correlation (as formulated in Eq. (2) and Eq. (3)), we design a learnable correlated fusion network to generate a jointly enhanced tracking result abiding by two principles: 1) the eventually output tracking result mainly depends on the response to the correlation with the initial template to ensure robustness; 2) the output tracking result from the response to the correlation with the transient template plays a part of fine tuning to promote accuracy in the fusion process. Therefore, we innovate a residual network simulated to the update strategy in literature [21] to realize our fusion network (FNet), which is shown in the right-down corner of Fig. 1. Furthermore, we give the formulation of FNet as follows:

$$F(f_o, f_t) = f_o(x, z_o) + \psi(f_o(x, z_o) \parallel f_t(x, z_t))$$ (4)

where $\psi()$ represents a fusion function and $\parallel$ is an operation of concatenation. The obtained value of $F$ is a tracking result by associated enhancement from the fusion network, which is learnable based on the ground truth.

3.4 Adaptive Optimal Selection for Transient Template

The fundamental idea of dual-template setting is to advance tracking accuracy meanwhile reserve tracking robustness. With the previous analysis, the key problem is to design a template capable of seizing the transient variation of target. The simplest is the linear combination of various informations like target examplar or varying appearance, etc., but undoubtedly lack of dynamically nonlinear adaption ability. In order to gain the better stochastic characteristic, we innovate an adaptive optimal selection scheme inspired by work [22] to generate a transient template for accuracy promotion.

It is known that the up-to-date varied feature lies in the current frame. Firstly, take the initial template as a base template. When we obtain an accurate tracking result at frame $T_{i-1}$, we can make use of it to help the base template fine tune with the variation of target. On the contrary, when the tracking happens to drift, it will produce negative effect on the base template for tracking. Thus, we propose an adaptive transient template optimizing network (named as ATTONet) to guarantee the tracking results guided to the desired direction. To be formulated, we introduce an adaptive factor $\lambda$ to modulate the influence of the tracking result from last frame:

$$\lambda = \exp(-\alpha \mathcal{L}_{ce}(R_{T_i-1}, Y))$$ (5)

where $R_{T_i-1}$ denotes the tracking patch which is tailored from the tracking result at frame $T_{i-1}$ and $Y$ is the ground truth. $\mathcal{L}_{ce}$ is the cross entropy loss. The operation $\exp$ ensure the
value of $\lambda$ greater than 0. And the hyperparameter $\alpha$ is set to 70 to limit the value of $\lambda$ ranging from 0 to 1.

Obviously, the value of $\lambda$ is inversely related to the cross entropy loss between the tracking result from the last frame and the ground truth, which means that when the tracking result is more accurate, the value of $\lambda$ is more closer to 1. With the template updated by a linear combination, the hyperparameter $\lambda$ is generally set to a fix value. If we replace it with the hyperparameter $\lambda$ in Eq. (4) indicating tracking accuracy, we are able to create an adaptive optimal selection scheme for the transient template. Correspondingly, the complete loss function of ATTONet is written as:

$$L_{trans} = \lambda L_{kl}(\phi(t) | \phi(r) \| \phi(o)) + (1 - \lambda)L_{trans}(\phi(t), \phi(o))$$  \hfill (6)

where $\phi(\ast)$ denotes the feature map, $o$ and $t$ denote the initial template and the transient template respectively, $r$ is the last-frame tracking result tailored to the size of template. $L_{kl}$ is the Kullback-Leibler divergence loss, which contains a temperature hyperparameter $T$ with a fixed value. All loss operations are performed on channel wise. By introducing the adaptive hyperparameter $\lambda$, ATTONet is able to generate a desired template for the better performance of dual-pipeline correlated fusion tracking. As long as the tracking results are accurate, the varying feature of target is integrated with the initial template to fine tune for transient template optimized. But when tracking drift occurs, the transient template mainly depends on the initial template for robust tracking.

3.5 ADF-SiamRPN

Our proposed ADF-SiamRPN is built on the base network DaSiamRPN which is an anchor-based tracker with RPN. Then we expand the network by double Y-shaped cross correlations and learn a syncretic-enhance tracking result from an following embedded correlated fusion network. ADF-SiamRPN inherits the excellent properties from DaSiamRPN like high discriminability from various kinds of training data, distractor suppression against the hard negative samples, and long-term tracking benefited from a local-to-global search region strategy. However, to increase the perceptive range of receptive field to a wider context, we adopt atrous convolution with the rate $k = 2$ in this work. In addition, we substitute the original UP-XCorr in DaSiamRPN layer with a lightweight DW-XCorr layer proposed in SiamRPN++.

**Loss Function:** Owing to the dual-pipeline correlated fusion network, ADF-SiamRPN is composed of two parallel tracking pipelines and a subsequently cascaded fusion network for tracking. Firstly considering the paralleled layer, we can choose the identical loss functions for the two tracking pipelines, with softmax loss for classification loss function $L_{cls}$ and smooth loss for regression loss function $L_{reg}$ as follows:

$$loss_{p(o,t)} = L_{cls}(o,t) + \eta_p L_{reg}(o,t)$$  \hfill (7)

where $p$ denotes the paralleled layer, $o$ and $t$ in set $\{o,t\}$ denote the initial-template pipeline and the transient-template pipeline respectively. $\eta$ is a weight to balance loss and set to 0.2 empirically. We can give the loss function to the fusion layer in the same manner, but replace $p$ with $f$ (denoted for fusion layer) in Eq. (7) without $\{o,t\}$. Because we focus on the final tracking results from the response fusion network, we give a loss function of ADF-SiamRPN to emphasize the role of cascaded layer, which is defined as:

$$loss_{ADF-SiamRPN} = loss_f + \gamma \sum_{\{o,f\}} loss_{p(o,f)}$$  \hfill (8)

where $\gamma$ acts similarly to $\eta$, but is empirically set to 0.1 here, which means a little influence of the paralleled layer on the final tracking results but indispensable.

**Training:** The training of ADF-SiamRPN is performed on the same video datasets like in DaSiamRPN, keeping up with the diversity of category of positive pairs, more semantic negative pairs and different variations of data by customized data augmentation. On the other hand, the correlated fusion network needs the responses from two different tracking pipelines, thus it is necessary to save the intermediate results from different cross correlations separately with the initial template and the transient template for the subsequent training of fusion layer. The loss function in Eq. (8) makes the ADF-SiamRPN capable to be trained in an end-to-end manner.

4. Experimental Results

4.1 Implementation Details

Despite the current works like SiamRPN++ built on ResNet-50 or other deeper CNNs to raise accuracy, we still select the lightweight AlexNet [23] to initialize our backbone network with the weights pre-trained on ImageNet [24] for the future sake of deployment on UAV. Similar to DaSiamRPN, the parameters of the first three layers are frozen and the ones of last two layers are fine tuned. Our network is trained with stochastic gradient descent (SGD) on the image pairs sampled from the same training dataset in DaSiamRPN like COCO [25], ImageNet DET, ImageNet VID and YouTube [26]. Owing to the correlated fusion network feedforward to the final tracking results and feedback to the transient template, we adopt a three-stage iterative training method which means the training results from last stage are ready for the training data in next stage. We train 50 epochs for each stage successively and the learning rate decayed exponentially from 0.001 to 0.0005, 0.0005 to 0.00005, 0.00005 to 0.000005 separately corresponding to each stage. The scales of two different template are unified to 127 pixels and the scale of search patch is set to 255 pixels. Our approach is implemented in Python using PyTorch on a PC with intel i7 8700, 32GB RAM and a NVIDIA GTX-1050Ti GPU.

We conduct extensive experiments on seven datasets like OTB100 [27], UAV123 [28], VOT2016 [29],
Fig. 2 Comparisons with state-of-the-art trackers on precision and success plots on OTB100 dataset.

Fig. 3 Comparisons with state-of-the-art trackers on precision and success plots on UAV123 dataset.

VOT2018 [30], GOT-10k [31], TrackingNet [32] and LaSOT [33] to compare the tracking performance of our approach with state-of-the-art trackers like XF-SiamRPN [34], SiamRPN++, C-RPN, DaSiamRPN, SiamRPN, UpdateNet, GradNet [7], ECO, C-COT, CFNet, SiamFC, SRDCF etc.

4.2 Comparison with State-of-the-Art Trackers

It is demonstrated by lots of experiments that the proposed ADF-SiamRPN excels the compared state-of-the-art trackers, and the processing speed of our method is 65 FPS, almost equivalent to 68 FPS by the XF-SiamRPN which is one of the state-of-the-art methods, under the same experiment condition.

Results on OTB100. OTB100 is a benchmark for tracking which contains 100 video sequences. Figure 2 shows the tracking performance of our ADF-SiamRPN by experiments on this dataset, with the precision plot and success plot in comparison with state-of-the-art trackers. The precision plot reflects the percentages of frames whose estimated locations lying within a given threshold distance to the ground-truth center, while the success plot represents the ratios of successful frames whose Overlap Scores (OS) are larger than a given threshold. Red curve in Fig. 2 represents our ADF-SiamRPN, which performs best with the precision score of 0.907 and the success score of 0.698. Although SiamRPN++ driven by deeper ResNet-50, ADF-SiamRPN with AlexNet surpasses SiamRPN++ on two evaluation metrics. In addition, ADF-SiamRPN behaves more better than XF-SiamRPN and the baseline DaSiamRPN, gaining the improvements of precision by 4.2% and success by 3.8% separately in contrast with XF-SiamRPN.

Results on UAV123. UAV123 is a short video dataset containing 123 sequences. Figure 3 reports the experimental results by the precision plot and success plot on UAV123 in comparison with state-of-the-art trackers. It is obvious that ADF-SiamRPN excels all the other compared trackers under two evaluation metrics, achieving the precision score of 0.809 and the success score of 0.615. SiamRPN++ ranks the second, slightly inferior to our tracker. Meanwhile, ADF-
SiamRPN surpasses XF-SiamRPN by 3.3% on precision and by 3.1% on success, and surpasses the baseline DaSiamRPN by 4% on precision and by 3.7% on success respectively.

Results on VOT2016 and VOT2018. We evaluate our ADF-SiamRPN on Visual Object Tracking challenge 2016 (VOT2016) and 2018 (VOT2018) separately, which are both fit for assessing the single-object tracking performance.

Firstly, we conduct experiments on VOT2016 which contains 60 video sequences, to test the performances of different trackers by accuracy, robustness and Expected Average Overlap (EAO). The evaluation results from different trackers on three metrics are summarized in Table 1.

VOT2018 also contains 60 video sequences, but of which more than 1/6 are substituted by other sequences, and the real-time experiments are introduced in VOT2018. Thus we adopt the realtime evaluation results to compare the performance of ADF-SiamRPN with state-of-the-art trackers, and report the respective scores on three realtime evaluation metrics in Table 2. The tracker with top scores is ADF-SiamRPN, which possesses the identical accuracy to SiamRPN++ and the highest score of 0.426 on EAO and the lowest score of 0.201 on robustness. Undoubtedly, the performance of ADF-SiamRPN excels XF-SiamRPN and the baseline DaSiamRPN as well. For another thing, we further exhibit the concrete behaviors of different trackers by six attributes of influence like occlusion, illumination change, motion change, size change etc. which is depicted in Fig. 4. In accordance with the results in Table 2, each behavior of ADF-SiamRPN on five influence attributes ranks the first, with one identical to SiamRPN++ on occlusion.

Results on GOT-10k. GOT-10k is a large dataset containing 10k videos with 563 categories. There is no overlapped category between the training set and the test set, which enable to promote the generalization of trackers. By experiments on GOT-10k, we compare the performance of ADF-SiamRPN with state-of-the-art trackers as depicted in Fig. 5. The success score by our ADF-SiamRPN is the highest 0.546, which excels SiamRPN++ by 2.9% and XF-SiamRPN by 8.1% and the baseline DaSiamRPN by 10.2%. It is validated that the best tracking performance of ADF-SiamRPN derives from the intrinsic architecture designment of dual-pipeline correlated fusion network.

Table 1 Results on VOT2016 dataset with accuracy, robustness and expected average overlap (EAO).

| Tracker   | EBT | DDC | Staple | MLDF | SSAT | TCNN | C-COT | C-RPN | Da-SiamRPN | XF-SiamRPN | SiamRPN++ | ADF-SiamRPN |
|-----------|-----|-----|--------|------|------|------|-------|-------|------------|------------|------------|------------|
| Accuracy  | 0.465 | 0.542 | 0.547 | 0.492 | 0.579 | 0.555 | 0.541 | 0.594 | 0.618 | 0.631 | **0.642** | 0.641 |
| Robustness| 0.252 | 0.345 | 0.378 | 0.233 | 0.291 | 0.268 | 0.238 | 0.266 | 0.238 | 0.235 | 0.2 | **0.144** |
| EAO       | 0.291 | 0.293 | 0.295 | 0.311 | 0.319 | 0.324 | 0.331 | 0.363 | 0.393 | 0.405 | 0.405 | 0.405 |

Table 2 Results on VOT2018 dataset with accuracy, robustness and expected average overlap (EAO).

| Tracker   | Update-Net | MBSiam | CSTEM | CSR-DCF | Siam-VGG | SA_Siam_P | SA_Siam_R | Da-SiamRPN | XF-SiamRPN | SiamRPN++ | ADF-SiamRPN |
|-----------|------------|--------|-------|---------|----------|-----------|-----------|------------|------------|------------|------------|
| Accuracy  | 0.517 | 0.529 | 0.472 | 0.466 | 0.531 | 0.553 | 0.566 | 0.576 | 0.593 | **0.601** | 0.601 |
| Robustness| 0.534 | 0.44 | 0.379 | 0.318 | 0.337 | 0.342 | 0.258 | 0.29 | 0.288 | 0.234 | **0.201** |
| EAO       | 0.209 | 0.238 | 0.239 | 0.263 | 0.275 | 0.286 | 0.337 | 0.352 | 0.37 | 0.415 | 0.426 |

Fig. 4 The ranking plot of overlap scores under six attributes of video sequences on VOT2018.

Fig. 5 Comparisons with state-of-the-art trackers on success plot on GOT-10k dataset.
Results on LaSOT. LaSOT is composed of 1400 videos and fit for long-term tracking. We conduct experiments on LaSOT in comparison with 9 trackers, and contrast the respective tracking performances on two metrics as shown in Fig. 6. It is demonstrated that our ADF-SiamRPN behaves best with the normalized precision score 0.577 and success score 0.498 respectively. In comparison, SiamRPN++ is inferior to ADF-SiamRPN by 0.7% on precision and by 0.3% on success. Meanwhile, ADF-SiamRPN surpasses the following C-RPN, XF-SiamRPN and the baseline DaSiamRPN.

Results on TrackingNet. TrackingNet contains 30000 video sequences with 511 sequences in test, fit for long-term tracking. The experiments on TrackingNet focus on the performance comparison among 8 different trackers on evaluation metrics like precision and success, which is shown in Table 3. By contrast, our ADF-SiamRPN achieve the best scores 0.694 on precision among 8 trackers and 0.799 on normalized precision, 0.732 on success nearly identical to SiamRPN++ but with a little gap of 0.03% which is not shown by rounding. The reason may be the deeper ResNet-50 enabling SiamRPN++ more discriminative.

4.3 Ablation Study

We study the impact of individual components in ADF-SiamRPN from the perspective of model architecture, and conduct ablation study on VOT2018. Based on the baseline DaSiamRPN, we test the variation of tracking performance with the change of architecture or the addition or substitution of modules one by one. The results are concluded in Table 4 and the baseline achieves the EAO score 0.352. When adding the other cross-correlation pipeline with the last-frame feedback tracking result and fusing the results of dual-pipeline correlation, the tracking accuracy declines and the EAO score is decreased by 0.5%. The main reason is that the inevitable tracking drift happens. If we use a linear combination of initial template, last-frame transient template and last-frame tracking result like in [34] to generate the transient template for the other correlation operation, the whole tracking performance is promoted and the EAO score is increased by 1.8%. Furthermore, we substitute the linear combination with our proposed adaptive optimal selection for transient template. Atr: atrous convolution.

### Table 3: Results on TrackingNet dataset with precision, normalized precision and success.

| Method     | Precision | Normalized Precision | Success |
|------------|-----------|----------------------|---------|
| SiamFC     | 0.533     | 0.666                | 0.571   |
| MDNet      | 0.557     | 0.705                | 0.606   |
| UPDT       | 0.613     | 0.751                | 0.655   |
| XF-SiamRPN | 0.622     | 0.771                | 0.703   |
| DaSiamRPN  | 0.648     | 0.798                | 0.732   |
| ATOM       | 0.693     | 0.799                |         |
| SiamRPN++  | 0.694     |                      |         |

### Table 4: Ablation study on VOT2018. DaSiamRPN is baseline. DPCF: dual-pipeline correlation cascaded by fusion network. LCTT: linear combination for transient template. AOSTT: adaptive optimal selection for transient template. Atr: atrous convolution.

| Method                  | A↑ | R↑ | EAO↑ | ΔEAO |
|-------------------------|----|----|------|------|
| Baseline                | 0.576 | 0.29 | 0.352 | -    |
| Baseline+DPCF           | 0.569 | 0.29 | 0.347 | -0.5%|
| Baseline+DPCF+LCTT      | 0.593 | 0.288 | 0.37 | 1.8% |
| Baseline+DPCF+AOSTT     | 0.6 | 0.242 | 0.403 | 5.1% |
| Baseline+DPCF+Atr (Ours) | 0.601 | 0.201 | 0.426 | 7.4% |
4.4 Qualitative Evaluation

In this section, we will give the qualitative tracking results of our ADF-SiamRPN in comparison with SiamRPN++, XF-SiamRPN and DaSiamRPN in two intuitive ways.

Firstly, we visualize the respective tracking results of four trackers and the corresponding heatmaps of tracking responses to some video clips of VOT2018, which is shown in Fig. 7. Besides, the cosine window is overlayed to correct the responses. Observing the tracking results at the leftmost column, the bounding boxes generated by ADF-SiamRPN and SiamRPN++ are nearest to the ground truth in each video clip except for the third one, because of the deviation from SiamRPN++. Although existing a certain error, the bounding box produced by ADF-SiamRPN contains the ground truth. Regarding heatmaps, the response peaks of XF-SiamRPN and the baseline DaSiamRPN are dispersive especially when distractors, cluttered background or motion blurring etc. happen, but the response peak of ADF-SiamRPN in each heatmap locates at the target exactly, the similar to SiamRPN++. Therefore, ADF-SiamRPN has a better ability to track the target.

Secondly, we exhibit some tracking samples from continuous frames of different video sequences as shown in Fig. 8. In the first-row video sequence, the cluttered background and distractors are the obstacle for tracking. The baseline DaSiamRPN appears deviation at the beginning, and XF-SiamRPN deviates from the ground truth at last. Relative to SiamRPN++, the bounding box generated by ADF-SiamRPN is more matching to the ground truth. In the second-row video sequence, with the tracking difficulty from the cluttered background, the bounding boxes generated by ADF-SiamRPN and SiamRPN++ nearly encompass the ground truth, demonstrating their real-time tracking ability, but the tracking results from XF-SiamRPN and DaSiamRPN are deviated. Similarly in the third-row video sequence, with the nearby distractor, the tracking results from ADF-SiamRPN and SiamRPN++ are all close to the ground truth but for XF-SiamRPN and DaSiamRPN. In conclusion, the proposed ADF-SiamRPN is able to effectively handle the visual tracking scenes with distracted foregrounds, cluttered background and motion blurring etc.

5. Conclusion

In this paper, we propose a novel Siamese tracking network ADF-SiamRPN, which promotes the tracking performance by fusing two responses from dual-pipeline cross correlation. To be specific, dual-pipeline cross correlation respectively contains one cross correlation operation of the search region with the initial template, and the other cross correlation operation of the search region with a novel transient template which is generated by an adaptive optimal selection strategy. The designmen of double Y-shape cross correlation and the learnable correlated fusion network enable ADF-SiamRPN to boost accuracy while keep robustness for tracking. We train ADF-SiamRPN in an end-to-end manner and conduct extensive experiments to test the tracking performance of ADF-SiamRPN on seven visual tracking benchmarks. By comparison with state-of-the-art trackers, it is verified that our method achieves new state-of-the-art tracking results with the best balance between accuracy and robustness.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China under Grant No.2017YFB1001901 and in part by the National Natural Science Foundation of China under Grant No.61906212.

References

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

[2] J.F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High-speed tracking with kernelized correlation filters,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.37, no.3, pp.583–596, 2014.

[3] M. Danelljan, A. Robinson, F.S. Khan, and M. Felsberg, “Beyond correlation filters: Learning continuous convolution operators for visual tracking,” European Conference on Computer Vision, pp.472–488, Springer, 2016.

[4] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, and P.H.S. Torr, “Fully-convolutional siamese networks for object tracking,” European Conference on Computer Vision, pp.850–865, Springer, 2016.

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

[6] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, “High performance visual tracking with siamese region proposal network,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.8971–8980, 2018.

[7] P. Li, B. Chen, W. Ouyang, D. Wang, X. Yang, and H. Lu, “Gradnet: Gradient-guided network for visual object tracking,” Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.6162–6171, 2019.

[8] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, “Distractor-aware siamese networks for visual object tracking,” Proceedings of the European Conference on Computer Vision (ECCV), pp.101–117, 2018.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.40, no.4, pp.834–848, 2017.

[10] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, “Accurate scale estimation for robust visual tracking,” British Machine Vision Conference, Nottingham, Sept. 1-5, 2014, BMVA Press, 2014.

[11] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang, “Hedged deep tracking,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.4303–4311, 2016.

[12] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, “Multicue correlation filters for robust visual tracking,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.4844–4853, 2018.

[13] J. Choi, H.J. Chang, T. Fischer, S. Yun, K. Lee, J. Jeong, Y. Demiris, and J. Young Choi, “Context-aware deep feature compression for high-speed visual tracking,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.479–488, 2018.

[14] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, “Visual tracking via adaptive spatially-regularized correlation filters,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.4670–4679, 2019.

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

[16] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, “Siamese tracking with very deep networks,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4282–4291, 2019.

[17] H. Fan and H. Ling, “Siamese cascaded region proposal networks for real-time visual tracking,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.7952–7961, 2019.

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

[19] H. Law and J. Deng, “Cornernet: Detecting objects as paired keypoints,” Proceedings of the European Conference on Computer Vision (ECCV), pp.734–750, 2018.

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

[21] L. Zhang, A. Gonzalez-Garcia, J.V.D. Weijer, M. Danelljan, and F.S. Khan, “Learning the model update for siamese trackers,” Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.4010–4019, 2019.

[22] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, “A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.9060–9069, 2020.

[23] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, vol.25, pp.1097–1105, 2012.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” International Journal of Computer Vision, vol.115, no.3, pp.211–252, 2015.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, “Microsoft coco: Common objects in context,” European Conference on Computer Vision, pp.740–755, Springer, 2014.

[26] E. Real, J. Shlens, S. Mazzucchelli, X. Pan, and V. Vanhoucke, “Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.5296–5305, 2017.

[27] Y. Wu, J. Lim, and M.-H. Yang, “Object tracking benchmark,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.37, no.9, pp.1834–1848, 2015.

[28] M. Mueller, N. Smith, and B. Ghanem, “A benchmark and simulator for uav tracking,” European Conference on Computer Vision, pp.445–461, Springer, 2016.

[29] F. Battistone, V. Santopietro, and A. Petrosino, “The visual object tracking vot2016 challenge results,” Proceedings of European Conference on Computer Vision, pp.777–823, 2016.

[30] M. Kristan, J. Matas, A. Leonardis, T. Vojíř, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin, “A novel performance evaluation methodology for single-target trackers,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.38, no.11, pp.2137–2155, 2016.

[31] L. Huang, X. Zhao, and K. Huang, “Got-10k: A large high-diversity benchmark for generic object tracking in the wild,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.43, no.5, pp.1562–1577, 2019.

[32] M. Müller, A. Bibi, S. Giancola, S. Alsubaibii, and B. Ghanem, “Trackingnet: A large-scale dataset and benchmark for object tracking in the wild,” Proceedings of the European Conference on Computer Vision (ECCV), pp.300–317, 2018.

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

[34] N. Wang, D. Shi, Y. Kang, Z. Fan, S. Jin, and Y. Zhang, “Dual-template siamese network with cross-correlation fusion for tracking,” 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), pp.1194–1201, IEEE, 2020.
Ying Kang received the B.S. degree in communication engineering from Inner Mongolia University, Hohhot, China, in 2007, and the M.S. degree in electromagnetic field and microwave technology from Tianjin University, Tianjin, China, in 2010, and the Ph.D. degree in computer architecture from Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, in 2015. Since 2019, she has been an Assistant Professor in Defense Innovation Institute, Beijing, China. Her research interests include artificial intelligence, computer vision, visual tracking.

Cong Liu received the B.S. degree in computer science and technology from North China Electric Power University, Baoding, China, in 2019. He is currently pursuing the M.S. degree in computer science and technology at National University of Defense Technology, Changsha, China. His research interests include computer vision, visual tracking, optical flow estimation and event camera.

Ning Wang received the B.S. degree in software engineering from Shandong University, Shandong, China, in 2018. Since 2018, he has been pursuing the M.S. degree in computer science and technology at National University of Defense Technology, Changsha, China. His research interests include computer vision and its application in visual object tracking.

Dianxi Shi is a Researcher and the Deputy Director of the Artificial Intelligence Research Center in Defense Innovation Institute. He received the B.S. degree, M.S. degree and Ph.D. degree in computer science from National University of Defense Technology, Changsha, China, in 1989, 1996 and 2000 respectively. His research interests include distributed object middleware technology, software component technology, adaptive software technology and intelligent unmanned cluster system software architecture.

Mengmeng Li received the B.S. degree in information system engineering from National University of Defense Technology, Changsha, China, in 2018. She is currently pursuing the M.S. degree in computer science and technology at Defense Innovation Institute, Beijing, China. Her research interests include data quality, evolutionary algorithms.

Ning Zhou received the B.S. degree in communication engineering from Xi’an Communication College of PLA, Xi’an, China, in 2008. He has been engaged in the research and application of wired communication and wireless communication for ten years. His current research interests include wired communication, wireless communication, artificial intelligence.

Yunlong Wu received the B.S. degree in information security from Wuhan University, Wuhan, China, in 2011, and the M.S. degree in computer science and technology and the Ph.D. degree in microelectronics and solid electronics from National University of Defense Technology, Changsha, China, in 2013 and 2017 respectively. Since 2018, he has been an Assistant Professor in Defense Innovation Institute, Beijing, China. His main research interests include robotic communications, multi-agent systems, machine behavior, and swarm intelligence.

Mengmeng Li received the B.S. degree in information system engineering from National University of Defense Technology, Changsha, China, in 2018. She is currently pursuing the M.S. degree in computer science and technology at Defense Innovation Institute, Beijing, China. Her research interests include data quality, evolutionary algorithms.