Received December 11, 2020, accepted December 16, 2020, date of publication December 23, 2020,
date of current version January 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3046731

Siamese Attentional Cascade Keypoints Network for Visual Object Tracking

ERSHEN WANG, DONGLEI WANG, YUFENG HUANG, GANG TONG, SONG XU, AND TAO PANG
School of Electronic and Information Engineering, Shenyang Aerospace University, Shenyang 110136, China

Corresponding author: Yufeng Huang (yufengh_sau@sina.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61571309 and Grant 61703287, in part by the Key Research and Development projects of Liaoning Province under Grant 2020JH2/10100045, in part by the Liaoning Education Department of Science and Technology Research Project under Grant JYT2020142/2020030, in part by the Natural Science Foundation of Liaoning Province under Grant 2019-MS-251, in part by the Talent Project of Revitalization Liaoning under Grant XLYC1907022, and in part by the High-Level Innovation Talent Project of Shenyang under Grant RC190030.

ABSTRACT Visual object tracking is urgent yet challenging work since it requires the simultaneous and effective classification and estimation of a target. Thus, research on tracking has been attracting a considerable amount of attention despite the limitations of existing trackers owing to deformation, occlusion and motion. For most current tracking methods, researchers have proposed various ways to adopt a multi-scale search or anchors for estimation, but these methods always need prior knowledge and too many hyperparameters. To address these issues, we proposed a novel Siamese Attentional Cascade Keypoints Tracking Network named SiamACN to exactly track the object by using keypoints prediction instead of anchors. Compared to complex target prediction, the anchor-free method is performed to avoid plaguy hyperparameters, and a simplified hourglass network with global attention is considered the backbone to improve the tracking efficiency. Further, our framework uses keypoints prediction around the target with cascade corner pooling to simplify the model. To certificate the superiority of our framework, extensive tests are conducted on five tracking benchmarks, including OTB-2015, VOT-2016, VOT-2018, LaSOT and UAV123. Our method achieves the leading performance with an accuracy of 61.2% on VOT2016 and favorably runs at 32 FPS against other competing algorithms, which confirms its effectiveness in real-time applications.

INDEX TERMS Visual object tracking, siamese network, hourglass network, global attention, cascade corner pooling.

I. INTRODUCTION

Visual object tracking (VOT) is a foundational and critical challenge that has received considerable attention as it is widely applied in unmanned driving, intelligent surveillance and video editing [1]. Given the initial target size or position, the tracking task needs to predict an object in each subsequent frame. Although substantial breakthroughs have been achieved in VOT algorithms [2], VOT remains a difficult topic due to unconstrained conditions and suffers from scale and position variation, complex background and heavy occlusion [3], [4]. Based on the estimated initial bounding boxes, the training samples for model updating could be accurate, which would gradually cause model degradation over time [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Vivek Kumar Sehgal.

Many traditional visual tracking methods utilize discriminative correlation filter (DCF) [6]–[8] frameworks. Considering the lack of computational cost and capacity, a DCF-based framework is chosen for its highly effective calculation in the Fourier domain. The main reason for this choice is that manually produced features can represent targets and target modeling is rigid [9]. However, these features are less effective in complex environments. Deep learning-based tracking [5], [10]–[15] has been widely employed due to its strong capacity of learning powerful deep features. Subsequently, inspired by deep learning breakthroughs, a substantial amount of work in object detection [16]–[18] and Visual Object Tracking (VOT) has been conducted [19], [20]. Deep learning-based methods can dominate online learning [21]–[25] and one-shot learning [26] in the short-term tracking field. Online learning trackers [21], [23] are training with less data while becoming more accurate with the help of new incoming data. Some of them (e.g., ECO [21], TFCR [25]) use improved DCF

frameworks for online training. One-shot learning, such as SiamFC [10] and SiamRPN [27], may serve the VOT as a target matching problem and attempt to learn the similarity valuation between the search region and the target.

Focusing on offline tracking, Siamese network-based trackers [10], [11], [27]–[31] are trained by collecting pairs of frames, and there are some difficulties in accurately tracking the targets with fast motion, large-scale variation or occlusions [32]. To address these problems, SiamFC [10] estimated the bounding box using a multi-scale search mechanism. To handle aspect ratio changes, SiamRPN [27] introduced the region proposal networks [33] into the tracking process and obtained higher precision on bounding box estimation. However, the design of anchors is crucial to tracking performance in the region proposal. To cope with changes to the shape and scale of the target, the anchors need to be designed in advance for quantity, different sizes, and aspect ratios. The anchors need prior knowledge to define, which introduces too many hyperparameters. There are a large number of anchors, but only a small number of fractions actually have a high overlap with the ground truth, which introduces computational complexity.

In this paper, we design an efficient anchor-free Siamese network-based visual tracking framework to address the challenge of state estimation and achieve excellent performance. First, an hourglass network is involved in the feature exaction process, so that we can obtain the multiscale features to help in the subsequent tracking process. Second, global attention is used to predict the rough target position as well as improve the efficiency and accuracy of the tracking framework. Inspired by the related anchor-free detectors method, our framework applies the anchors and adopts corner detection to accurately predict the bounding box. We can use only $O(wh)$ keypoints to present the possible anchor boxes that correspond to $O(w^2h^2)$ in a feature map of size $w \times h$.

Our main contributions can be summarized as follows:

(1) The global attention hourglass network is designed for feature extraction. The hourglass network can obtain the multiscale information for the initial target, which can help to obtain detailed and comprehensive features. The attention mechanism is performed to obtain the global information and compressed integration information, so that the method can effectively obtain important feature information and increase the processing efficiency.

(2) We introduced the advance corner detection network to improve the subsequent tracking process, so that it can improve the tracking speed and properly identify the missing target. Anchor-free detection is employed to predict the bounding box. To accelerate the prediction step, the top-left and bottom-right corner information are obtained to confirm the bounding box. Using cascade corner pooling, we can construct the heatmaps, embeddings and offset corners of the tracking target.

(3) Numerous tracking experiments are carried out on several classical datasets, including OTB-2015 [34], VOT-2016 [19], VOT-2018 [20], LaSOT [35] and UAV123 [36]. By the quantitative and qualitative analysis, our proposed tracking framework outperforms other outstanding methods in a certain area. Further, related ablation studies have been applied to verify the reasonable parts and parameters in the proposed framework.

II. RELATED WORK

Various frameworks have been investigated to increase the tracking process. In this section, we summarize the most related two aspects: Siamese network based visual trackers and corner detection-based detectors.

A. SIAMESE NETWORK BASED VISUAL TRACKERS

Generally, visual object tracking involves several aspects, such as feature extraction [7], [37], classifier design [38] and bounding box regression [23]. We assume that the template patch is z, the search patch is x, and $f(\bullet)$ is a function to measure similarity. We denote $\varphi(x)$ and $\varphi(z)$ as the output feature maps of the Siamese subnetwork and their similarity as $f(\varphi(x), \varphi(z))$. In online tracking, the initial frame is employed as a template, and the target is search in following video sequences. Current studies indicate that the Siamese-based tracking framework have achieved success due to its strong training capabilities and high accuracy [10], [12], [27], [29], [39], [40]. SiamFC [10] first employed the Siamese network to extract feature information and combined feature maps with the correlation layer. Owing to its success in tracking, the researchers design some advanced models and obtain better tracking results. By using the correlation filter, the CFNet [11] improved the Siamese network in the feature exaction layer to increase the tracking accuracy. In [12], DSiam investigated feature transformation to modify the Siamese branches and improve the accuracy by suppressing the background. The varied attention mechanisms are involved in the RASNet [39], so that the tracking model can adapt to the target. To obtain a more precise bounding box, SiamRPN [27] introduced the RPN [33] in the SiamFC, so that it can avoid complex multiscale computations. Inspired by the SiamRPN, DaSiamRPN [31] improved the discrimination of the tracker by adding hard negative data in the training process. RAR [41] uses LSTM to integrate the DCF framework as a correlation layer into the Siamese network. Advanced Siamese networks, such as the SiamRPN++ [29], SiamMask [42] and SiamDW [40], optimized the architecture by using modern deep networks. SPM-Tracker [32] combined coarse and fine matching to improve the robustness and power of discrimination. The designs of the anchors in these trackers avoid time-consuming multiscale feature extraction.

B. OBJECT DETECTORS FRAMEWORK

Due to similar characteristics, visual object tracking may follow the tracking-by-detection paradigm [27], [43]–[45]. Many recent approaches have improved the efficiency of tracking since the introduction of the detector, which may consider tracking as a whole or parts in detection problems [27], [46]. Inheriting from Faster-RCNN [33], the RPN structure achieves excellent accuracy in the SiamRPN [27].

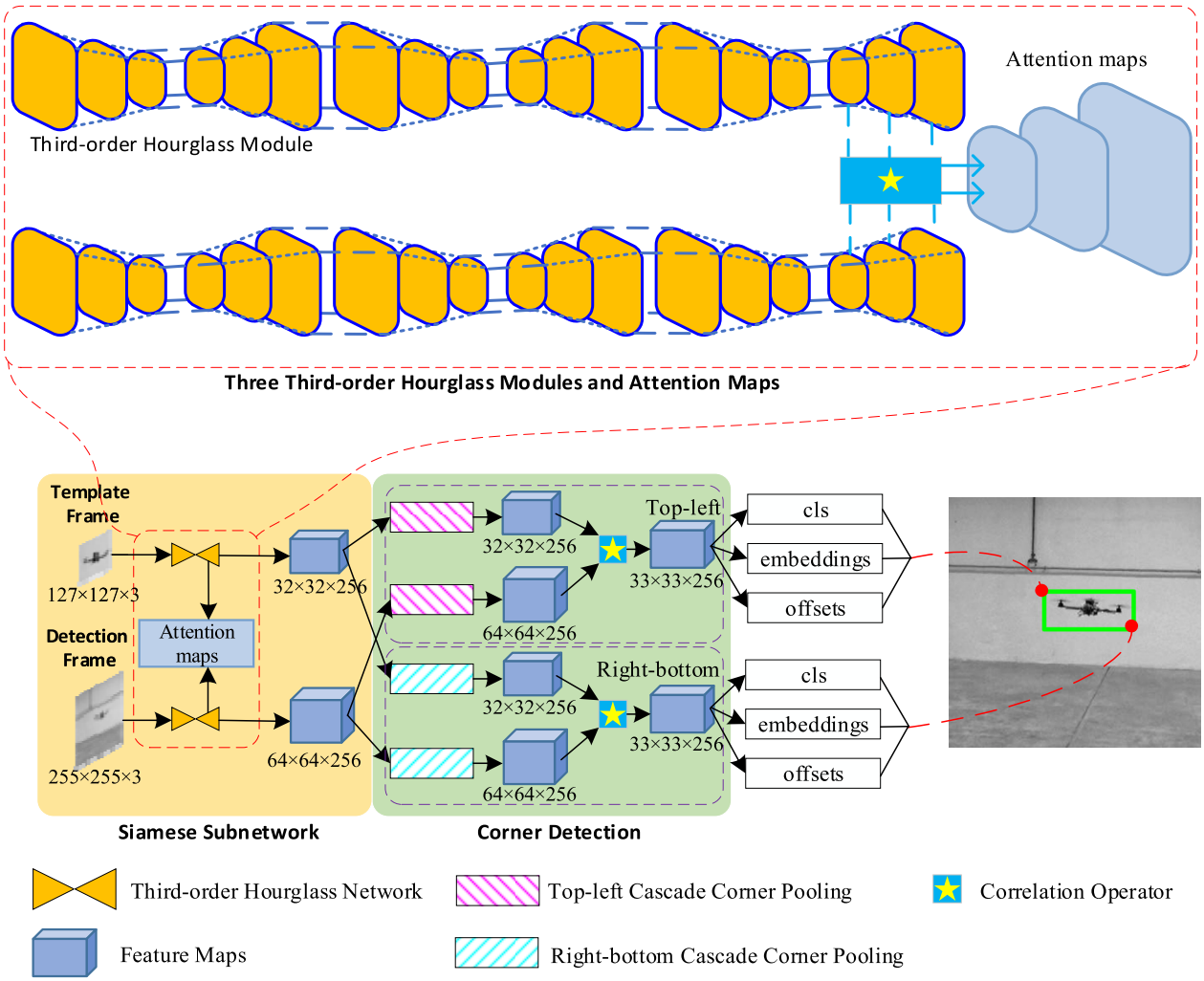


FIGURE 1. Framework of SiamACN: left side shows the Siamese subnetwork for feature extraction. We adopt three third-order hourglass modules as the backbone network. In the middle, the green box is the corner detection module, which has two branches: one branch on the top left and another branch on the bottom right. The 'cls', 'embeddings', and 'offsets' denote that in each branch, we predict the heatmap of classification, embeddings and offsets of the keypoints.

SATIN [47] introduces CornerNet [48] and spatiotemporal attention mechanisms to directly track target corners. DCF-based tracking methods [7], [44] detect the targets via matching a search for the highest score on similarity score maps. To obtain similarity score maps, some methods [6], [7] obtain results by transferring the candidate feature maps into a trained correlation filter and other trackers [21], [22] directly calculate the correlation between the example and candidate feature maps. The anchor-based detectors are extensively employed in the tracking methods, which classify the anchors as negative or positive [33], [49], [50]. These detectors set the anchor to obtain an extra offset regression to refine the bounding box prediction. Although anchors can help in the tracking or detection process, the hyperparameters of an anchor can have a substantial effect on the final accuracy [51], [52]. Thus, the anchor-free detectors attract the attention of tracking researchers, who predict bounding boxes at certain points [53], [54] or detect and classify a pair of corners in the proper way [48]. CornerNet [48] is an effective anchor-free detection method that abandons the design of the anchors and directly detects an object by detecting a couple of corners. Simplifying the hourglass network in CornerNet from 104 layers to 54 layers, CornerNet-Saccade [55] improves the speed of detection and adds a saccade mechanism to maintain the detection accuracy. CenterNet [56] adds center prediction and proposes cascading corner pooling to improve the accuracy and perform better tracking.

III. METHODS

In this section, we introduce the proposed SiamACN framework. Fig. 1 shows that our SiamACN uses the global attention hourglass network instead of the former backbone Siamese network for feature extraction. Followed by the corner detection modules, cascade corner pooling is employed for bounding box prediction. Thus, the corner detection network can produce classification, regression and embedding information for the tracking process.

A. SIAMESE FEATURE NETWORK

The Siamese network [59] has been proven to be effective in the tracking process. Here, we adopt a fully convolution network without padding to extract the features. In the tracking

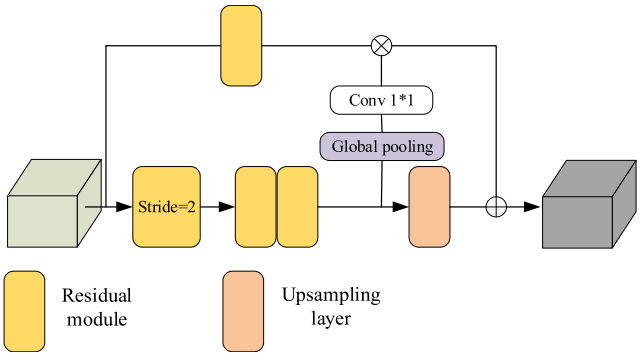


FIGURE 2. First-order global attention hourglass module.

process, an hourglass network is applied to the feature extraction network, which will be detailed in the next section. The Siamese network generally consists of two equal branches: the first branch is a template branch and the second branch is the search branch. These two branches share parameters to ensure a similar transformation to each branch. To reduce the computational cost, we add the convolution stages to reduce the image feature resolution. The details of each branch are discussed in the next part.

B. HOURGLASS BASED SIAMESE TRACKING

The hourglass network [58] is introduced as the backbone of the Siamese tracking process. This network contains at least one hourglass module, which can preserve low- and high-level information across different resolutions by using a series of down-sample and up-sample processes. We further modify the architecture of the hourglass module inspired by the PAN [59], which is the global attention hourglass module shown in Fig. 2. Instead of using a single residual module, we employ the improved global attention module in each skip connection for upsampling. We use a residual module with stride 2 to downsample the input feature maps and another residual module to change the channels of the feature maps. Our global attention hourglass module performs global pooling on the high-level features after downsampling to provide global context. The skip layer is composed by the global context, which is obtained via a 1×1 convolution, and then multiplied by the processed input feature. We apply Nearest Neighbor (NN) Interpolation for high-level feature maps to upsample the features across scales. This module can refine the comprehensive information of the category and provide more precise resolution details. Our hourglass network consists of 3 third-order hourglass modules. We apply a convolution layer with stride 2 and a residual module to downsample the image feature resolution 4 times.

We utilize 3 attention maps at different scales to predict large-, medium- and small-sized objects in the upsampling layers of the last hourglass module. As shown in Fig. 1, we separately correlate the last three upsampling layers of the template and detection branch and then employ a 3×3 Conv-ReLU module and a 1×1 Conv-Sigmoid module to obtain the attention maps. The attention maps are employed

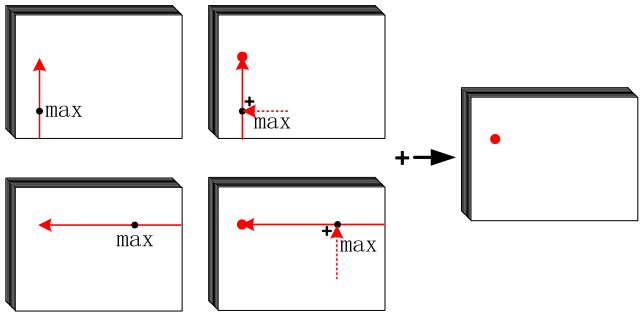


FIGURE 3. Example of top-left cascade corner pooling by grouping corner pooling in multiple directions.

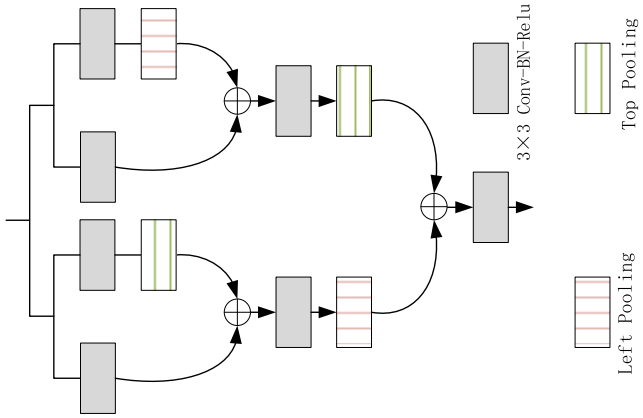


FIGURE 4. Structure of cascade top-left corner pooling.

to obtain a suitable size for the foreground area as the next stage of the fine inspection search frame.

C. CORNER DETECTION

1) CASCADE CORNER POOLING

Two corner prediction modules follow the backbone network. There are often no local appearance features for the corners, which are located outside the object. CornerNet [48] uses corner pooling to solve this problem. However, corner pooling can easily render the corner features affected by the edges and lacks object information. To address these problems, we introduce cascade corner pooling from CenterNet [56]. As shown in Fig. 3, first, the cascade corner pooling scans from the boundary for a boundary max-pooling. Second, it scans inside the location of the boundary maximum value for the internal max-pooling. Last, it combines two maximum values.

We combine the corner pooling in different directions to achieve cascade top-left corner pooling and cascade bottom-right corner pooling. The structure of the cascade top-left corner pooling module is shown in Fig. 4. For top pooling, we add a left pooling layer before the origin boundary top pooling in corner pooling; for left pooling, we add a top pooling layer before the origin boundary left pooling in corner pooling. We combine top pooling and left pooling together to obtain cascade top-left corner pooling. The same process is performed for bottom-right corner pooling. For bottom pooling, we add a right pooling layer before the

origin boundary bottom pooling in corner pooling; for right pooling, we add a bottom pooling layer before the origin boundary right pooling in corner pooling. We combine bottom pooling and right pooling to obtain cascade bottom-right corner pooling. In this way, the corners can learn richer object information.

For the feature map of the output of the cascade corner pooling layer, we compute the correlation on both the top-left branch and the bottom-right branch, As shown in Fig. 1, we predict the heatmaps as 'cls' for each branch to represent the locations of the corner keypoints that contain targets. The heatmap in each branch has only one channel for foreground background classification. We simultaneously predict two embedding vectors as 'embeddings' for the top-left branch and the bottom-right branch to group corner keypoints. In each branch, an embedding vector is used to separate the background and shorten the distance between the target corner pairs. Specifically, if a pair of top-left and bottom-right corner keypoints originate from the same bounding box, then the distance between their embedding vectors must be small. Offsets have two channels for each branch to fine-tune the horizontal and vertical locations of the corners. We adopt simple post-processing to locate the final bounding box.

2) TRAINING LOSS

During training, we adopt the focal loss L_{cls} to detect the classification of keypoints for heatmaps. The smooth L1 loss L_{off} is applied to predict the offsets between the prediction corners and the ground truth corner locations. We apply the ''pull'' loss L_{pul} to group the corners that belong to an object and the ''push'' loss L_{pus} to detach the corners between the foreground and the background. We take the template frame and detection frame via the three up-sampling layers of the last hourglass module of the backbone network and obtain three attention maps after the correlation operation to pre-detect the possible object position. We also utilize the focal loss L_{att} to predict the attention maps.

We optimize the full training loss function that is obtained by combining the introduced loss functions to train our network end-to-end:

$$L = \alpha L_{att} + \gamma L_{cls} + \lambda L_{emb} + \eta L_{off}, \tag{1}$$

where $L_{emb} = \theta L_{pul} + \varpi L_{pus}$ denotes the loss of embedding, α, β, λ and η denote the weights to balance the full training loss.

D. TRACKING DETAILS

The SiamACN uses attention maps to obtain the possible object locations. Different object sizes determine different zoom ratios in these possible locations. We set $S_s = 4$ for the ratio of a small-sized object, $S_m = 2$ for the ratio of a medium-sized object and $S_l = 1$ for the ratio of a large-sized object. We enlarge the image by regulating S_i, and then detect the object at the possible object detections by cascade corner pooling.

During the tracking process, we add global searching for a lost target when an object is not in the cropped search frame. Instead of using the cropped search frame, we track by searching from the original size of an image. We set the threshold $t = 0.2$. If the score of the final bounding box is less than the threshold, we will preserve the bounding box and use global search re-track. After corners prediction, we adopt the Soft-NMS [60] to remove redundant locations and improve the accuracy.

IV. EXPERIMENTS

In this section, we perform a detailed implementation and comprehensively compare our results with the state-of-the-art methods on five benchmark datasets: OTB-2015 [34], VOT-2016 [19], VOT-2018 [20], LaSOT [35] and UAV123 [36]. Necessary ablation studies have not been carried out to prove the effectiveness of the designed components.

A. IMPLEMENTATION DETAILS

Our SiamACN is implemented using PyTorch [61] in python 3.7 and runs on one NVIDIA Tesla V100 GPU with 16 GB of VRAM. We apply the modified hourglass network as the backbone network with no pretraining on any datasets, and the sample image pairs are picked from YouTube-BB [62], VID [63], DET [63] and COCO [64] datasets to train the whole network. Adaptive moment estimation (Adam) is carried out to train the network with 20 epochs. According to the SiamRPN [27], the parameter is set to a warmup learning rate that increases from 0.001 to 0.005 in the first 5 epochs, and a learning rate decays exponentially from 0.005 to 0.00005 for the last 15 epochs. We set the input size of the template patches to 127×127 and the size of the search patches to 255×255. The code will be released on the GitHub.

B. COMPARISONS WITH THE STATE-OF-THE-ART

The network is initialized with its default setting without any pretraining on other external datasets. Here, we evaluate the tracking methods on the five benchmark datasets. The detailed results are presented as follows:

1) OTB2015

OTB2015 is one of the most widely employed VOT benchmark datasets, which are composed of approximately 100 challenging videos. One-pass evaluation (OPE) is an important evaluation index that has two metrics: precision score (PS) and area under curve (AUC). The PS is the percentage of frames whose tracking results lie in a 20-pixel distance to ground truth centers. The AUC of a success plot is the area under the plot that contains ratios of successfully tracked frames at the thresholds that range from 0 to 1. We compare our tacker with 9 state-of-the-art trackers, including ATOM [23], Da-SiamRPN [31], SiamRPN [27], CREST [15], SINT [65], CFNet [11], SiamFC [10], Staple [66], and HCF [67]. The PS and AUC results are shown in Fig. 5. From the figure, we obtain the third-best precision of 87.5% and an AUC of 64.5%. Although these results are not optimal, they are

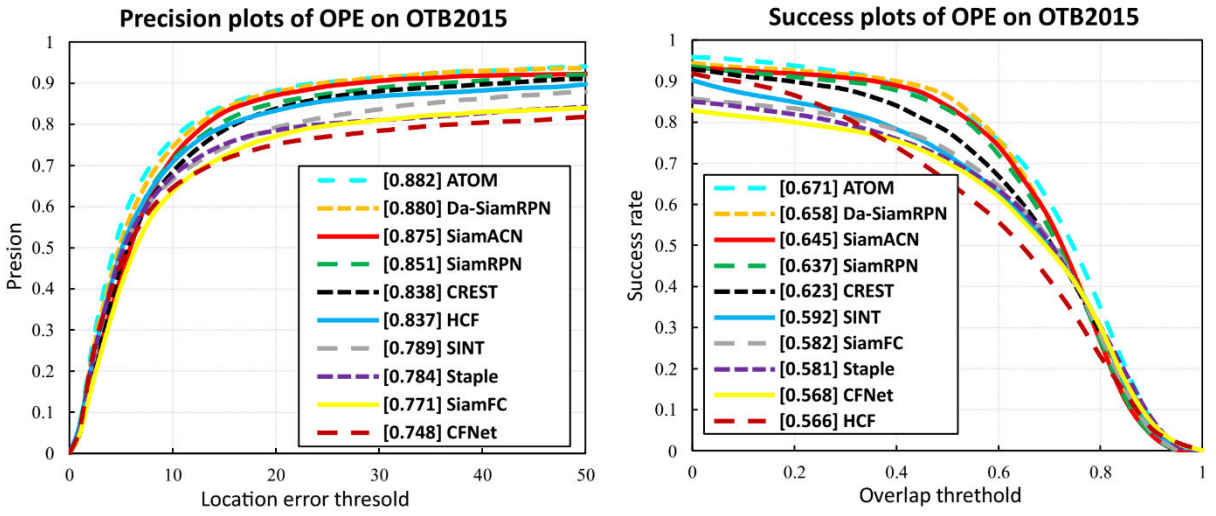


FIGURE 5. Comparison on OTB2015 with the evaluation metrics of precision and success plots in one-pass evaluation.

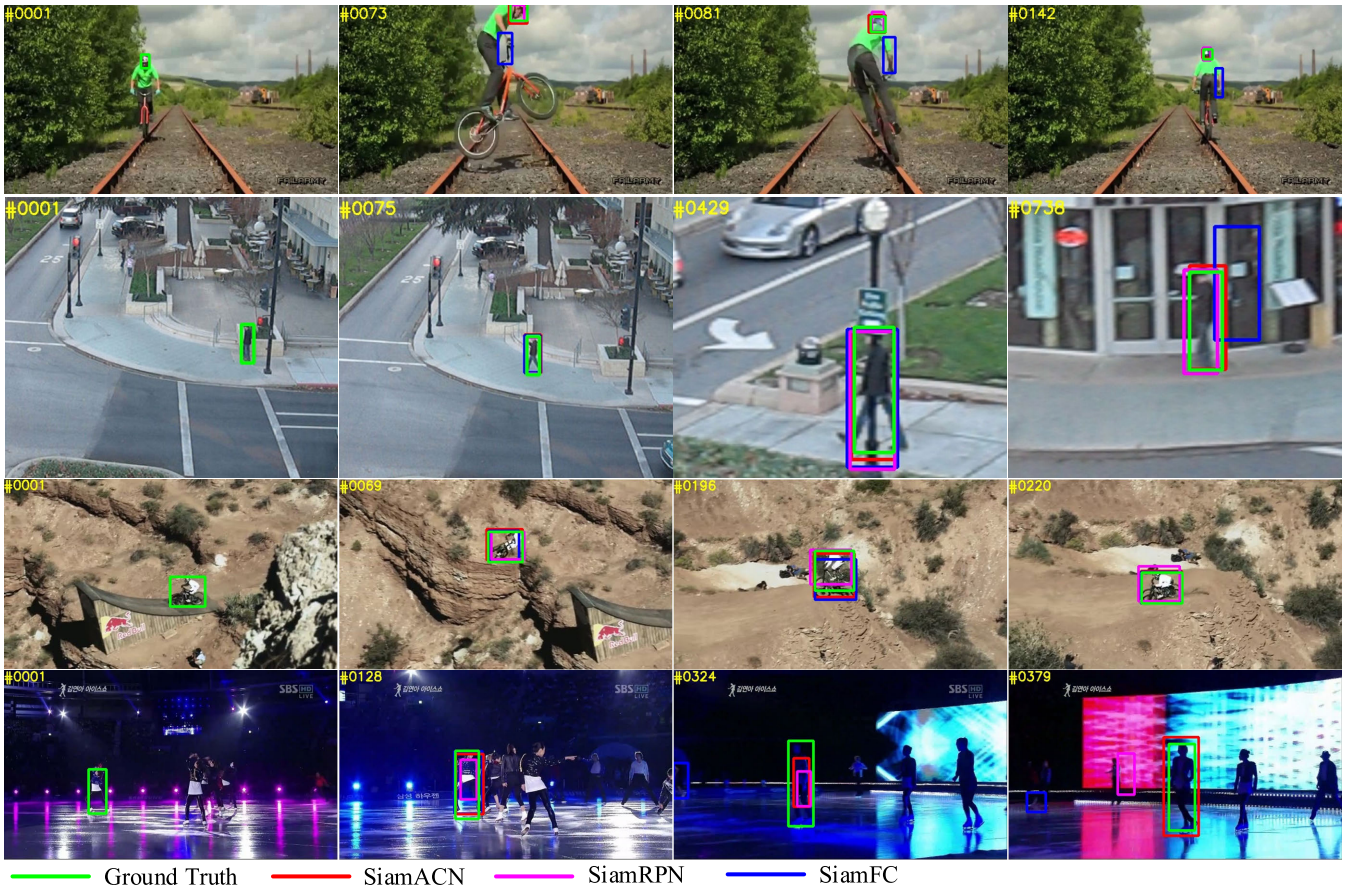


FIGURE 6. Example tracking results of our approach with other Siamese network based trackers for three challenging sequences (from top to bottom: *Biker*, *Human6*, *MountainBike* and *Skating1*).

satisfactory and improve the scores of the precisions by 2.4% compared with the SiamRPN.

The qualitative analysis results for OTB2015 are shown in Fig. 6. Eleven attributes are used to flag the challenges faced by the sequences in OTB2015, including Background Clutters (BC), Deformation (DEF), Fast Motion (FM), In-Plane Rotation (IPR), Illumination Variation (IV), Low Resolution (LR), Motion Blur (MB), Occlusion (OCC), Out-of-Plane Rotation (OPR), Out-of-View (OV), and Scale Variation (SV) [68]. We compare our SiamACN with SiamFC and SiamRPN, which use the Siamese network for four challenging sequences, including *Biker* (has the following properties:

TABLE 1. Results on VOT2016 and VOT2018 with the evaluation metrics of accuracy (A), robustness (R), and expected average overlap (EAO). Respectively, Red, blue and green denote 1st, 2nd, 3rd performance.

Tracker		SiamFC	MDNet	C-COT	SiamRPN	DaSiamRPN	ATOM	Ours
VOT2016	A↑	0.532	0.54	0.54	0.56	0.61	-	0.612
	R↓	0.461	0.34	0.24	0.26	0.22	-	0.287
	EAO↑	0.235	0.257	0.331	0.344	0.375	-	0.351
VOT2018	A↑	0.50	-	0.49	0.588	0.59	0.59	0.593
	R↓	0.59	-	0.32	0.276	0.28	0.203	0.255
	EAO↑	0.188	-	0.267	0.384	0.383	0.401	0.381

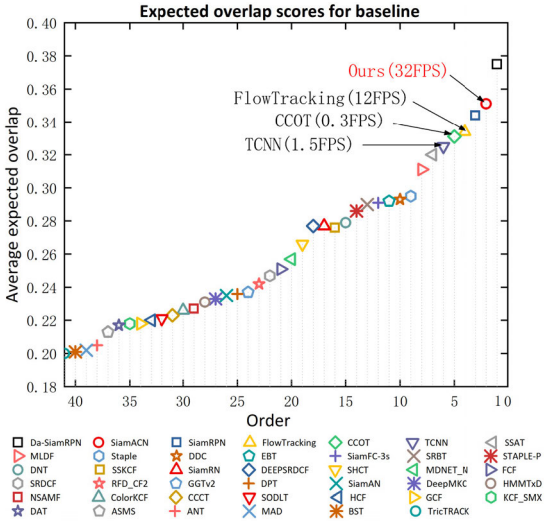


FIGURE 7. Expected averaged overlap performance on VOT2016.

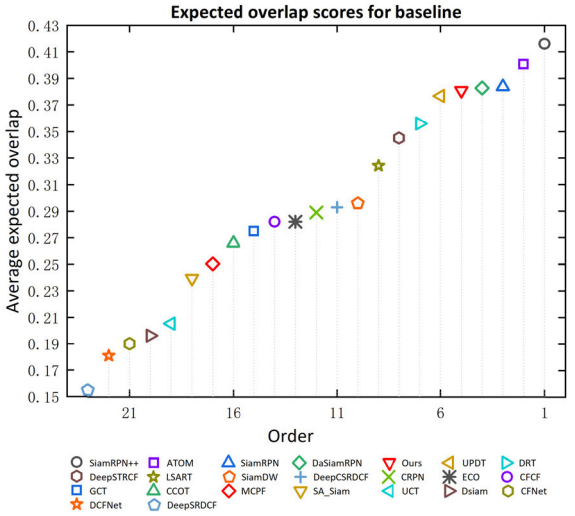


FIGURE 8. Expected averaged overlap performance on VOT2018.

LR, FM, MB, OCC, OPR, OV and SV), *Human6* (has the following properties: DEF, FM, OCC, OPR, OV, and SV), *MountainBike* (has the following properties: BC, IPR, and OPR) and *Skating1* (has the following properties: BC, DEF, IV, OCC, OPR, and SV). The results show that our approach can handle fast-moving targets as well as complex scenarios. To address fast-moving small targets, our SiamACN shows excellent performance for the sequence of *Biker*. As shown in Fig. 6 (*Human6*), the target changes by a large scale and our attentional maps are well suited to cope with the scale change. In the sequence of *MountainBike*, our approach can also solve the problem of occlusion and In-Plane rotation well benefiting from the keypoints detection. The tracking results in *Skating1* verified that by relying on our strong feature extraction backbone network, our method can distinguish blurred, dim backgrounds and many similar objects. Especially, we still track well when SiamFC and SiamRPN have already lost the target at frame #379. However, since our tracker does not update the templates online, when the background is too complex and the target undergoes numerous changes, our SiamACN can classify correctly but cannot predict the target state very well.

2) VOT

VOT2016 has approximately 60 sequences, and VOT2018 has another 10 different sequences with VOT2016. There are three commonly employed evaluation metrics in the VOT datasets: accuracy (A), robustness (R), and expected average overlap (EAO). Accuracy is used to evaluate the average overlap between the predicted bounding and the ground truth box during successful tracking periods; robustness is used to evaluate the failure rate; and EAO merges the accuracy and robustness. We evaluate our SiamACN on VOT2016 and VOT2018 with the participants in the challenges. Fig. 7 illustrates that our SiamACN ranks 2nd in 41 advanced tracking frameworks according to the EAO criterion on VOT2016. Fig. 8 shows that our SiamACN ranks 5th in 23 advanced tracking frameworks according to the EAO criterion on VOT2018. Table 1 demonstrates the detailed comparison results with some participants, including SiamFC [10], MDNET [14], C-COT [22], SiamRPN [27], Da-SiamRPN [31] and ATOM [23], on the two datasets. Da-SiamRPN achieves the top EAO with 0.375 in VOT2016, which exploits high-quality training datasets and distractor-aware modules. Our tracker achieves the best accuracy despite the second EAO and the similar robustness as SiamRPN in VOT2016.

For VOT2018, our trackers achieve not only a rank of 5th with an EAO score of 0.381 but also a similar EAO and robustness as ATOM, which uses online template updates. However, the accuracy increased by 0.3%.

3) UAV123

UAV123 is known as a new aerial video dataset, which is captured by low-altitude UAVs and contains 123 sequences. It has the characteristic that the tracked target is usually small size. By following OTB2015, we use precision and success plots to present evaluation results. We present a state-of-art

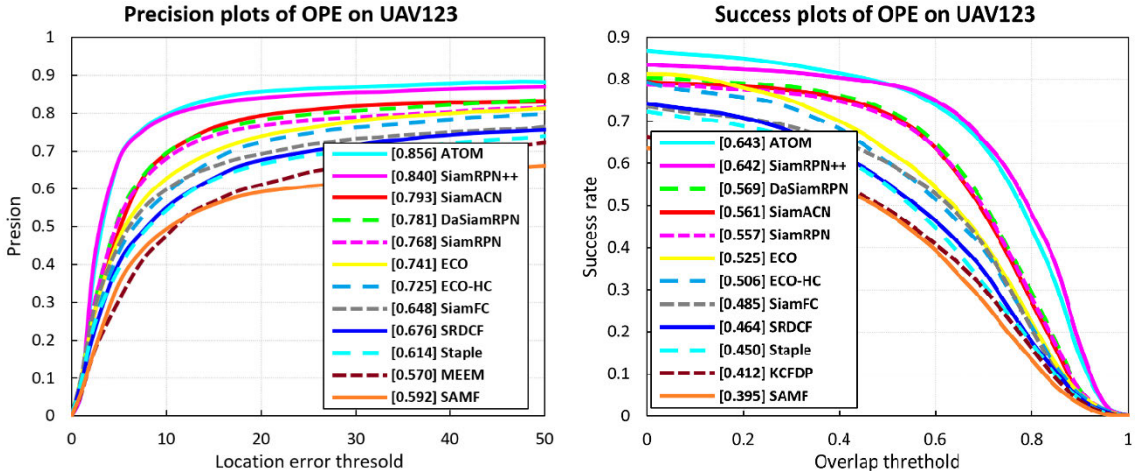


FIGURE 9. Results on UAV123 with the evaluation metrics of precision and success (AUC).

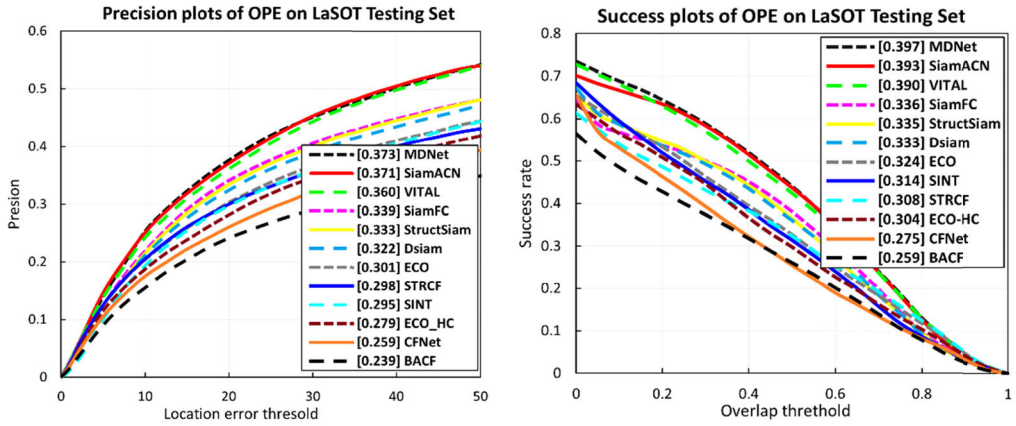


FIGURE 10. Results on LaSOT Testing Set with the evaluation metrics of precision and success (AUC).

comparison with partial advanced methods including ATOM [23], SiamRPN++ [29], Da-SiamRPN [31], SiamRPN [27], ECO [21], ECO-HC [21], SiamFC [10], SRDCF [69], Staple [66], MEEM [70] and SAMF [71]. As shown in Fig. 9, we obtain the precision of 79.3% and AUC of 56.1% higher than SiamRPN 2.5% and 0.4%, and has a comparable precision score with Da-SiamRPN.

4) LASOT

LaSOT is a large-scale dataset of 1400 high-quality sequences that have recently been frequently employed for single object tracking. A total of 280 sequences in LaSOT are combined to form a testing set. We also use the same evaluation metrics as OTB2015. We appraise our SiamACN on the testing set and compare it with numerous excellent trackers, such as MDNET [14], SiamFC [10], VITAL [72], and StructSiam [73]. Fig. 10 demonstrates the comparison success plots and normalized precision plots. Our tracker is able to achieve the second-best results, with a success rate of 37.1% in the success plot and a precision of 39.3% in the precision plot.

Unlike other tracking networks, we also qualitatively test our tracking method on the flying objects to verify the validity of our tracking framework. The tracking of flying objects often suffers from poor separation from the background, more complex scale morphology, and large inter-frame position differences. We select *drone-2* and *airplane-13* from the LaSOT testing set and compare with SiamRPN and SiamFC. The results show that our SiamACN can obtain a better performance.

As shown in Fig. 11, with the help of attention maps and corner detections, our framework can solve the more complex problem of scale estimation and out-of-plane rotation. For poor separation from the background and motion blur, our global attention hourglass module and cascade corner pooling can also help to obtain better feature maps. A low-threshold global search strategy can solve the problem of large inter-frame position differences.

C. ABLATION STUDY

1) BACKBONE ARCHITECTURE

We explore the impacts of backbone architecture and the attention maps in our tracker execute ablation study on VOT2016.

We compare the origin Houglass-54 in CornerNet-Lite [55] as the backbone and the modified Houglass-54, which add global attention to the hourglass module, and the effect of whether to add attention maps to the trackers. We train

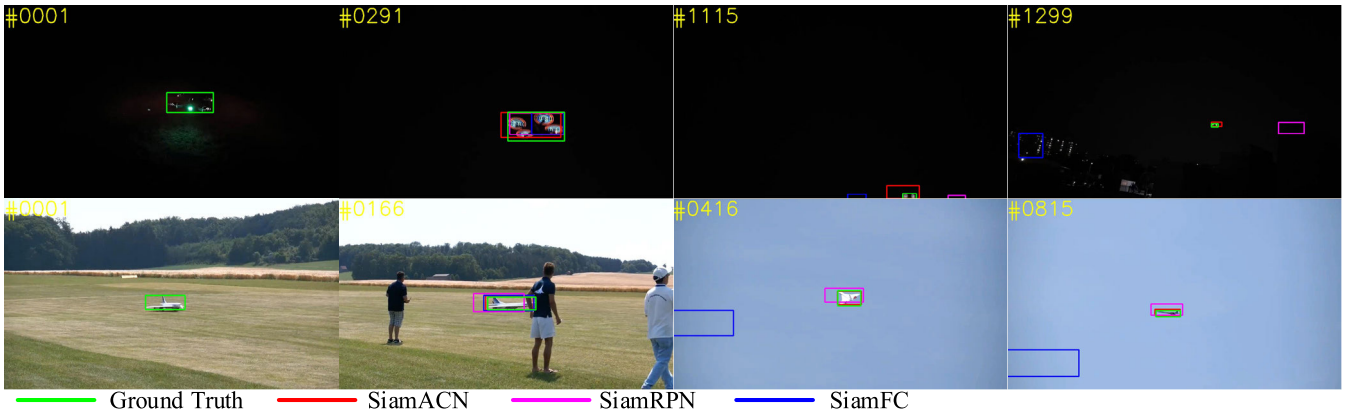


FIGURE 11. Example tracking results of our approach with other Siamese network-based trackers for two flying object sequences in LaSOT (from top to bottom: *drone-2*, *airplane-13*.).

FIGURE 12. Absolution study of corner detection with corner pooling and cascade corner pooling. (a) First frame. (b) Cascade corner pooling versus corner pooling in occlusion. (c) Cascade corner pooling versus corner pooling in scale estimation.

TABLE 2. Results performance on VOT2016 of major components: impacts of global attention (GA) Hourglass-54 as backbone network and attention maps (AM) in our tracker.

NO.	ATTENTION MAPS	BACKBONE	EAO↑	△EAO	FPS
1	×	Hourglass-54	0.336	-	36
2	×	Hourglass-54+GA	0.345	+0.006	34
3	√	Hourglass-54	0.344	+0.007	33
4	√	Hourglass-54+GA	0.351	+0.012	32

these networks with a batch size of 12 on NVIDIA Tesla V100 GPU. As shown in Table 2, our modified architecture improves the EAO by 0.6% and the FPS only decays 2 frames. Our integral tracker improves the EAO by 1.2% compared to only using hourglass-54.

2) IMPACT OF CASCADE CORNER POOLING

Corner prediction determines the accuracy of the state estimation to validate that the use of cascade corner pooling provides better access to corner information for more accurate state estimates, especially for large objects. We qualitatively analyze and compare the effects of corner pooling and cascade corner pooling on our tracker for the same backbone network. Fig. 12 shows the results for predicting corners using corner pooling and cascade corner pooling. A comparison with corner pooling indicates that the corner points of the car can be determined by using cascade corner pooling, even with a scale change.

3) ROBUSTNESS ANALYSIS

We employ attention maps to pre-detect the target in the currently tracked frame, which is very important, especially for frames that require global tracking. If the attention maps

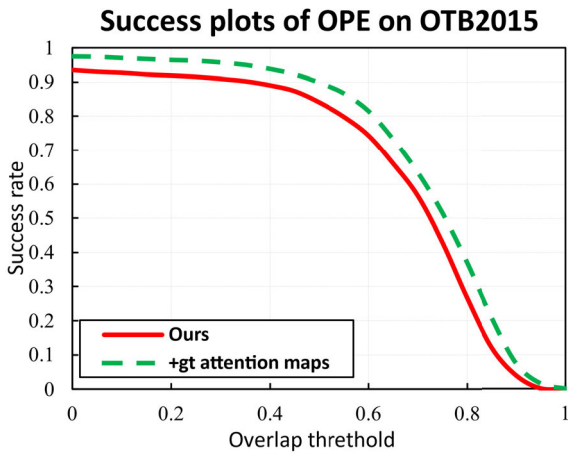


FIGURE 13. Analysis of attention maps via ground truth attention maps and predicting attention maps.

are imprecise, we could lose the target. To verify the importance of the quality of attention maps to the tracking quality, we replaced the predicted attention maps with ground truth attention maps and compared the two results by success plots on OTB2015. Fig. 13 shows that there is still improvement needed for prediction of attention maps.

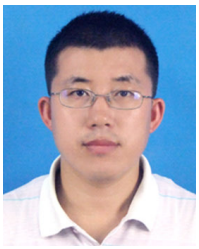
V. CONCLUSION

We proposed an effective Siamese attentional cascade keypoints network, named SiamACN, for visual object tracking. The hourglass network is employed to express the ability of the keypoints feature. By using a global attention hourglass module to integrate global and local information and contextual information and attention maps instead of a multiscale search, we further streamline our method. SiamACN directly predicts an object by using the state of a pair of corners with cascade corner pooling, and it requires no prior knowledge of the design anchors. Extensive experiments on 5 different tracking benchmarks verifies that our method achieves state-of-the-art performance and runs at 32 FPS and especially achieves the second precision of 37.1% on the LaSOT large-scale tacking dataset as well as the second precision of 79.3% on the aerial video dataset UAV123. We believe that the attention maps for pre-prediction in our framework still need improvement. Our future work will focus on lighter backbone networks to further improve the efficiency.

REFERENCES

[1] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1442–1468, Jul. 2014.
[2] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," *Neurocomputing*, vol. 74, no. 18, pp. 3823–3831, Nov. 2011.
[3] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.
[4] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3530–3538.
[5] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 FPS with deep regression networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 749–765.
[6] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.
[7] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.
[8] D. Yuan, X. Shu, and Z. He, "TRBACF: Learning temporal regularized correlation filters for high performance online visual object tracking," *J. Vis. Commun. Image Represent.*, vol. 72, Oct. 2020, Art. no. 102882.
[9] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, and S. Kasaei, "Deep learning for visual tracking: A comprehensive survey," 2019, *arXiv:1912.00535*. [Online]. Available: http://arxiv.org/abs/1912.00535
[10] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 850–865.
[11] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5000–5008.
[12] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic Siamese network for visual object tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1781–1789.
[13] X. Lu, C. Ma, B. Ni, X. Yang, I. D. Reid, and M.-H. Yang, "Deep regression tracking with shrinkage loss," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 369–386.
[14] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302.
[15] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, and M.-H. Yang, "CREST: Convolutional residual learning for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2574–2583.
[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[19] F. Battistone, V. Santopietro, and A. Petrosino, *The Visual Object Tracking VOT2016 Challenge Results* (Lecture Notes in Computer Science), 2016, pp. 777–823.
[20] M. Kristan *et al.*, "The sixth visual object tracking VOT2018 challenge results," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–53.
[21] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6931–6939.
[22] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, vol. 9909, 2016, pp. 472–488.
[23] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4655–4664.
[24] D. Yuan, W. Kang, and Z. He, "Robust visual tracking with correlation filters and metric learning," *Knowl.-Based Syst.*, vol. 195, May 2020, Art. no. 105697.
[25] D. Yuan, N. Fan, and Z. He, "Learning target-focusing convolutional regression model for visual object tracking," *Knowl.-Based Syst.*, vol. 194, Apr. 2020, Art. no. 105526.
[26] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.
[27] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with Siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.
[28] A. He, C. Luo, X. Tian, and W. Zeng, "A twofold siamese network for real-time object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4834–4843.
[29] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of Siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4282–4291.
[30] H. Fan and H. Ling, "Siamese cascaded region proposal networks for real-time visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7952–7961.
[31] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware Siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 101–117.
[32] G. Wang, C. Luo, Z. Xiong, and W. Zeng, "SPM-tracker: Series-parallel matching for real-time visual object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3643–3652.
[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[34] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.
[35] H. Fan, H. Ling, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, and C. Liao, "LaSOT: A high-quality benchmark for large-scale single object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5374–5383.
[36] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 445–461.
[37] H. Possegger, T. Mauthner, and H. Bischof, "In defense of color-based model-free tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2113–2120.
[38] L. Zhang, J. Varadarajan, P. N. Suganthan, N. Ahuja, and P. Moulin, "Robust visual tracking using oblique random forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5825–5834.
[39] 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," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4854–4863.

[40] Z. Zhang and H. Peng, "Deeper and wider siamese networks for real-time visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4591–4600.
[41] P. Gao, Q. Zhang, F. Wang, L. Xiao, H. Fujita, and Y. Zhang, "Learning reinforced attentional representation for end-to-end visual tracking," *Inf. Sci.*, vol. 517, pp. 52–67, May 2020.
[42] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1328–1338.
[43] P. Gao, Y. Ma, K. Song, C. Li, F. Wang, L. Xiao, and Y. Zhang, "High performance visual tracking with circular and structural operators," *Knowl.-Based Syst.*, vol. 161, pp. 240–253, Dec. 2018.
[44] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1561–1575, Aug. 2017.
[45] R. Liu, D. Wang, Y. Han, X. Fan, and Z. Luo, "Adaptive low-rank subspace learning with online optimization for robust visual tracking," *Neural Netw.*, vol. 88, pp. 90–104, Apr. 2017.
[46] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, "SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines," in *Proc. 34th AAAI Conf. Artif. Intell. (AAAI)*, 2020, vol. 34, no. 7, pp. 12549–12556.
[47] P. Gao, R. Yuan, F. Wang, L. Xiao, H. Fujita, and Y. Zhang, "Siamese attentional keypoint network for high performance visual tracking," *Knowl.-Based Syst.*, vol. 193, Apr. 2020, Art. no. 105448.
[48] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," *Int. J. Comput. Vis.*, vol. 128, no. 3, pp. 642–656, Mar. 2020.
[49] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[50] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "DetNet: Design backbone for object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 334–350.
[51] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[52] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.
[53] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[54] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "DenseBox: Unifying landmark localization with end to end object detection," 2015, *arXiv:1509.04874*. [Online]. Available: http://arxiv.org/abs/1509.04874
[55] H. Law, Y. Teng, O. Russakovsky, and J. Deng, "CornerNet-lite: Efficient keypoint based object detection," 2019, *arXiv:1904.08900*. [Online]. Available: http://arxiv.org/abs/1904.08900
[56] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Object detection with keypoint triplets," 2019, *arXiv:1904.08189*. [Online]. Available: https://arxiv.org/abs/1904.08189
[57] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, no. 4, pp. 669–688, 1993.
[58] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.
[59] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," in *Proc. BMVC*, 2018, p. 285.
[60] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 5562–5570.
[61] B. Steiner, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 8026–8037.
[62] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, "YouTube-BoundingBoxes: A large high-precision human-annotated data set for object detection in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7464–7473.
[63] 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," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[64] 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," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[65] R. Tao, E. Gavves, and A. W. M. Smeulders, "Siamese instance search for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1420–1429.
[66] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1401–1409.
[67] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3074–3082.
[68] G. Li, M. Peng, K. Nai, Z. Li, and K. Li, "Reliable correlation tracking via dual-memory selection model," *Inf. Sci.*, vol. 518, pp. 238–255, May 2020, doi: 10.1016/j.ins.2020.01.015.
[69] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4310–4318.
[70] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 188–203.
[71] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 254–265.
[72] Y. Song, C. Ma, X. Wu, L. Gong, L. Bao, W. Zuo, C. Shen, R. W. H. Lau, and M.-H. Yang, "VITAL: VIsual tracking via adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8990–8999.
[73] Y. Zhang, L. Wang, J. Qi, D. Wang, M. Feng, and H. Lu, "Structured Siamese network for real-time visual tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 355–370.



ERSHEN WANG is currently pursuing the Ph.D. degree with Shenyang Aerospace University.

He is also a Professor with Shenyang Aerospace University. His research interests include UAV visual tracking and global navigation satellite system (GNSS) positioning algorithms.

DONGLEI WANG received the B.S. degree in electronic and information engineering from Shenyang Aerospace University, Liaoning, China, where she is currently pursuing the degree in information and communication engineering.

She is also a Graduate School Student with Shenyang Aerospace University.

YUFENG HUANG received the B.S. degree from the Hefei University of Technology, Anhui, China, and the M.S. and Ph.D. degrees from the China University of Science and Technology, Anhui.

She is currently a Teacher with Shenyang Aerospace University. Her research interests include image enhancement and artificial intelligence.

GANG TONG is currently a Professor with Shenyang Aerospace University. His research interests include UAV design and visual tracking.

SONG XU received the M.S. degree in information and communication engineering from Shenyang Aerospace University, Liaoning, China, in 2016.

He is currently a Teacher with Shenyang Aerospace University. His research interests include visual object tracking and satellite navigation.

TAO PANG received the Ph.D. degree in control science and engineering from the Beijing University of Technology.

She is currently a Teacher with Shenyang Aerospace University. Her research interests include pattern recognition and artificial intelligence.