Visual Communications and Image Processing (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/vcip.2022.010101
ISSN 2522-6770 Vol. 1 Num. 1

Siamese Network for Fast Visual Tracking of Rotating Targets

Yibo Gao

Guidance Control and Information Perception Laboratory of High Overload Ammunition, Army Artillery and Air Defense Academy of PLA, Hefei, Anhui, 230031, China

Keywords: Oriented target tracking, feature extraction, RPN, Feature fusion

Abstract: Directional target tracking is an important task in the field of target tracking, which has great application prospects in geography, agriculture and military. The current algorithm for rotating target detection relies on the detection frame after regression and then uses the segmentation mask for further accuracy, which is obviously too cumbersome. In this paper, a scheme of direct generation of directional detection frame (Siamese-ORPN) is proposed. Specifically, improve the alternative box strategy so that the orpn can directly generate a high-quality directional detection box proposal in a low consumption manner. In addition, a top-down feature fusion network is proposed as the backbone of feature extraction and feature fusion, which can obtain substantial benefits from the diversity of visual semantic levels. Siamese-ORPN realizes lightweight and real-time detection, and achieves leading performance on benchmark data sets, including vot2018 (44.6% EAO) and vot2019 (39.6% EAO).

1. Introduction

Object tracking is a topic of fundamental research in the field of computer vision with many applications such as video vision, visual guidance, and human computer interaction. Given the position of the first frame of any object of interest in the video, the goal of the visual tracking is to estimate its position in all subsequent frames with the most possible accuracy [1]. Lighting, obstruction, deformation, and challenge to the difference of large appearance by background clutter. And, the speed is important in the practical application. Based on correlation filters [2-4] and deep learning networks [5-7], modern trackers can be roughly divided into two categories. They rely on horizontal annotations and return a simple axis lined bounding box. However, in real world applications, many objects are not horizontal. And applying an axis alignment box representing non horizontal objects may be inaccurate. The segmentation mask can describe objects more accurately than axis aligned boundary boxes [8]. Therefore, some segmentation based tasks have been proposed for directional - oriented visual tracking. Sim-mask [9] takes advantage of the strength from the segmentation data set of the video object, predicts the set of agnostic binary segmentation masks in the class and trains the shomnet to rotate the bounding box on the object. Siam-mask E [10] improves the boundary box fitting process of the mask by elliptical fitting for better rotation angle and tight border box ratio.D3S [11] has proposed an identification single object segmentation tracker that combines two object models with complementary geometric features to achieve online object

segmentation and object tracking. Although a significant advance has been made in localization accuracy, it is often time consuming by some operations such as mask generation and mask refinement. It is challenging for large appearance differences caused by lighting, occlusion, deformation, and background clutter. In addition, its speed is also important in practical applications. Based on correlation filters [2-4] and deep learning networks [5-7], modern trackers can be roughly divided into two categories. They rely on horizontal annotations and return simple axially aligned bounding boxes. However, in real-world applications, a large number of objects are not horizontal, and applying axis-aligned boxes representing non-horizontal objects may be inaccurate. Segmentation masks can describe objects more accurately than axis-aligned bounding boxes [8]. Therefore, some segmentation-based works have been proposed to address orientation-oriented visual tracking. Siam-Mask [9] leverages strengths from a video object segmentation dataset and trains a Siamese net to predict a set of class-agnostic binary segmentation masks and rotated bounding boxes on objects. Siam-Mask_E [10] improves the mask's bounding box fitting process by ellipse fitting for better rotation angles and tighter bounding box ratios. D3S [11] proposed a discriminative single-object segmentation tracker that combines two object models with complementary geometric properties to achieve online object segmentation and object tracking. Although great progress has been made in localization accuracy, these methods are often time-consuming due to several operations such as mask generation and mask refinement.

Visual tracking requires a rich representation, and feature extraction networks range to high level, low scale, and low level for coarse solutions. Architecture evolved from a shallow network to more complex networks such as RESNET or deeper networks. And it can fuse different features. Siamrpn + + [6] proposes a hierarchical feature intensive structure for cross - correlation arithmetic and predicts similarity maps from features learned at multiple levels. However, it builds rpns in multiple layers of feature maps to obtain tracking predictions that introduce many hyperparameters and computational complexity.

In this paper, we propose a simple and effective visual - oriented tracking network (orpn) based on region - based network (RPN) and feature fusion. Inspired by object-oriented detection [12-14], we can directly generate high quality oriented suggestions with direct oriented RPN, almost zero cost. Therefore, this extraction is performed on the associated feature map. Indeed, generating a generated bounding box directly is faster than fitting a masking mask. We also propose a top down feature fusion method using deconvolution and deformable convolution, and obtain high resolution and strong semantic feature maps. In summary, the main contributions of this work are as follows:

1) unlike the segmentation based method, Siamese orpn is proposed to generate the direction of high quality tracking tasks;

2) a top down feature fusion network is proposed as a backbone for predicting similarity graphs from features with different level features, thereby improving tracking accuracy.;

3) this algorithm achieves the main performance at 85 frame / s speed in VOT 2018 and VOT 2019, and proves the superiority of accuracy and efficiency.

2. Siamese-ORPN

A fully convoluted Siam network architecture [12] is adopted considering the operability and robustness of the network. Other components such as orpn and top-down feature fusion networks are built into this architecture. As shown in Figure 1.A solid box is a real box and the dash box is an estimation box.

The proposed top down feature fusion network is adopted as a backbone network for feature extraction and feature fusion. Given the target template and search area, the Siam network outputs a multichannel analogous map using depth level cross correlation [15].The directed RPN is executed

on the similarity map to extract the prediction box scheme. Oriented RPN has two branches, a regression branch that generates a proposal and a classification branch that estimates the corresponding object / background score for each indicated proposal. Large image pairs are supplied to the framework and the whole system is trained end-to-end offline.

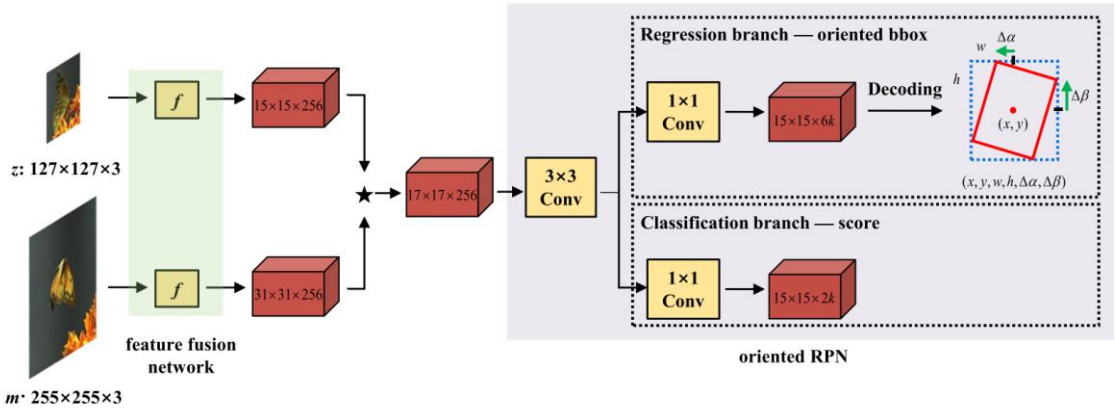

Figure 1: Main framework of Siamese-ORPN

2.1 Fully connected Siamese network with directed RPN

Considering a pair of images and templates, the target template image Z is a crop centered on an object, and the search image m is a large harvest centered on the last estimated position of the target. A fully convoluted Siam network that compares Z and m and returns a dense response map is used. These two inputs are processed by the same transformation F.:

$$g(z,m) = f(z) \otimes f(m) \tag{1}$$

Among them $\otimes$ is the correlation operator. Improved RESNET - 50 was used as f. A top down function fusion structure is introduced. The details of this network are shown in Figure 2. A deep cross correlation method is used to generate a multichannel reaction map. Following object oriented detection. The directional RPN can output a sparse set of suggestions for the input image orientation of any size. Considering its computational efficiency. A proposed target box for directional RPN is extracted on the similarity map. As shown in Figure 1, Take similarity maps $g(z,m)$ as inputs. To obtain similarity scores and bounding box coordinates $\times 3$ convolutional layers and two twins $\times$ One convolution layer was connected. Directional RPN is a lightweight complete convolution network with much fewer parameters than the segmentation based approach [14].

Each spatial location in $g(z,m)$ is assigned k horizontal anchors, so the regression branch has 6k outputs and the classification branch has 2k outputs. One of the two Siamese 1×1 convolutional layers is the regression branch, which outputs the offsets dx, dy, dw, dh, dα, dβ of the orientation proposals relative to the anchors. Here, we use (x, y, w, h, Δα, Δβ) to describe point shifts similar to [14]. Finally, the orientation offset is regressed by decoding. The decoding process is described as follows:

$$\begin{cases} x = a_w \cdot dx + a_x, y = a_h \cdot dy + a_y \\ w = a_w \cdot e^{dw}, h = a_h \cdot e^{dh} \\ \Delta\alpha = w \cdot d\alpha, \Delta\beta = h \cdot d\beta \end{cases} \tag{2}$$

Where (x, y) are the center coordinates of the estimated orientation scheme, w, h are the width and height of the outer rectangular box of the estimated orientation scheme, and Δα, Δβ are the offsets

relative to the midpoint of the top and right side of the outer rectangle Offset, (ax, ay, aw, ah) represents the anchor. Therefore, we propose orientation alternatives in terms of (x, y, w, h, Δα, Δβ). Another sibling convolutional layer is the classification branch, which estimates object/background scores for each orientation scheme, and finally, the results of orientation tracking can be inferred from the above two branches.

2.2 Top-down functional fusion network

Different layers of deep networks like RESNET make sense because the receptive fields vary widely. While the function of the latter layer is invariant to the complex appearance changes and clutter while encoding high level semantic information, the functional layer contains rich low-level visual information that is beneficial for localization. This paper proposes a feature - fusion module based on deconvolution and deformable convolution considering the output of different layers of convolutional layers to be complementary to each other.

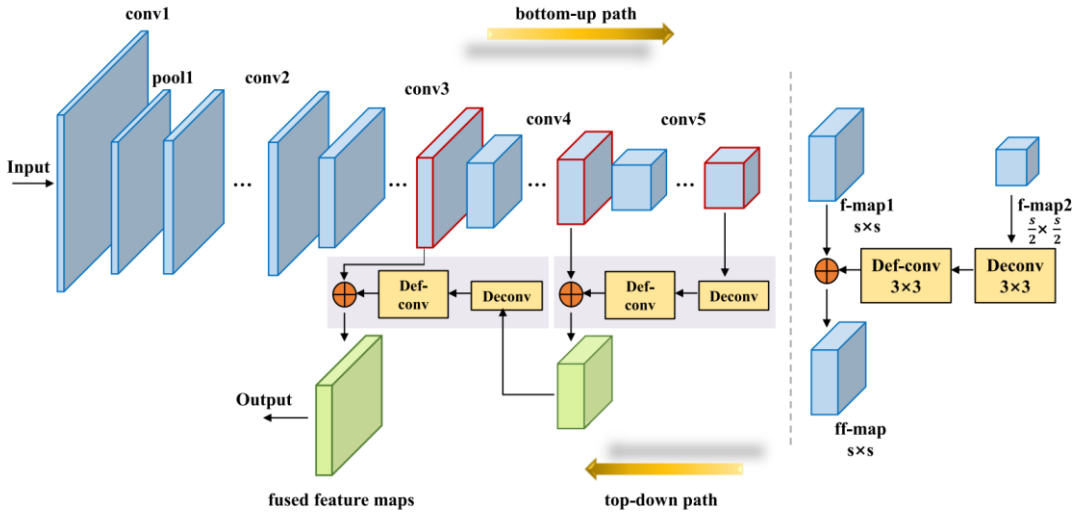

Figure 2: The feature fusion network (left) The feature fusion module (right)

In this module, first 3 × 3 deconvolution layers are employed. Secondly, deformable convolution is used to mitigate aliasing effects caused by up sampling. It has the ability to model various geometric transformations of scale, aspect ratio and rotation. Finally, the feature map obtained by deconvolution by f-map 1 and the deformable convolution are added to generate the fusion feature map FF map.

Use feature fusion modules to design backbones of Siam networks. Figure. 2 shows a top-down path of a network structure for feature fusion. In bottom-up paths, multilevel features are extracted by renet-50 for classification using features extracted from the last three remaining blocks. Our feature is fusion.In the top-down path, two fusion modules are in series. These features are then supplied to one feature fusion module, and the outputs of these modules and convn3 blocks are supplied to other feature fusion modules. Thus, a final fusion feature map is generated. The top-down fusion strategy can obtain high-resolution, informative feature maps and increase the receptive field of the last layer of the network, which is suitable for dense network prediction. By leveraging the proposed fusion network, we can gain substantial gains from visualizing the diversity of semantic hierarchies.

The top down fusion strategy yields high resolution, informative feature maps, and increases the receptive field of the last layer of the network suitable for high density network prediction. By utilizing the proposed fusion network, we can obtain significant gain by visualizing the diversity of

semantic hierarchy.

3. Experiments

3.1 Experimental details

The environment used by this tracker is portorch, GPU is 2rtx2080 Ti, and backbone network is pretrained in the imagenet1k classification task. Siamese orpn was trained using a stochastic gradient descent method (SGD) using coco [16] and Imagenet vid [17]. These training datasets are labeled with a rotating boundary box adapted by standard segmentation. A total of 20 repetitions were performed. In the first five iterations, use the warm-up strategy to set the learning rate from 0.001 to 0.005 training the indicated RPN. In the last 15 iterations, the entire network adopts training losses used in directed r-cnn [14] by exponentially decaying the learning rate from 0.005 to 0.00005. The loss classification is cross entropy loss and the regression loss is the smooth L1 loss using affine transformation. Training losses are the sum of classification loss and regression loss.

3.2 Ablation experiment

The ablation analysis of this method was performed on VOT 2018 [18] and VOT 2019 [19] data sets to verify the effectiveness of the designed network structure. Table 1 shows the effects of using different backbone networks and RPN.

Orpn: compares the proposed oriented RPN network and the original siam RPN network on the same backbone. As shown in Table 1, the Siam orpn has a large advantage in the VOT challenge (7).5% increase in VO 2018 and 6.7% increase in VOT 2019 indicating that directional boundary boxes can describe objects more accurately.

Functional fusion: To study the effects of feature fusion, six variants are trained using different hierarchical feature aggregation, which can be seen with the proposed top down feature fusion score 0.498 and 0.320eao score of vo2018 and vo2019 higher than single layer 5.5% and 5.6%.

Table 1: Ablation Experiment of Siamese-ORPN

Backbone	Conv layer	RPN	VOT 2018	VOT 2019
ResNet-50	Conv3	ORPN	0.401	0.300
	Conv4	ORPN	0.442	0.341
	Conv5	ORPN	0.392	0.296
	Conv3,Conv4	ORPN	0.421	0.318
	Conv3,Conv5	ORPN	0.412	0.313
	Conv4,Conv5	ORPN	0.457	0.352
	Conv3,Conv4,Conv5	RPN	0.423	0.328
	Conv3,Conv4,Conv5	ORPN	0.498	0.395

3.3 Comparative experiment

Since the VOT Challenge series is the only object tracking benchmark labeled with a rotating boundary box, we proposed the proposed method on VOT 2018 and VOT 2019 data sets. The two datasets contain 60 sequences with different tasks, including occlusion, lighting, motion blur, scale variation. Monitoring tracking is adopted to evaluate metrics' accuracy, vulnerability and expected average overlap. Compare different trackers. A detailed comparison is shown in Table 2.

Table 2: Comparative experiment of Siamese-ORPN

	VOT2018			VOT2019			FPS
	A↑	R↓	EAO↑	A↑	R↓	EAO↑	
SPM[20]	0.581	0.300	0.337	0.577	0.506	0.274	**120fps**
SiamFC++[21]	0.587	0.182	0.425	-	-	-	90fps
ATOM[3]	0.591	0.202	0.400	0.602	0.410	0.291	31fps
DIMP-50[22]	0.596	0.152	0.440	0.593	0.277	0.378	42fps
SiamBAN[7]	0.596	0.176	0.451	0.601	0.395	0.326	40fps
SiamRPN++[6]	0.600	0.232	0.413	0.598	0.481	0.284	36fps
STMTrack[23]	0.590	0.158	0.446	-	-	-	38fps
SiamMask[9]	0.608	0.275	0.380	0.593	0.460	0.286	54fps
SiamRN[24]	0.594	**0.130**	0.471	0.592	0.305	0.340	19fps
Siamese-ORPN	**0.611**	0.142	**0.498**	**0.604**	**0.225**	**0.395**	85fps

As can be seen in Table 2, the accuracy of the Sims RPN of this paper on vot.2018 is zero.611 vulnerabilities are 0.Numeral 142 denotes the expected average overlap amount. The accuracy of 498 and vot.2019 is 0.604, the vulnerability is 0.225 expected average overlap is 0.The overlap of 604 is set to 0.395 these scores are new records with other trackers. Although the result of siam-rpn is comparable to our algorithm, it can only run in 18 frames per second for its computational complexity.

3.4 Qualitative experiment

Algorithm and tracking result of the suboptimal algorithm on vot2019 data set of Table 2 are shown. As shown in Figure.3, red box is an algorithm of this algorithm, the yellow box is the siam-rpn algorithm, and the green box is GT. This method can obtain accurate directional boundary boxes in complex scenes.

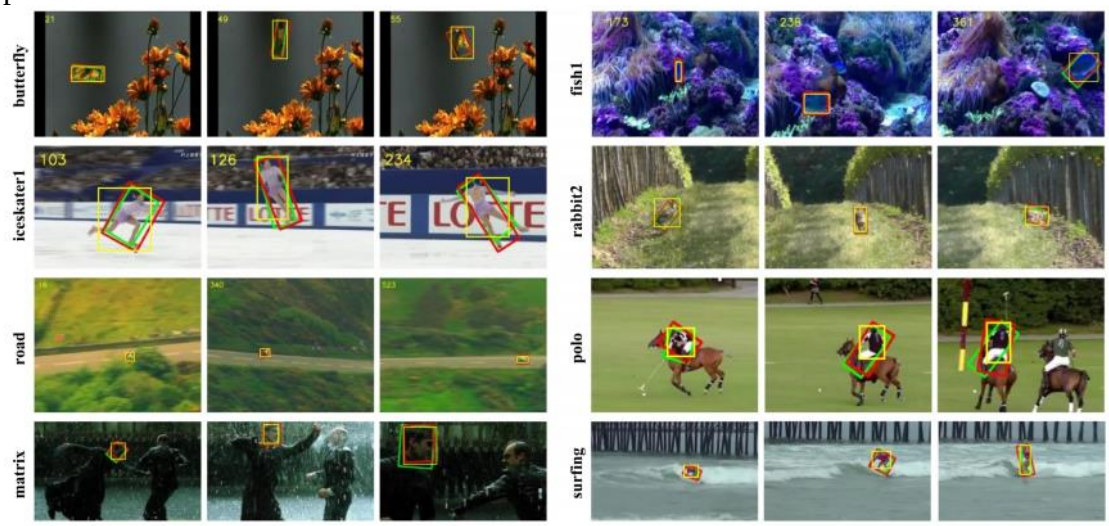

Figure 3: Simese RPN and siam-rpn experimental results

4. Conclusion

In this paper, we propose a Siam network based on directional RPN and feature fusion. Siamese - orpn can obtain a low cost and precisely oriented localized box. We also propose a feature - fusion network using deconvolution and deformable convolution as a backbone network to obtain high - precision images and improve accuracy. Extensive experiments on the tracking dataset task for two

objects labeled with a rotating boundary box are carried out. Experimental results show that this method achieves fairly high tracking efficiency while maintaining the accuracy comparable to the latest segment based tracker.

References

[1] A. W. 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] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg. (2017) ECO: Effificient convolution operators for tracking.in Proc. IEEE Conf. Comput. Vis.Pattern Recognit, pp. 6638–6646.
[3] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg. (2019) ATOM: Accurate tracking by overlap maximization.in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 4660–4669.
[4] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li. (2019) Visual tracking via adaptive spatially-regularized correlation fifilters.in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 4670–4679.
[5] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr. (2016) Fully-convolutional Siamese networks for object tracking.in Proc. Eur. Conf. Comput. Vis. Workshop, pp. 850–865.
[6] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan. (2019) SiamRPN++: Evolution of Siamese visual tracking with very deep networks.in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 4282–4291.
[7] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji. (2020) Siamese box adaptive network for visual tracking.in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 6668–6677.
[8] R. Yao, G. Lin, S. Xia, J. Zhao, and Y. Zhou. (2020) Video object segmentation and tracking: A survey. ACM Trans. Intell. Syst. Technol., vol. 11, no. 4, pp. 1–47.
[9] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr. (2019) Fast online object tracking and segmentation: A unifying approach.in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 1328–1338.
[10] B. Chen and J. K. Tsotsos. (2019) Fast visual object tracking with rotated bounding boxes, arXiv: 1907. 03892.
[11] A. Lukeži˘c, J. Matas, and M. Kristan. (2020) D3S–A discriminative single shot segmentation tracker. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 7133–7142.
[12] J. Ma et al. (2018) Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans. Multimedia, vol. 20, no. 11, pp. 3111–3122, Nov.
[13] J. Ding, N. Xue, Y. Long, G. Xia, and Q. Lu. (2019) Learning RoI transformer for oriented object detection in aerial images.in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 2849–2858.
[14] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han. (2021) Oriented R-CNN for object detection.in Proc. IEEE Int. Conf. Comput. Vis, pp. 3520–3529.
[15] L. Bertinetto, J. F. Henriques, J. Valmadre, P. H. S. Torr, and A. (2016) Vedaldi. Learning feed-forward one-shot learners, arXiv: 1606. 05233.
[16] T. Y. Lin et al. (2014) Microsoft COCO: Common objects in context.in Proc. Eur. Conf. Comput. Vis, pp. 101–115.
[17] O. Russakovsky et al. (2015) ImageNet large scale visual recognition challenge. Int. J. Comput. Vis., vol. 115, pp. 211–252.
[18] M. Kristan et al. (2018) The sixth visual object tracking VOT2018 challenge results.in Proc. Euro. Conf. Comput. Vis. Workshops, pp. 3–53.
[19] M. Kristan et al. (2019) The seventh visual object tracking VOT2019 challenge results.in Proc. IEEE/CVF Int. Conf. Comput. Vis, pp. 2206–2241.
[20] G. Wang, C. Luo, Z. Xiong, and W. Zeng. (2019) SPM-tracker: Series-parallel matching for real-time visual object tracking.in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 3643–3652.
[21] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu. (2020) SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines.in Proc. AAAI Conf. Artif. Intell, pp. 12549–12556.
[22] G. Bhat, M. Danelljan, L. V. Gool, and R. (2019) Timofte. Learning discriminative model prediction for tracking.in Proc. IEEE Int. Conf. Comput. Vis. pp. 6182–6191.
[23] Z. Fu, Q. Liu, Z. Fu, and Y. Wang. (2021) "STMtrack: Template-free visual tracking with space-time memory networks. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 13774–13783.
[24] S. Cheng et al. (2021) Learning to fifilter: Siamese relation network for robust tracking.in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 4421–4431.