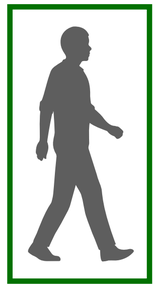


Object Tracking and SiamRPN

Computer Vision Lab, Hanyang University.

Paper Review, 25 Nov 2019.

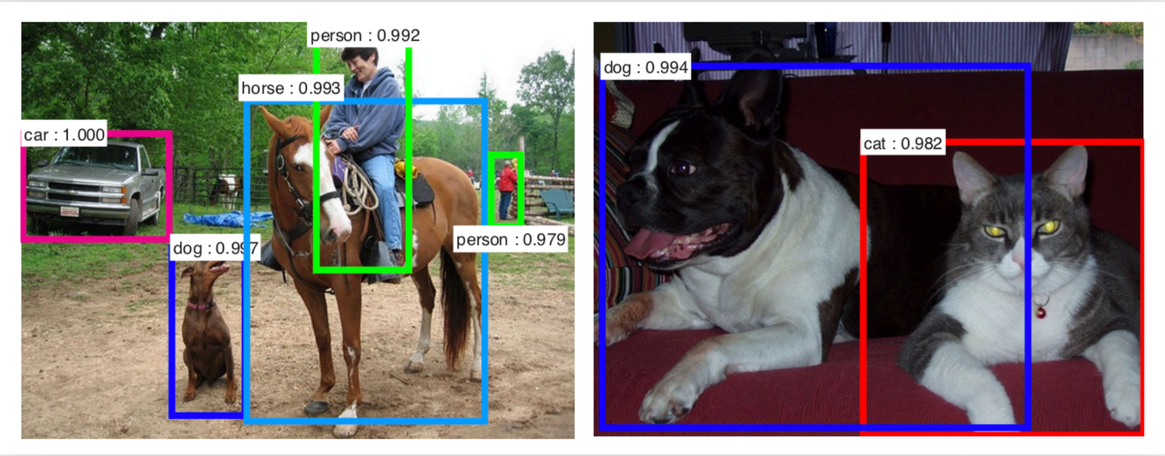
Jihun Kim

RPN

Region Proposal Network

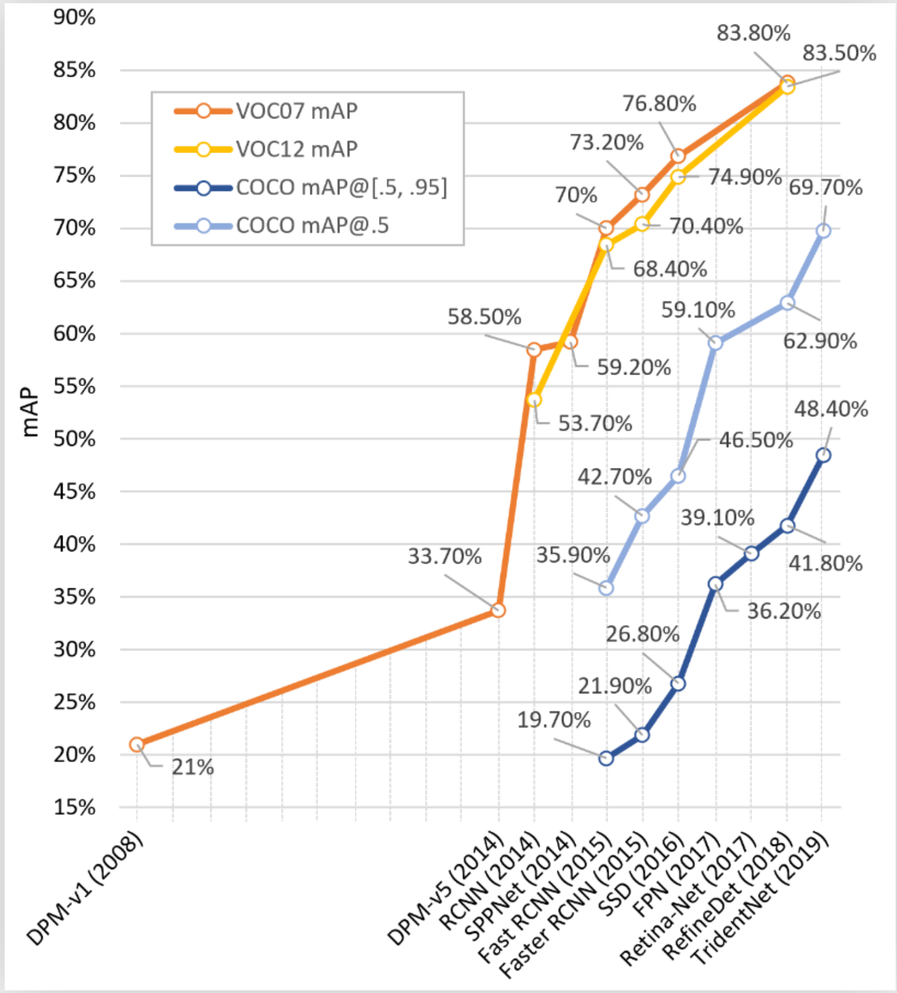
Object Detection

Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos.



Example of object detection.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.



Object detection accuracy improvments.

Zou, Z., Shi, Z., Guo, Y., & Ye, J. (2019). Object Detection in 20 Years: A Survey.

Faster R-CNN

Faster R-CNN is composed of two modules.

The first module is a deep fully convolutional network that proposes regions, and the second module is the Fast R-CNN detector that uses the proposed regions.

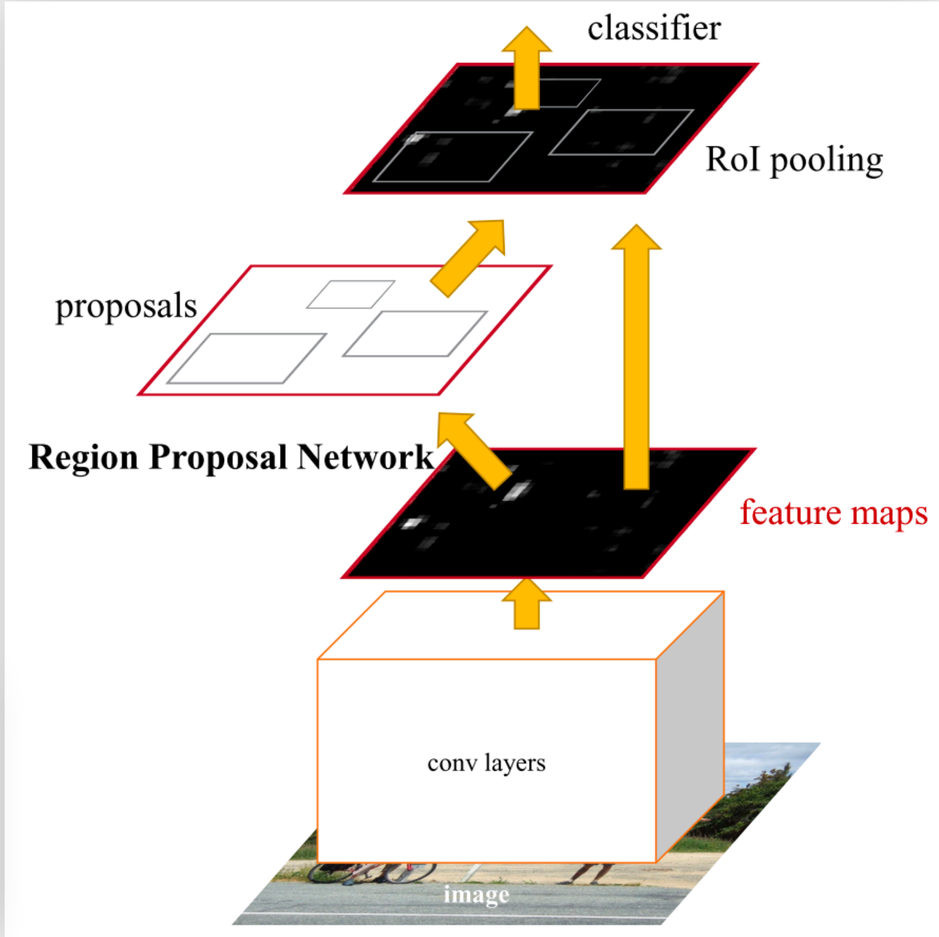
Faster r-cnn: Towards real-time object detection with region proposal networks

S Ren, K He, R Girshick, J Sun - Advances in neural information …, 2015 - papers.nips.cc

State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPPnet and **Fast R-CNN** have reduced the running time of these detection networks, exposing region proposal computation as a …

☆ 99 Cited by 13699 Related articles All 25 versions »

Google Scholar search result of Faster R-CNN.



Architecture of Faster R-CNN.

Region Proposal Network (RPN)

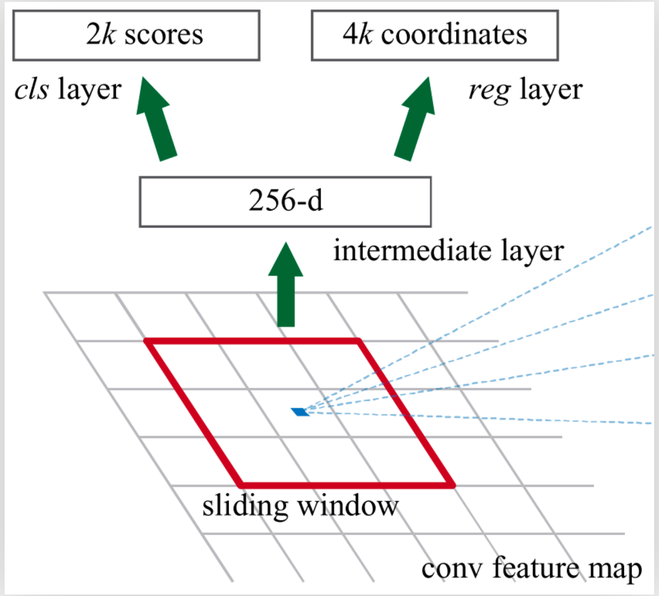
A Region Proposal Network (RPN) takes an image (of any size)
as input and outputs a set of rectangular object proposals,
each with an objectness score.

To generate region proposals,
we slide a small network over the convolutional feature map output.

This small network takes as input an $n \times n$ spatial window
of the input convolutional feature map.

Each sliding window is mapped to a lower-dimensional feature.

This feature is fed into two sibling fully-connected layers
— a box-regression layer (reg) and a box-classification layer (cls).



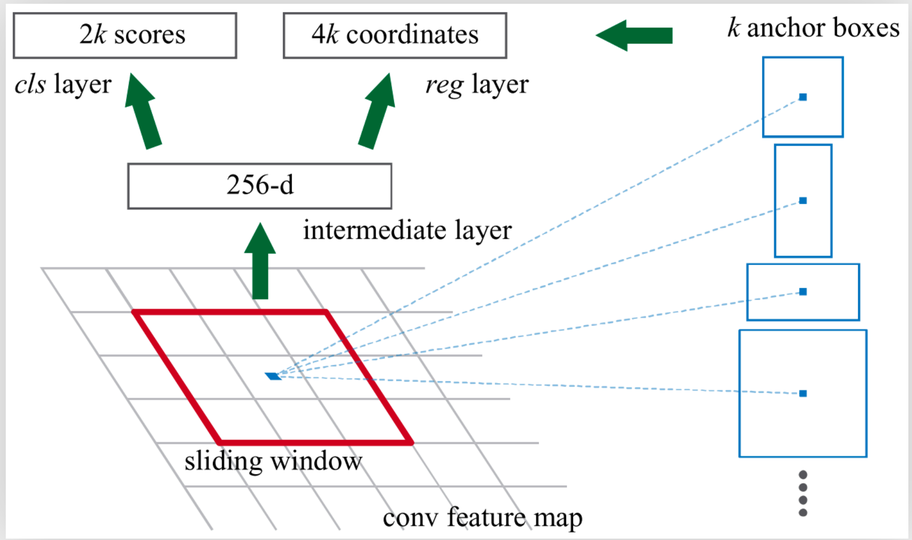
Region Proposal Network.

Anchors

At each sliding-window location, we simultaneously predict multiple region proposals,
where the number of maximum possible proposals for each location is denoted as k.

So the reg layer has $4k$ outputs encoding the coordinates of k boxes,
and the cls layer outputs $2k$ scores that estimate probability of object or not object for each proposal.

The k proposals are parameterized
relative to k reference boxes, which we call anchors.



Region Proposal Network.

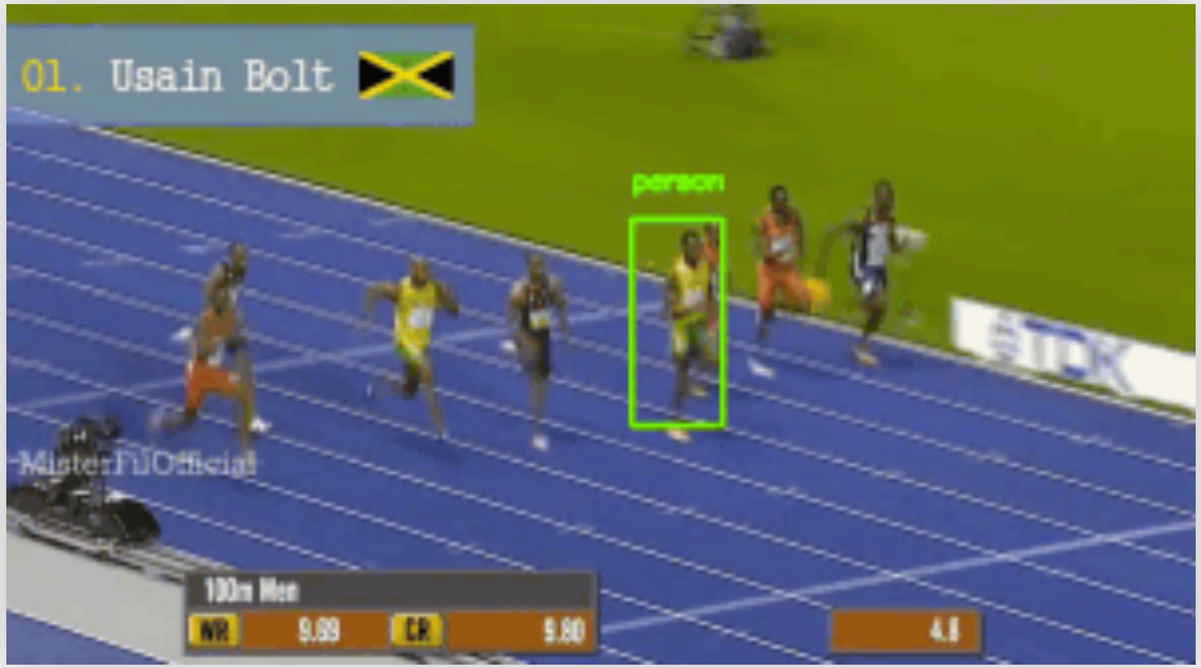
SiamRPN

Siamese Region Proposal Network

Object Tracking

Given the initialized state (e.g., position) of a target object in the first frame of one video, the goal of tracking is to estimate the states of the target in the following frames.

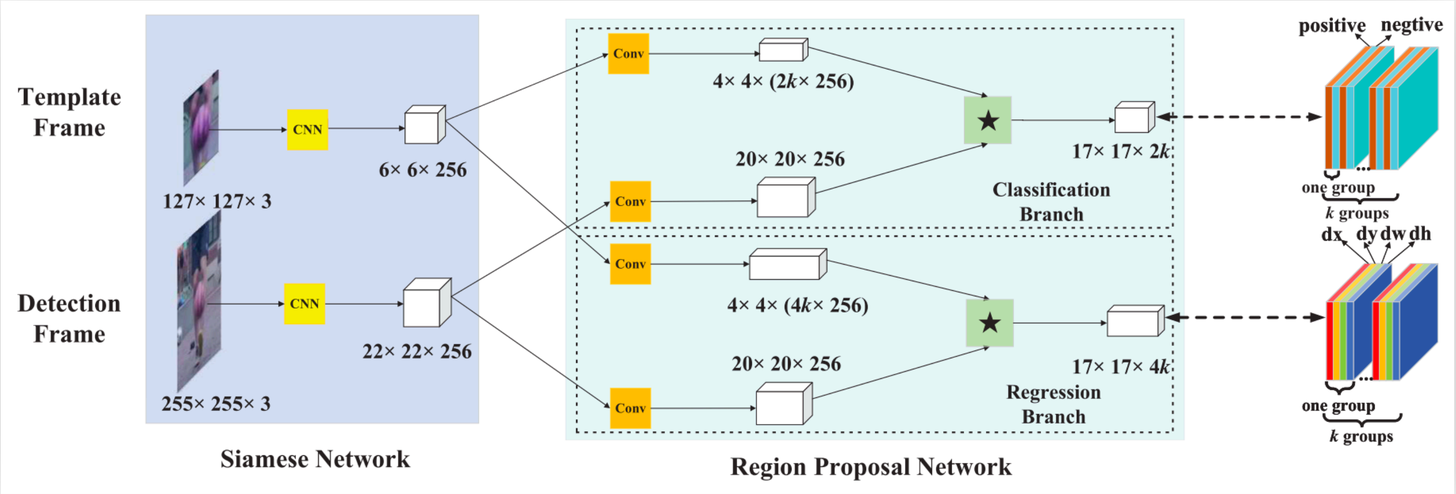
— Wu, Y., Lim, J., & Yang, M. H. (2013). Online object tracking: A benchmark.



Siamese-RPN Framework

The proposed framework consists of a Siamese subnetwork for feature extraction and a region proposal subnetwork for proposal generation.

Specifically, there are two branches in RPN subnetwork,
one is in charge of the foreground-background classification,
another is used for proposal refinement.



Main framework of SiamRPN.

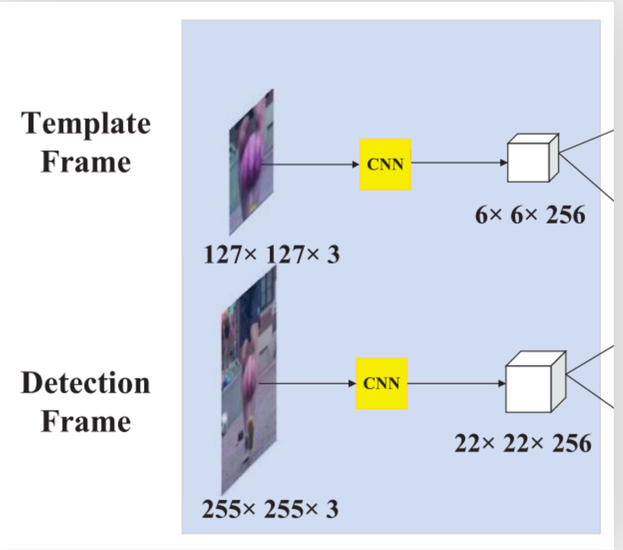
Siamese Feature Extraction Subnetwork

The Siamese feature extraction subnetwork consists of two branches.

One is called the template branch which receives target patch
in the historical frame as input (denoted as z).

The other is called the detection branch which receives target patch
in the current frame as input (denoted as x).

The two branches share parameters in CNN so that
the two patches are implicitly encoded by the same transformation.



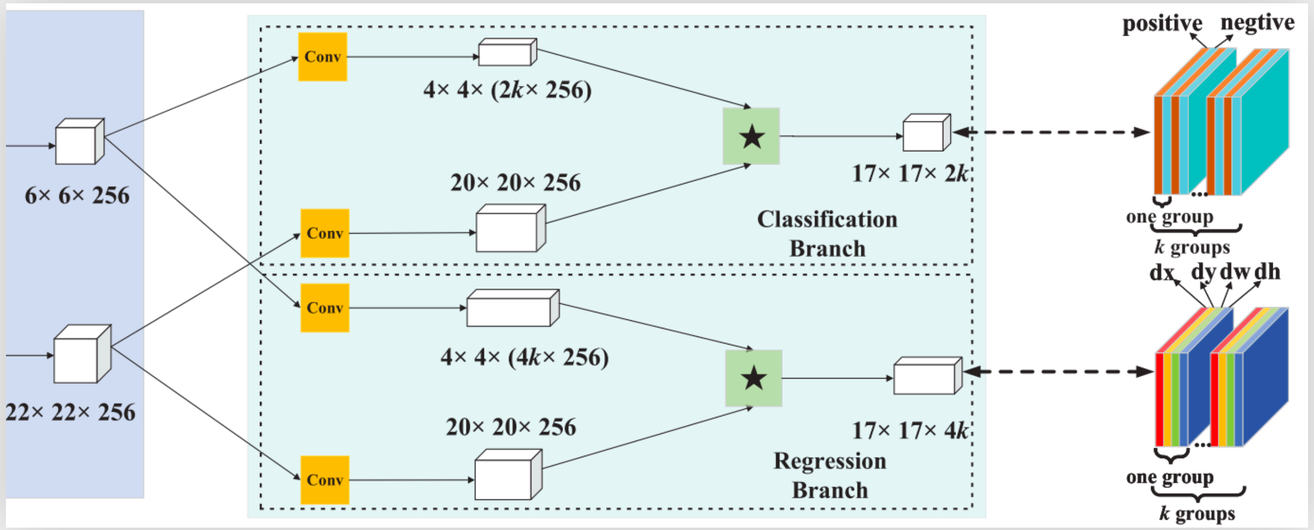
Siamese feature extraction subnetwork.

We denote $\phi(z)$ and $\phi(x)$ as the output feature maps of Siamese subnetwork.

Region Proposal Subnetwork

The region proposal subnetwork consists of a pair-wise correlation section and a supervision section.

The supervision section has two branches,
one for foreground-background classification and the other for proposal regression.



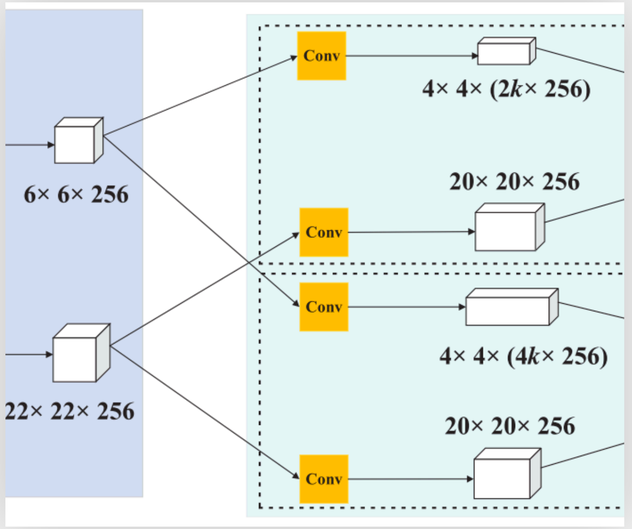
Region proposal subnetwork.

Supervision Section

If there are k anchors,
network needs to output $2k$ channels for classification and $4k$ channels for regression.

So the pair-wise correlation section first increase the channels of $\phi(z)$ to two branches $[\phi(z)]_{cls}$ and $[\phi(z)]_{reg}$ which have $2k$ and $4k$ times in channel respectively by two convolution layers.

$\phi(x)$ is also split into two branches $[\phi(x)]_{cls}$ and $[\phi(x)]_{reg}$
by two convolution layers but keeping the channels unchanged.

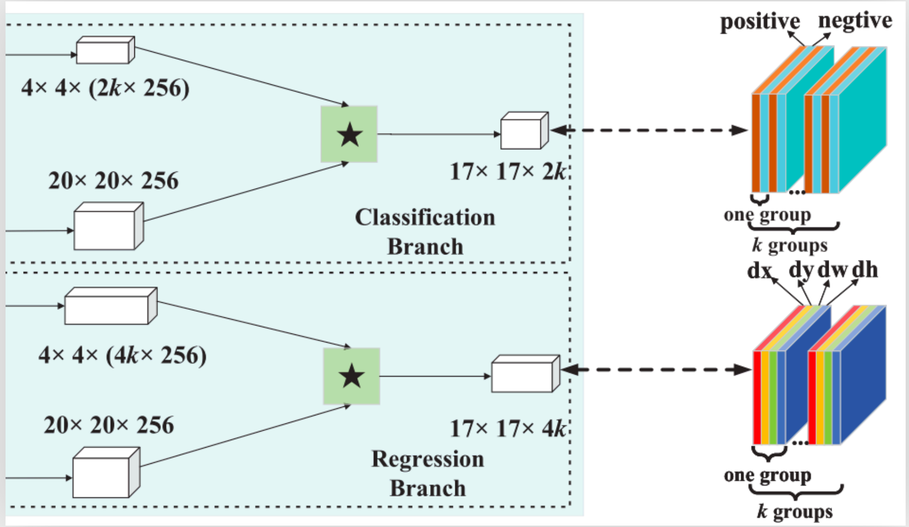


Supervision section of region proposal subnetwork.

Correlation Section

$[\phi(z)]$ is served as the correlation kernel of $[\phi(x)]$ in a “group” manner, that is to say,
the channel number in a group of $[\phi(z)]$ is the same as the overall channel number of $[\phi(x)]$.

The template feature maps $[\phi(z)]_{cls}$ and $[\phi(z)]_{reg}$ are used as kernels
and $\star$ denotes the convolution operation.



Correlation section of region proposal subnetwork.

Result on VOT2015

The VOT2015 dataset consists of 60 sequences.

The performance is evaluated in terms of accuracy (average overlap while tracking successfully) and robustness (failure times).

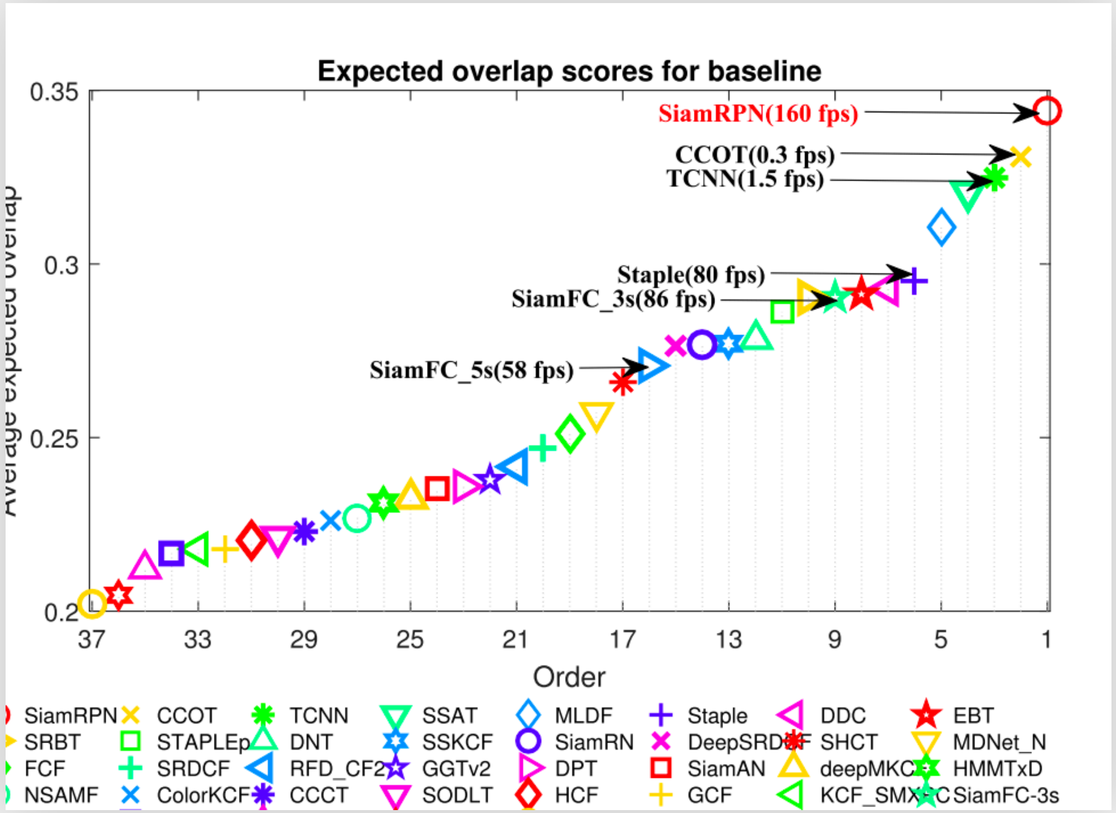
The overall performance is evaluated using Expected Average Overlap (EAO) which takes account of both accuracy.

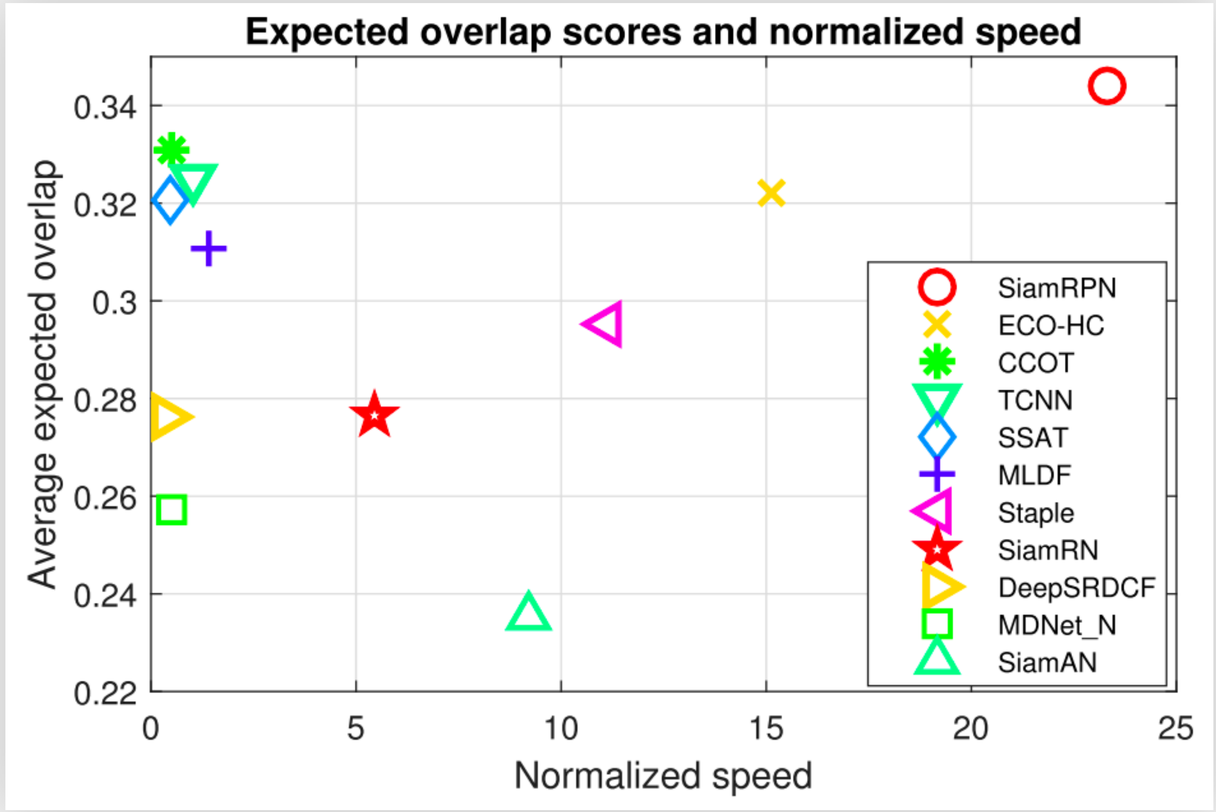
Besides, the speed is evaluated with a normalized speed (EFO).

Tracker	EAO	Accuracy	Failure	EFO
DeepSRDCF	*0.3181*	*0.56*	*1.0*	0.38
EBT	*0.313*	0.45	*1.02*	1.76
SRDCF	0.2877	0.55	1.18	1.99
LDP	0.2785	0.49	1.3	4.36
sPST	0.2767	0.54	1.42	1.01
SC-EBT	0.2548	0.54	1.72	0.8
NSAMF	0.2536	0.53	1.29	5.47
Struck	0.2458	0.46	1.5	2.44
RAJSSC	0.242	*0.57*	1.75	2.12
S3Tracker	0.2403	0.52	1.67	*14.27*
SiamFC-3s	0.2915	0.54	1.42	*8.68*
SiamFC-5s	0.275	0.53	1.45	7.84
SiamRPN	***0.358***	***0.58***	***0.93***	***23.0***

Results >

Result on VOT2016





References

Faster R-CNN

Ren, Shaoqing, et al. **"Faster r-cnn: Towards real-time object detection with region proposal networks."** Advances in neural information processing systems. 2015.

SiamRPN

Li, Bo, et al. **"High performance visual tracking with siamese region proposal network."** Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.