Deep Learning for Computer Vision. Paper Review

Serhii Tiutiunnyk

August 2019

1 Paper

- **Title:** Siamese Cascaded Region Proposal Networks for Real-Time Visual Tracking
- **Authors:** Heng Fan, Haibin Ling
- **Link:** http://openaccess.thecvf.com/content_CVPR$_2$019/*papers/Fan_Siamese_Cascaded_Region_Prop Time_Visual_Tracking_CVPR_*2019*_paper.pdf*
- **Tags:** Region Proposal Networks, Siamese network, Real-Time Visual Tracking, Feature Transfer Block
- **Year:** 2019

2 Summary

- **What:**
  - They introduce a novel multi-stage framework C-RPN for tracking, the Siamese Cascaded RPN (C-RPN), to solve the problem of class imbalance. C-RPN demonstrates more robust performance in handling complex background such as similar semantic distractors by performing hard negative sampling within a cascade architecture.
  - Use a feature transfer block (FTB) which enables to fuse the high-level features into low-level RPN, which further improves its discriminative power to deal with complex background.
  - C-RPN refines the target bounding box step-by-step using cascade of regressions, leading to more accurate localization.
- **How:**
  - Basic method

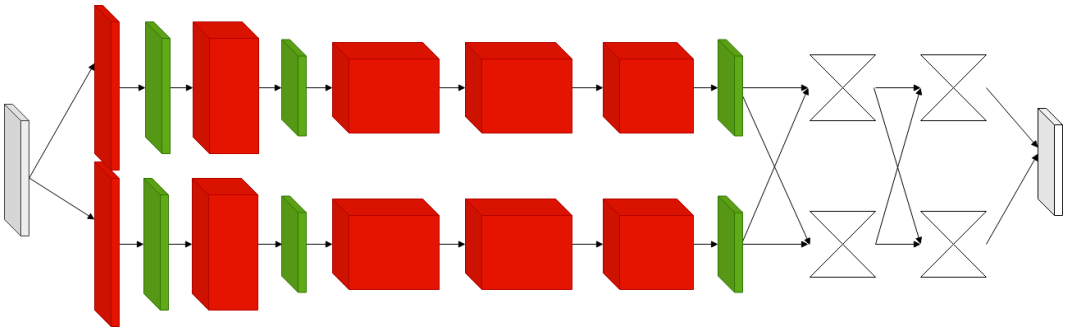

Figure 1: AlexNet Architecture.

  - Their approach cascades a sequence of RPNs to address the data imbalance by performing hard negative sampling.
  - Th eapproach progressively refines anchor boxes for better target localization using multi-regression.
  - Using multi-level features for tracking.
  - They propose a feature transfer block to fuse the features across layers for each RPN
- Architecture
  - C-RPN contains two subnetworks: Siamese network and the cascaded RPN.
  - As a Siamese network branch they adopted the modified AlexNet Figure [1]. The Siamese network comprises two identical branches, the z-branch and the x-branch, which are employed to extract features from z and x, respectively.
  - One-Stage RPN scheme is displayed on the Figure [2]. RPN consists of two branches of classification and regression for anchors. It takes as inputs the feature transformations $\phi(z)$ and $\phi(x)$ of z and x and outputs classification scores and regression offsets for anchors.
  - Feature Transfer Block is displayed in the Figure [3]. It leverages multi-level features.
  - The general architecture of C-RPN is shown in Figure [4]. It includes a Siamese network [1] for feature extraction, cascaded regional proposal networks [2] for sequential classifications and regressions and feature transfer blocks [3].
- They trained their network on a single Nvidia GTX 1080 with 8GB memory. Instead of training AlexNet they took model pretrained on ImageNet. During training, the parameters of first two layers are frozen. CRPN is trained end-to-end over 50 epochs using SGD and the learning rate is annealed geometrically at each epoch from 10^2 to

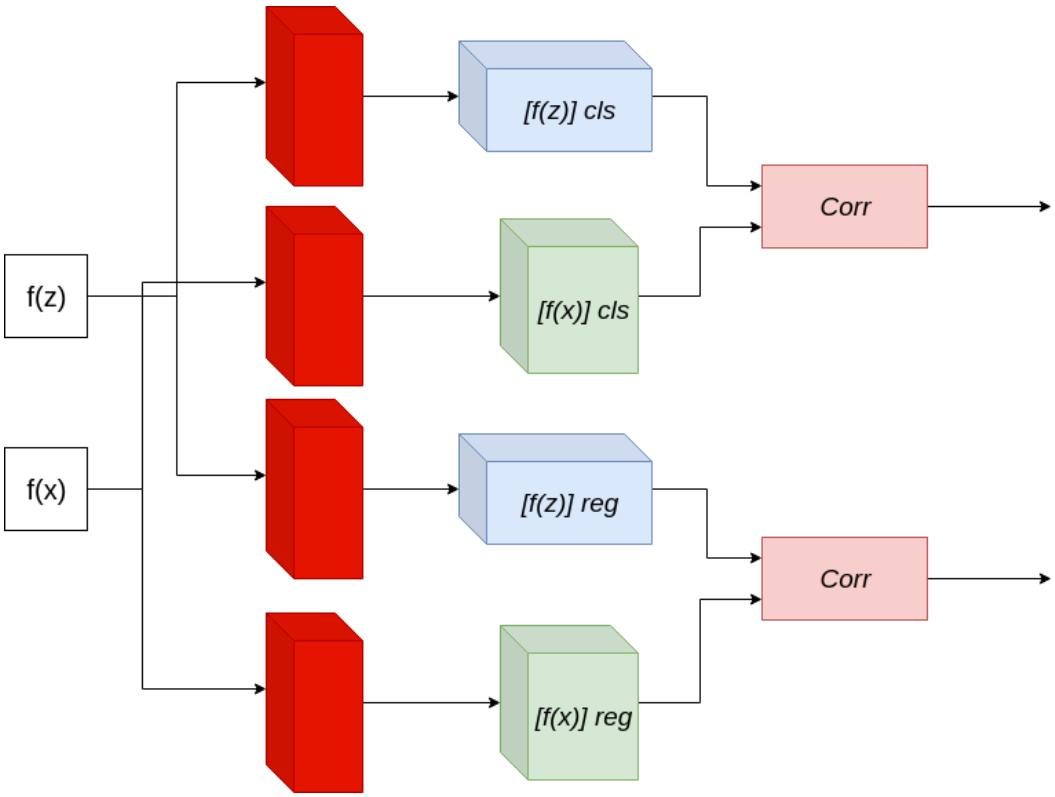

Figure 2: Architecture of RPN.

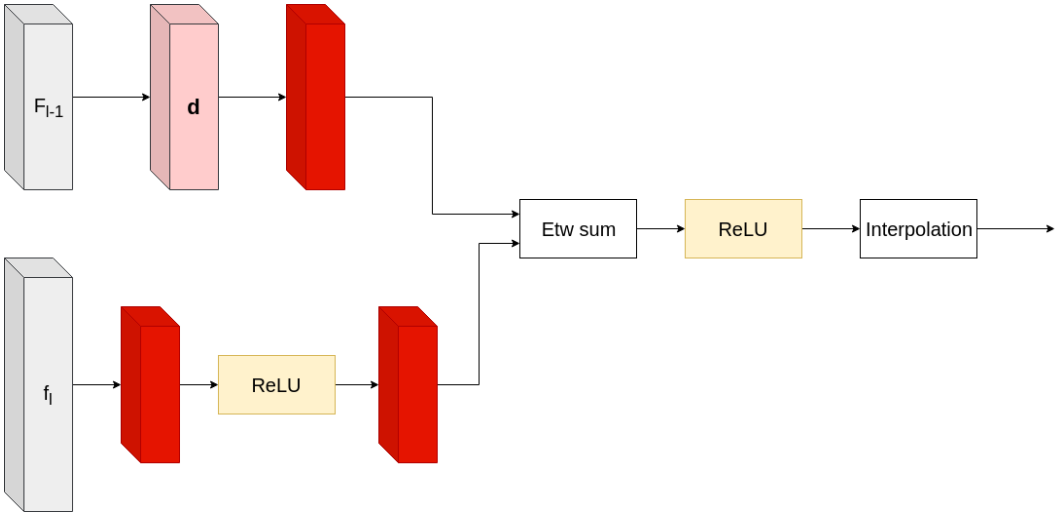

Figure 3: Architecture of feature transfer block.

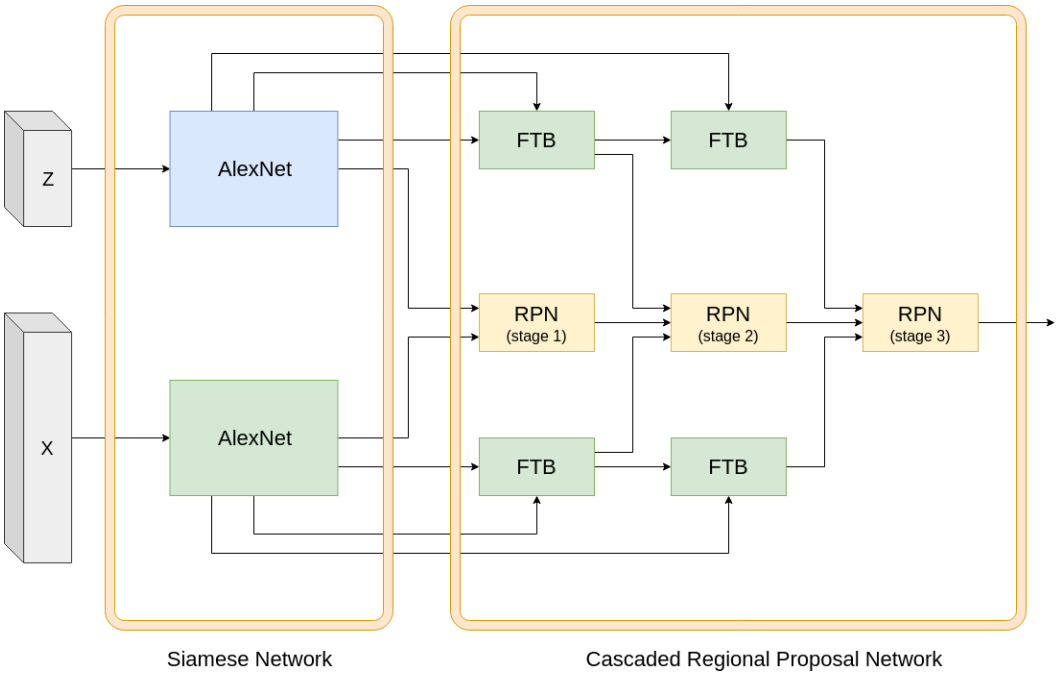

Figure 4: Illustration of the architecture of C-RPN.

10^6. They trained C-RPN using the training data from LaSOT for experiments and using VID and YT-BB for other experiments.

- **Results:**
  - They conducted experiments on the popular OTB-2013 and OTB-2015 which consist of 51 and 100 fully annotated videos, respectively. C-RPN runs at around 36 fps. They compared it with 15 state-of-the-art trackers (Figure [5]).
  - They evaluated C-RPN on VOT-2016, and compare it with 11 trackers including the baseline SiamRPN and other top ten approaches in VOT-2016 (Figure [7]). C-RPN significantly outperforms the baseline and other approaches. There is also detailed comparisons on VOT-2016 (Figure [8]). The best two results are highlighted.
  - The report of the results of success (SUC) for different state-of-the-art trackers is shown in Figure [6]. It shows that C-RPN outperforms all other trackers under two protocols.
  - And the last comparison is shown in Figure [9]. Comparisons on TrackingNet with the best two results highlighted. We can see here that C-RPN shows the best results on this test set.

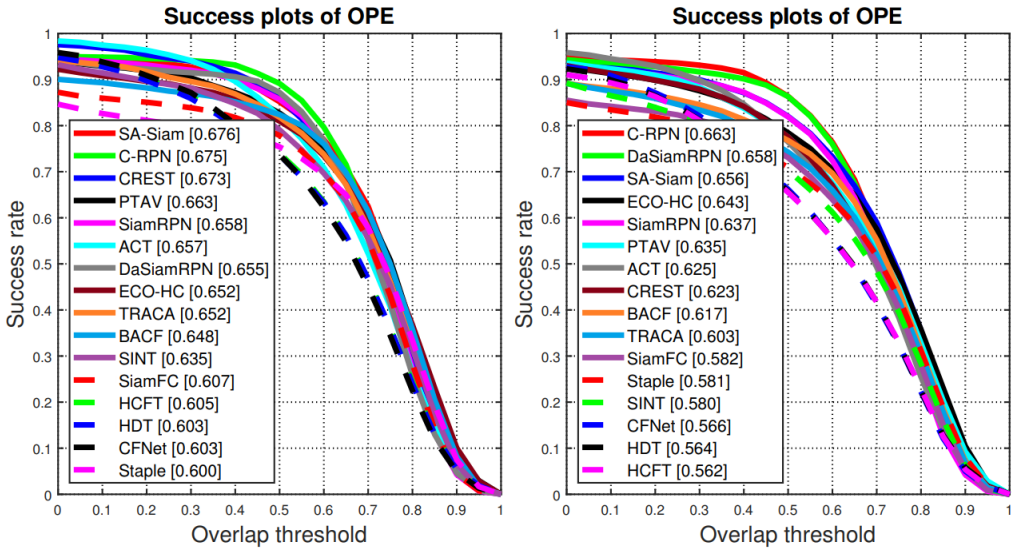

Figure 5: Comparisons with stage-of-the-art tracking approaches.

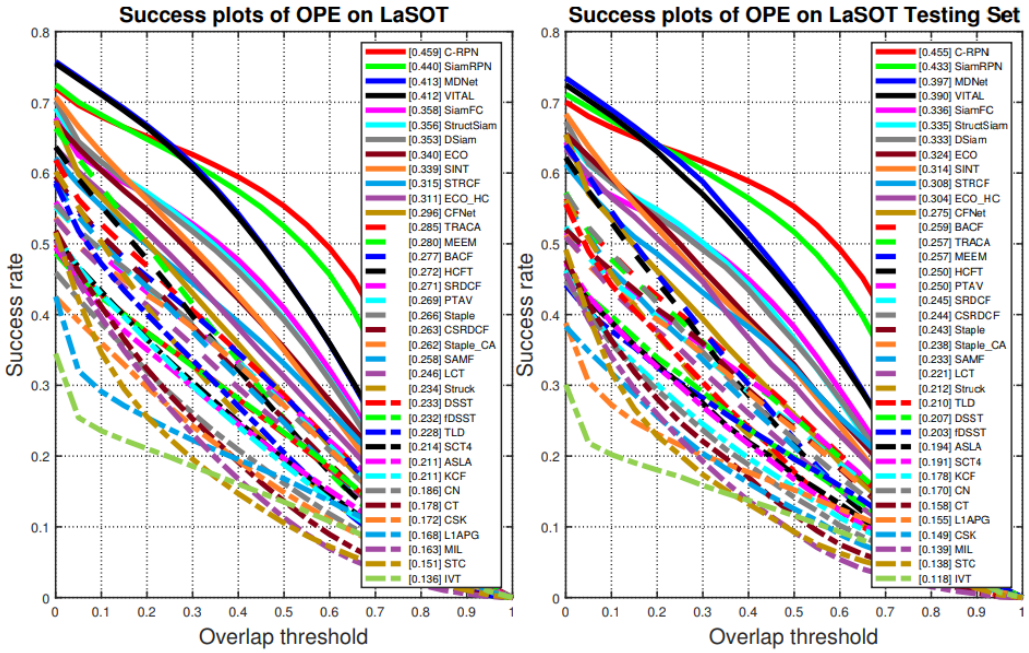

Figure 6: Comparisons with state-of-the-art tracking methods on LaSOT.

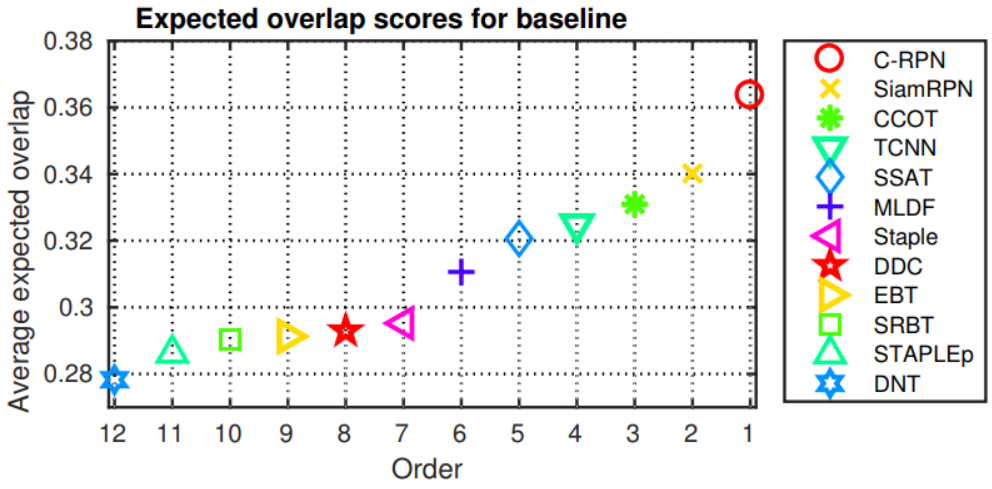

Figure 7: Comparisons on VOT-2016.

Tracker	EAO	Accuracy	Failure	EFO
C-RPN	**0.363**	**0.594**	0.95	9.3
SiamRPN [23]	**0.344**	0.560	1.12	**23.0**
C-COT [8]	0.331	0.539	**0.85**	0.5
TCNN [20]	0.325	0.554	0.96	1.1
SSAT [20]	0.321	**0.577**	1.04	0.5
MLDF [20]	0.311	0.490	**0.83**	1.2
Staple [1]	0.295	0.544	1.35	11.1
DDC [20]	0.293	0.541	1.23	0.2
EBT [58]	0.291	0.465	0.90	3.0
SRBT [20]	0.290	0.496	1.25	3.7
STAPLEp [20]	0.286	0.557	1.32	**44.8**
DNT [5]	0.278	0.515	1.18	1.1

Figure 8: Detailed comparisons on VOT-2016.

	PRE	NPRE	SUC
C-RPN	**0.619**	**0.746**	**0.669**
MDNet [35]	**0.565**	**0.705**	**0.606**
CFNet [46]	0.533	0.654	0.578
SiamFC [2]	0.533	0.663	0.571
ECO [7]	0.492	0.618	0.554
CSRDCF [31]	0.48	0.622	0.534
SAMF [26]	0.477	0.598	0.504
ECO-HC [7]	0.476	0.608	0.541
Staple [1]	0.470	0.603	0.528
Staple_CA [33]	0.468	0.605	0.529
BACF [13]	0.461	0.580	0.523

Figure 9: Comparisons on TrackingNet.