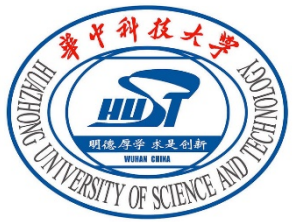
Proposal Tracking and Segmentation (PTS): A cascaded network for video object segmentation

Zilong Huang*, Qiang Zhou*, Lichao Huang, Han Shen, Yongchao Gong,
Chang Huang, Wenyu Liu, Xinggang Wang

speeding_zZ team

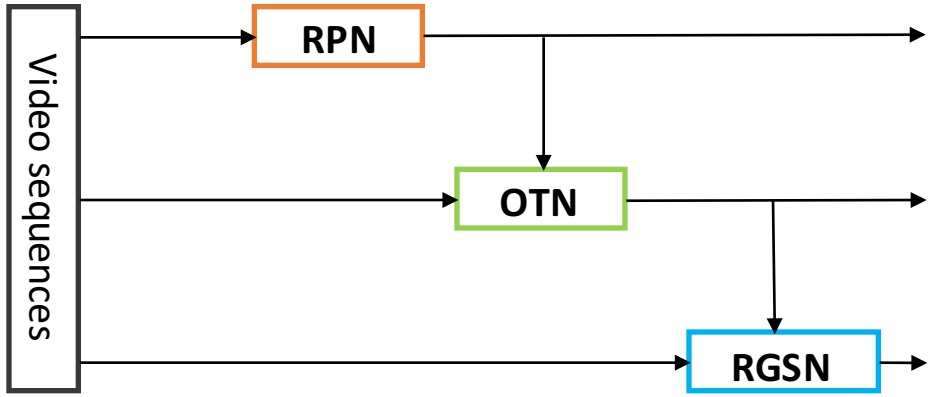
Huazhong University of Science and Technology (HUST) & Horizon Robotics

*equal contribution & interns of Horizon Robotics



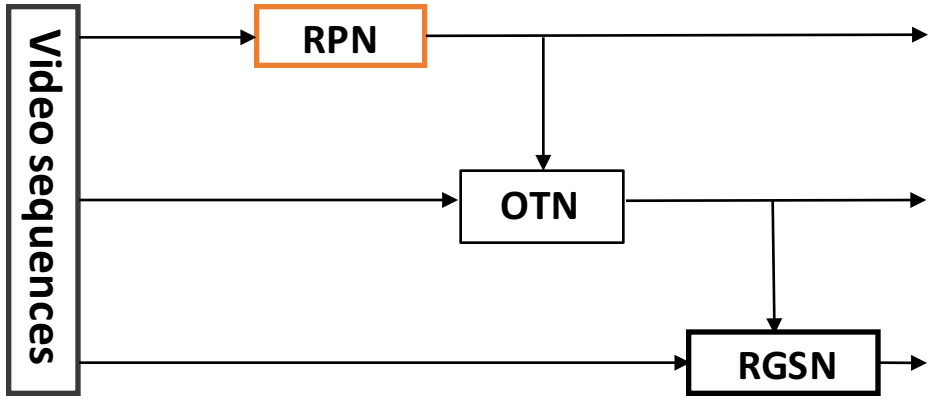
Horizon Robotics

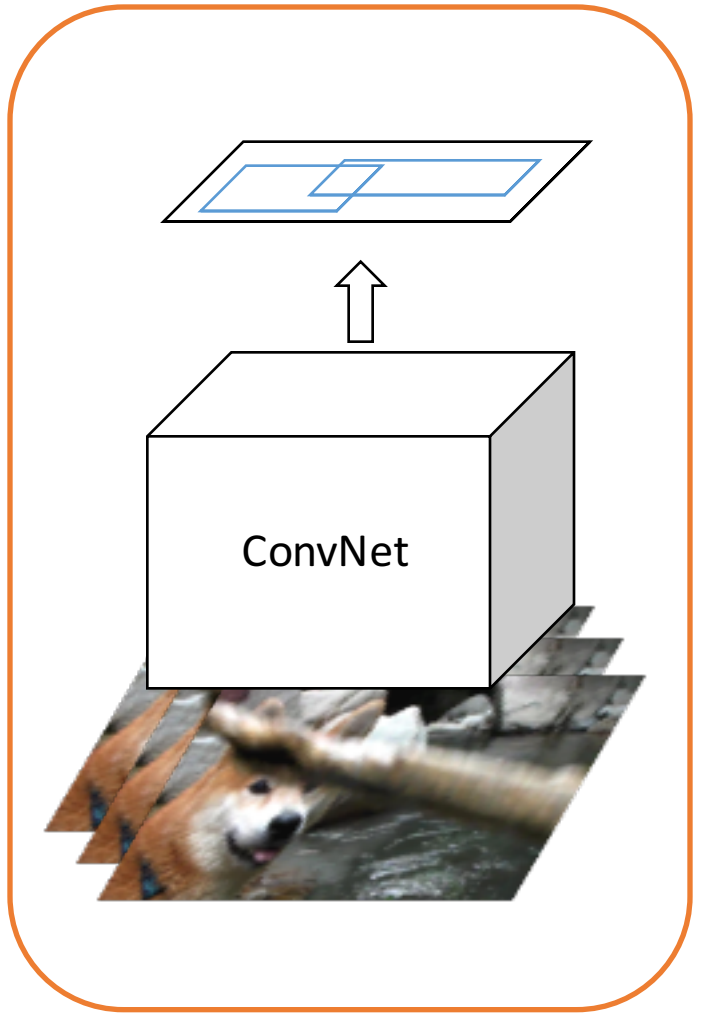
PTS: A cascaded network for video object segmentation



RPN: Region Proposal Network (2000 boxes)
OTN: Object Tracking Network (1 box)
RGSN: Reference-Guided Segmentation Network

RPN: Region Proposal Network

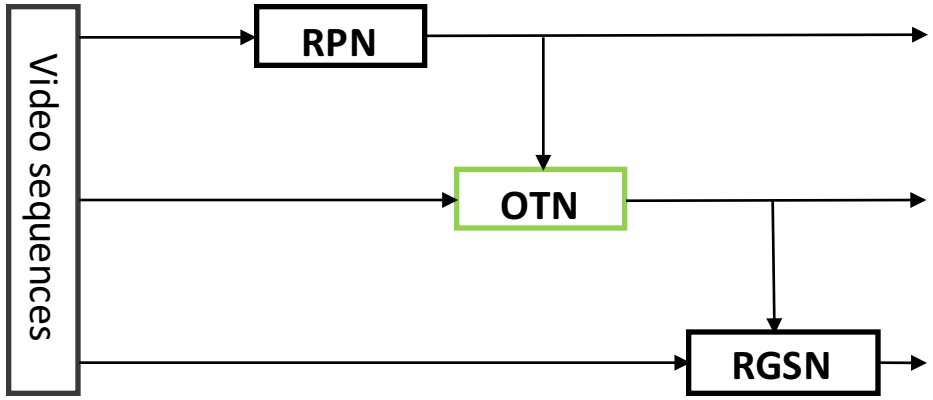


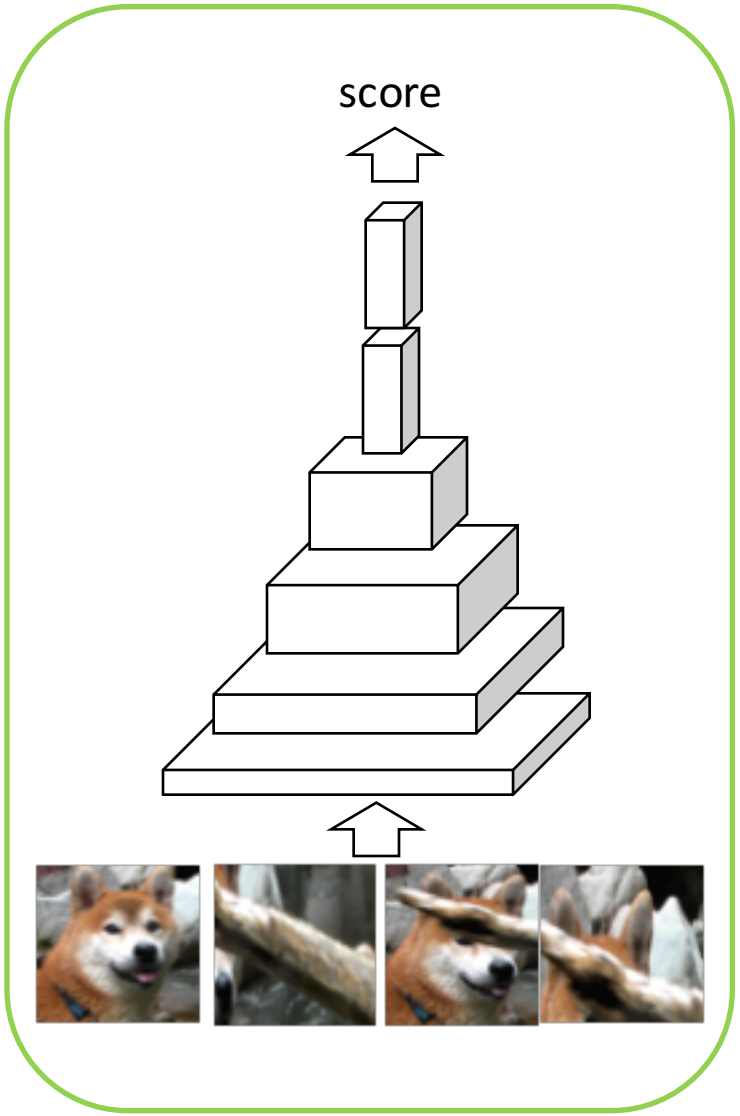


The Region Proposal Network is pre-trained on COCO and provides class-agnostic object candidate boxes. RPN could encode the instance(object) information into framework.

Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015.

OTN: Object Tracking Network





Inspired by MDNet, Object Tracking Network is designed to score the candidate boxes and updated online for adapting to large and fast changes in object appearance.

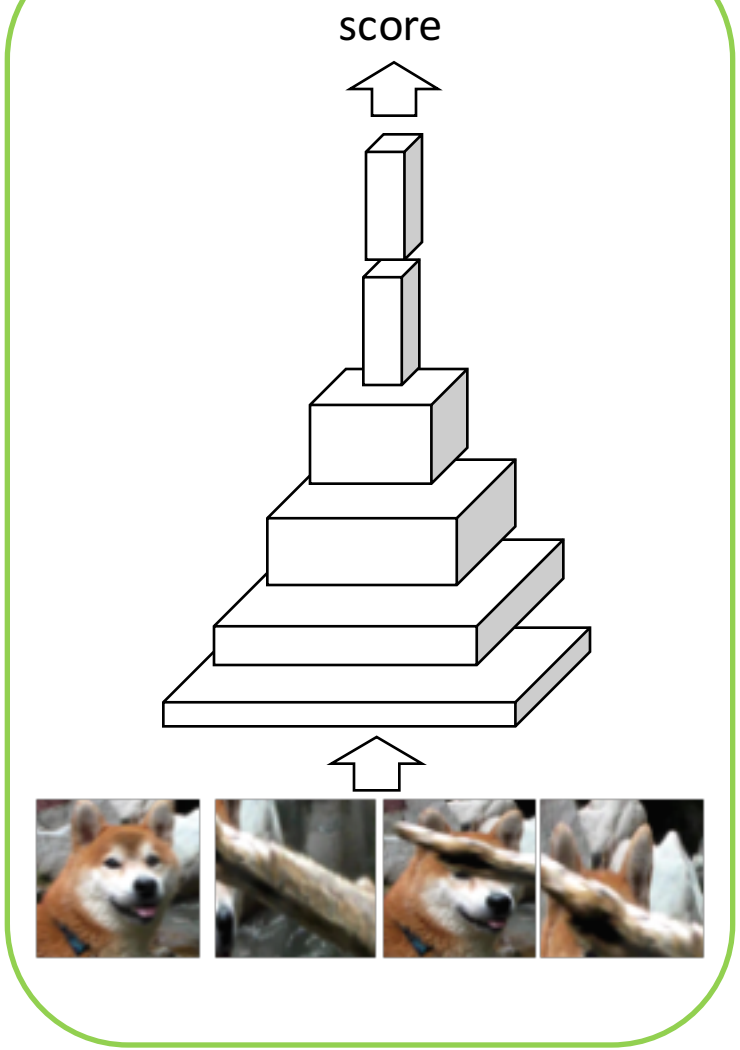
Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016.

Online Object Tracking Network

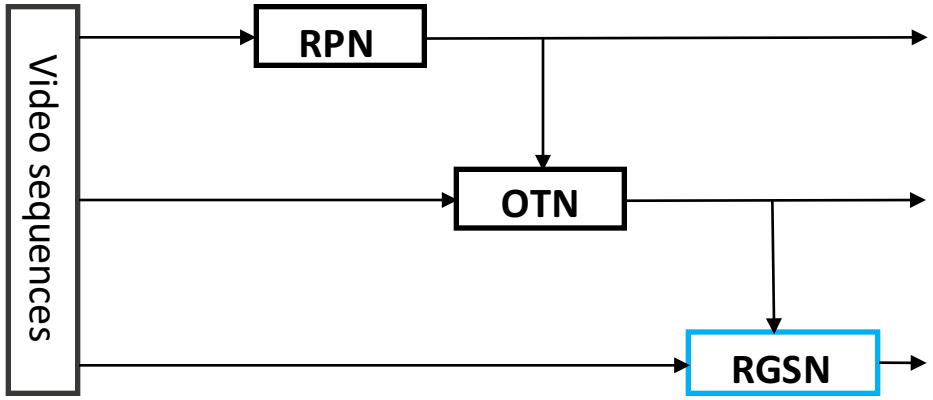
- **Long-term updates** are performed in regular intervals using the positive samples collected for a long period
- **short-term updates** are conducted whenever potential tracking failures are detected—when the score of the estimated target is less than 0.5 — using all the positive samples in the short-term period.

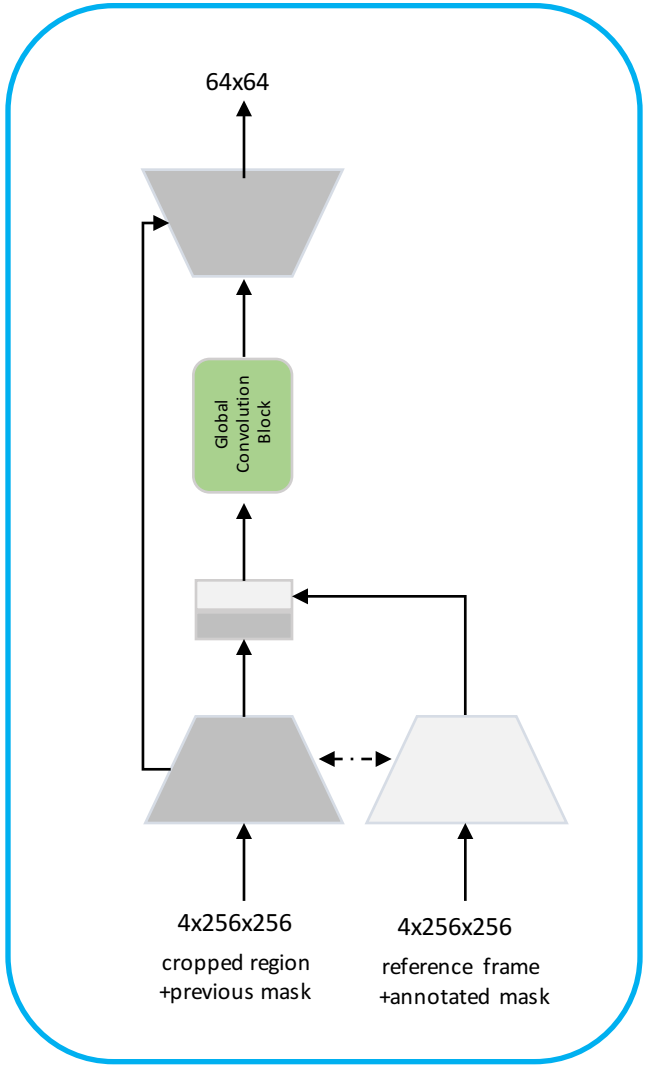
To estimate the target state in each frame, N=256 target candidates x^1,…, x^N sampled from candidate bounding boxes which are around the previous target state are evaluated using the network, and we obtain their scores $f(x^i)$. The optimal target state x^* is given by finding the example with the maximum score as

$$x^* = \underset{x^i}{\operatorname{argmax}} f(x^i)$$



RGSN: Reference-Guided Segmentation Network



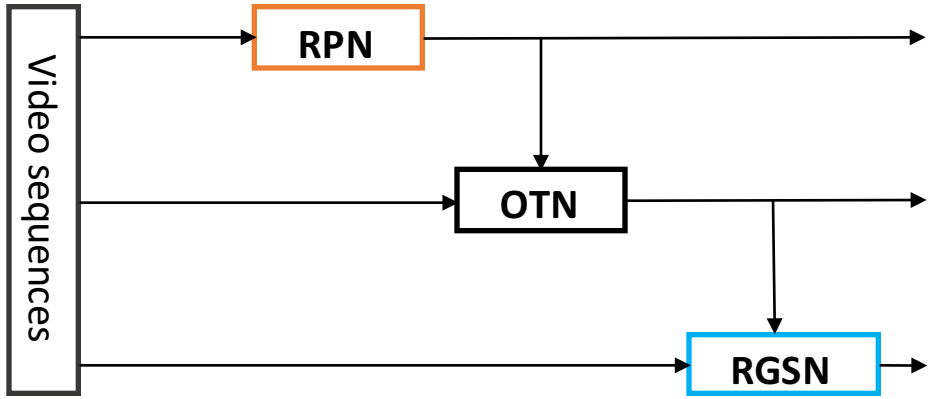


Then, the box with the highest score evaluated by OTN is selected to crop and resize the frame for normalizing the scale variation of objects.

Reference-Guided Segmentation Network will make use of both cropped region with previous mask and the reference frame to segment target object.

Wug Oh S, Lee J Y, Sunkavalli K, et al. Fast Video Object Segmentation by Reference-Guided Mask Propagation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018.

Offline Training



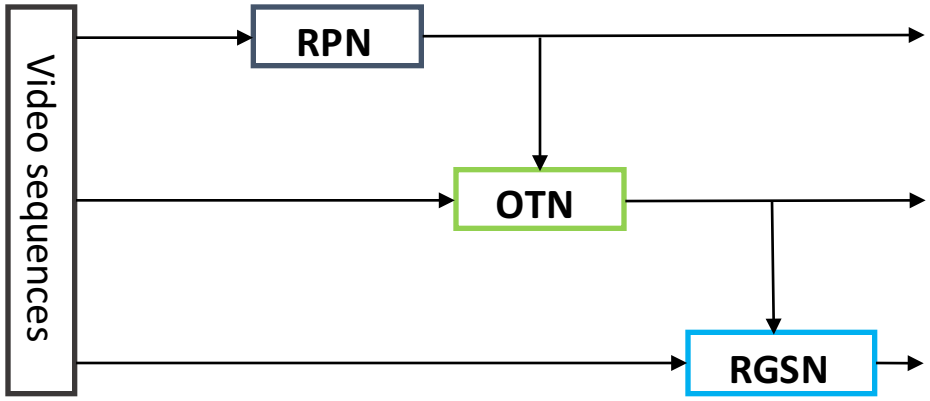
RPN RPN adapts Resnet-152 as backbone and is trained on COCO

RGSN RGSN adapts Resnet-50 as backbone and is trained on YouTube-VOS training dataset

AUG:

1. Random select two frames as a current frame and a reference frame.
2. Sample bounding boxes around the ground truth box and random scale from 1.5~2.0
3. Encode the previous mask as a heatmap with a two-dimensional Gaussian distribution

Online Training



OTN Update model during inference

RGSN Fine-tune with first annotated frame before inference for only one time

AUG:

1. Sample bounding boxes around the ground truth box and random scale from 1.5~2.0
2. Encode the previous mask as a heatmap with a two-dimensional Gaussian distribution

The influence of Reference-Guided Segmentation Network

Method	J seen	J unseen	F seen	F unseen	Mean
P + T+ naïve segmentation	61.3	50.5	61.9	55.3	57.1
P + T+ RGSN	66.3	51.2	69.2	57.2	**61.0**

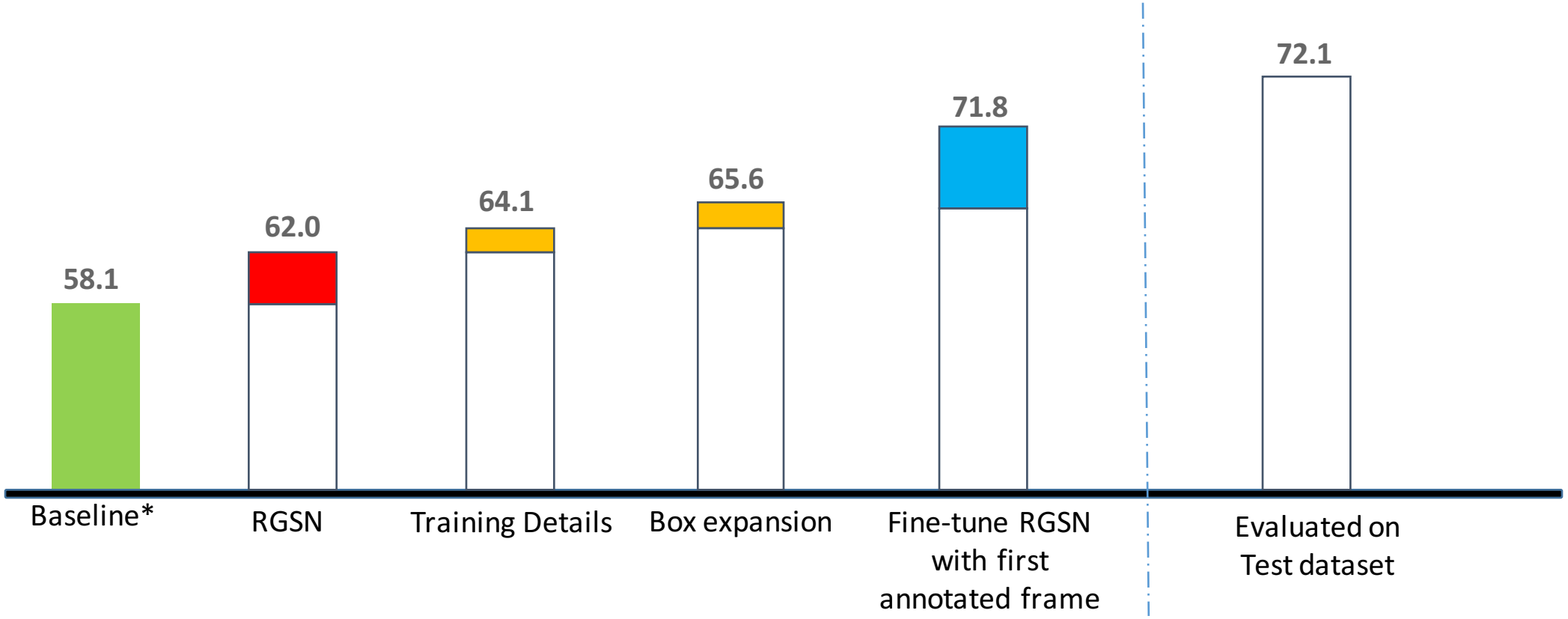
Reference-Guided Segmentation Network outperforms naïve segmentation Network

The influence of tracked box expansion

Method	J seen	J unseen	F seen	F unseen	Mean
PTS+1.0x tracked box	66.3	51.2	69.2	57.2	61.0
PTS+1.4x tracked box	67.9	52.7	70.6	58.6	62.4
PTS+1.5x tracked box	68.4	52.5	70.9	58.3	**62.5**
PTS+1.6x tracked box	68.5	52.3	70.9	57.8	62.4
PTS+1.7x tracked box	68.5	52.1	70.9	57.2	62.2

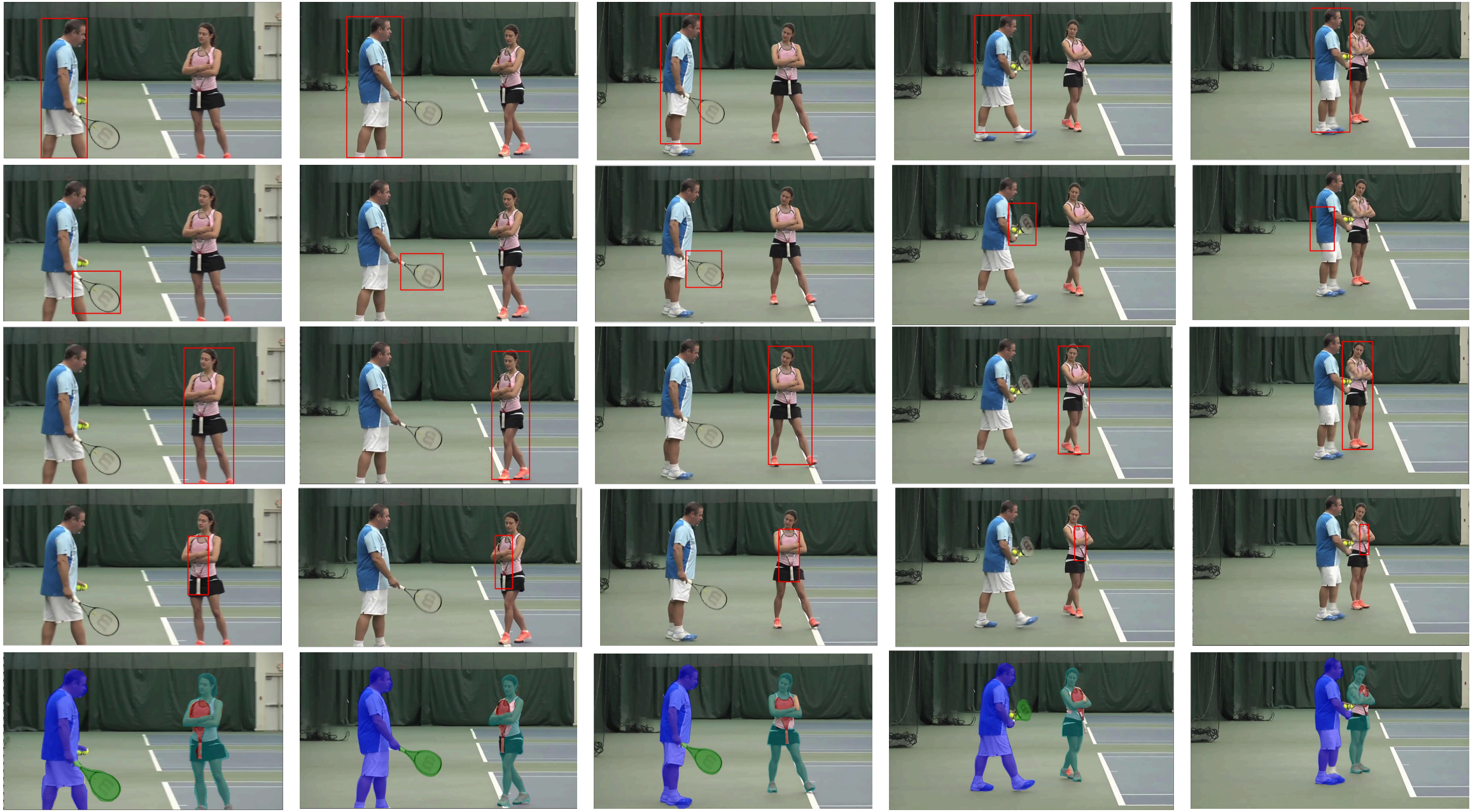
The proper box expansion can improves the result consistently

Summary



*Baseline: RPN + OTN + naïve segmentation network

Visualization



Visualization

Speed

- 30 hours for offline-training (RGSG)
- 0.9 second per frame for online-learning and inference
- Hardware: a single Titan X Pascal GPU
- Implemented using PyTorch

Conclusions

1. PTS is a unified, simple yet effective framework for video object segmentation.

2. The proposal network helps to bring objectness info for VOS by supervised pre-training.

3. PTS utilizes the SOTA video object tracking and video segmentation methods.

Future directions

1. Integrate long-term temporal features of OTN into RGSN
2. Joint training of three networks
3. Speedup

Thanks & Questions