JD Cloud & AI

Robust Visual Object Tracking with Two-Stream Residual Convolutional Networks

Ning Zhang[1], Jingen Liu[1], Ke Wang[2], Dan Zeng[3], Tao Mei[1]

1. JD AI Research, Mountain View, Beijing
2. Migu Culture & Technology, Beijing
3. Shanghai University, Shanghai

Paper ID:1155

JD Cloud & AI

Outline

01 **VOT Introduction**

Visual Object Tracking
Task, challenge and algorithm criteria

02 **Related Work**

VOT History, Baseline approach

03 **Proposed System**

Two-Stream Residual Convolutional Network (TS -RCN)

04 **Experiment and Result, Conclusion**

Results, Ablation, Discussion,
Video demo and Conclusion

VOT Introduction

Challenge and Criteria

- Challenges:

  complex object motion/shape, nonrigid nature of the objects, background noise, partial and full object occlusion, illumination, real-time process

- Criteria:

  - robustness: not affected by occlusion, noise, illumination, motion, deformation, blurring, etc.

  - accuracy: accurate capture of the bounding box target object.

  - tracking is a real-time task.

Yilmaz, Alper, Omar Javed, and Mubarak Shah. "Object tracking: A survey." *Acm computing surveys (CSUR)* 38.4 (2006): 13-es.

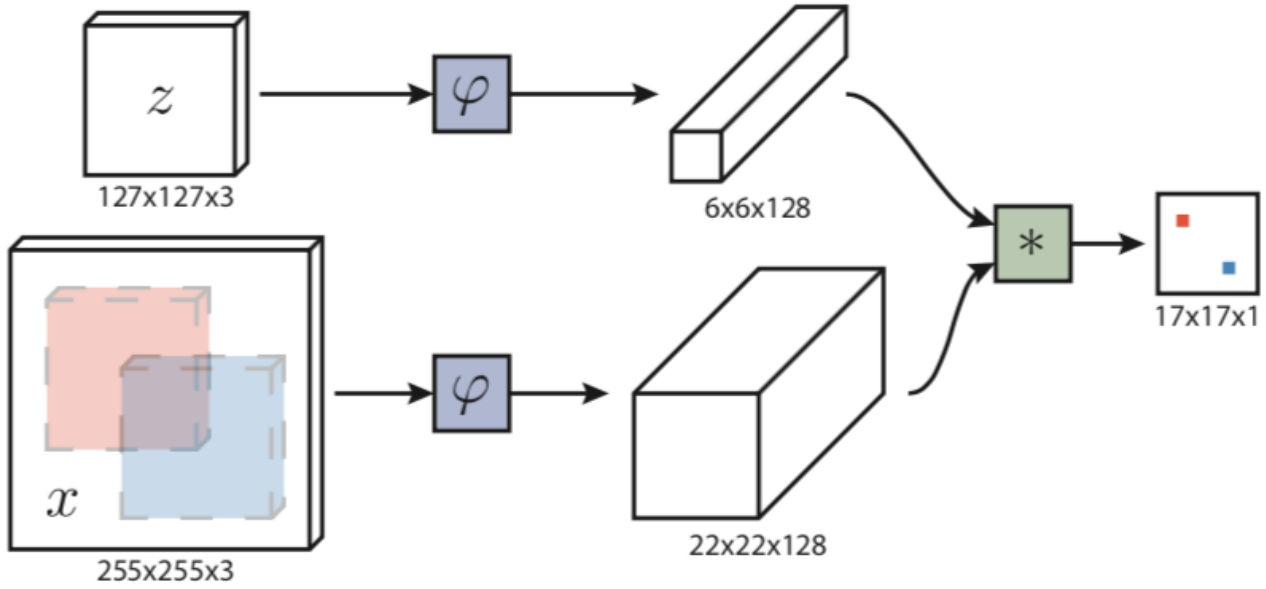
VOT Related Work

A brief history of VOT

JD Cloud & AI

- Previously (before 2016), mostly online, filter-based process, not much deep learning.

- Tracking can be viewed as a similarity match, and can be learned off-line.
- SiamFC
  - Large-training SiamFC is the first to use 2015 ILSVRC Object detection from video task (VID) , VID has 4417 videos，over 2 Million Annotation

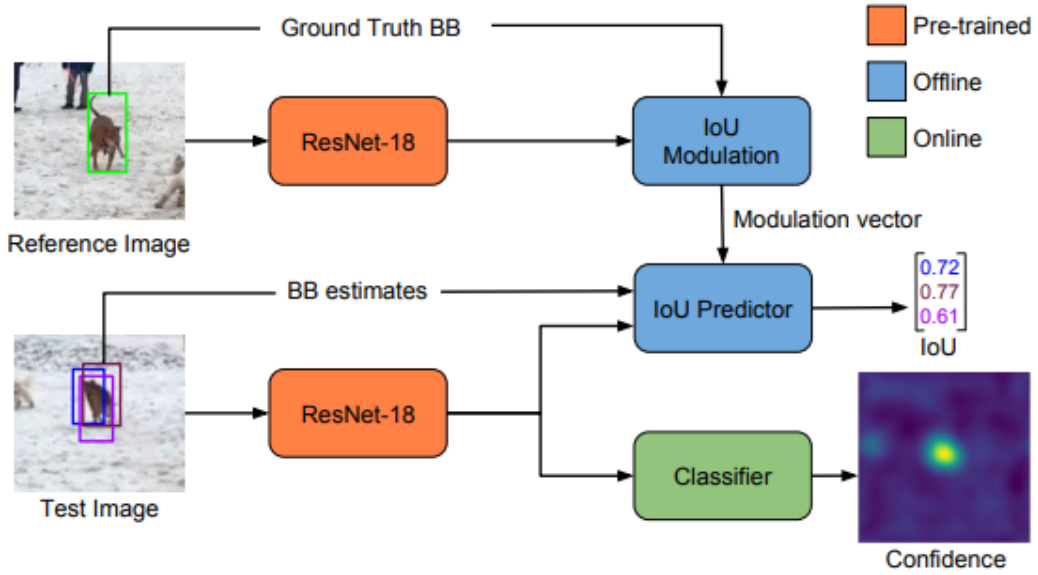


[1] Bertinetto, Luca, et al. "Fully-convolutional siamese networks for object tracking." *ECCV2016*

VOT Related Work

DiMP Tracker, a strong baseline

- Siamese Limitation:
  - Siamese offline discard the background appearance,
  - training set low distinction between object and distractor
- Solution:
  - Propose a prediction-model for classification, end-to-end offline training
  - Online fine-tune offline classification
- Tracking Speed: FPS 40

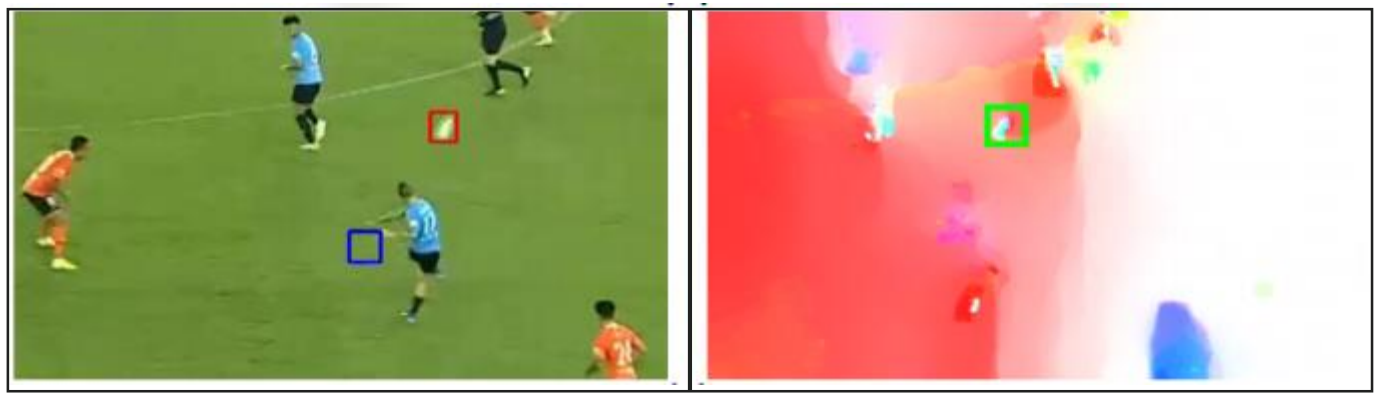


[1] Bhat, Goutam, et al. "Learning discriminative model prediction for tracking." *ICCV 2019*.

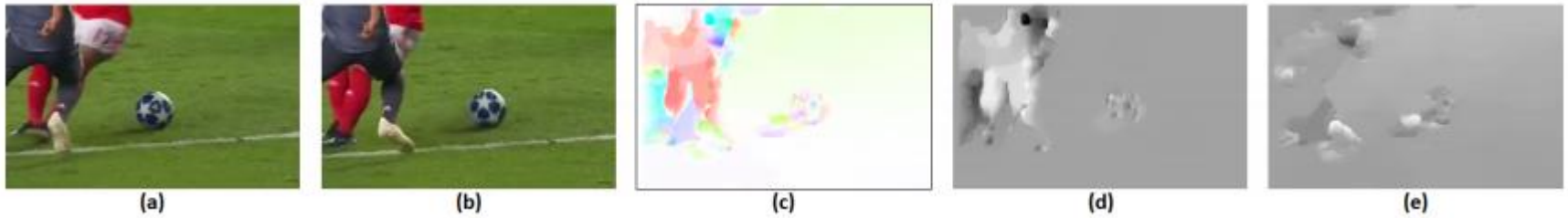
Proposed System: Motivation

Optical fow approach

- Motivation: deformed object, motion blur fails the appearance-based tracker



- Intuitive solution

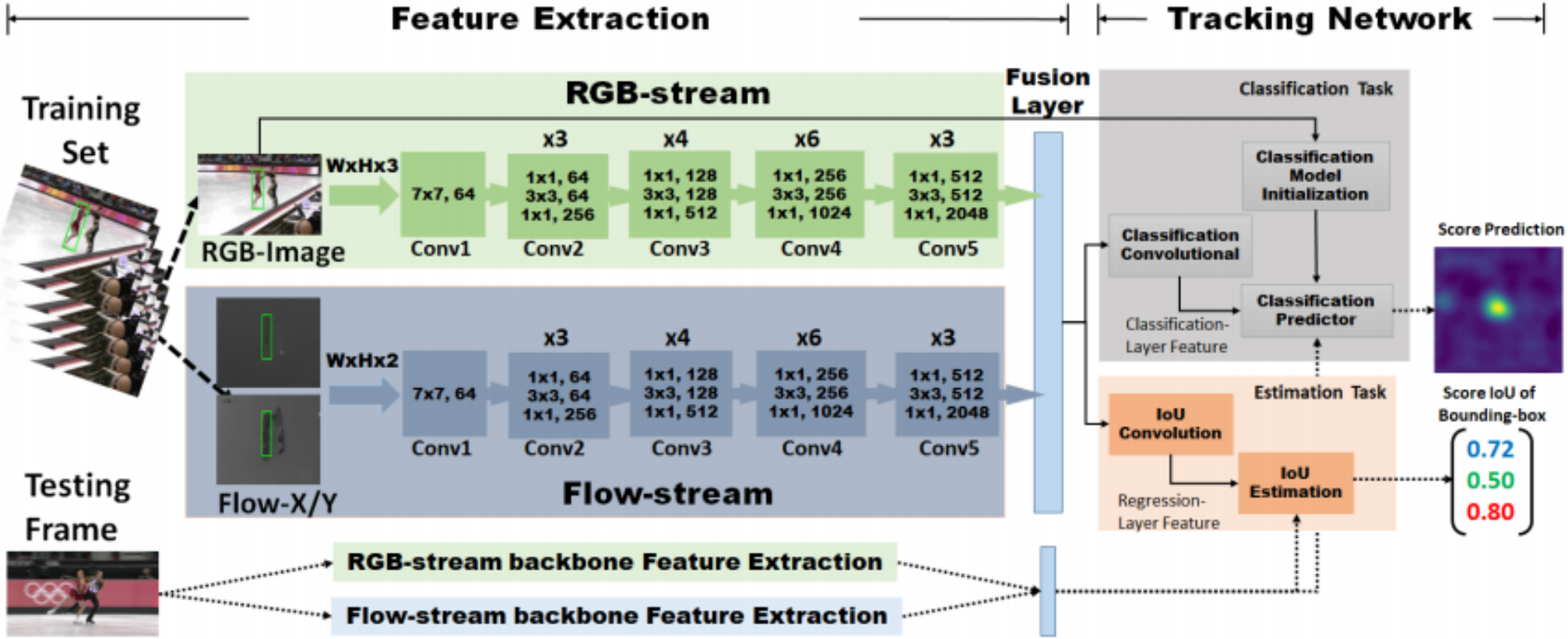


Optical flow visualization: (a-b) consecutive video frames of a targeted soccer ball. (c): Color visualization based on displacement vector's magnitude and direction, using the HSV color-space. (d-e): horizontal and vertical displacement vector fields d^t_u , and d^t_v, respectively, with higher intensity representing positive values.

Proposed System: TS-RCN

Two-stream architecture: appearance + motion (TS-RCN)

- Motivation: Solution: two-stream residual convolutional networks (TS-RCN)
  - End-to-end trainable Two-stream Optical-flow + Appearance
  - Based on residual network backbones
  - Up to speed 38FPS



Experiment and Result

JD Cloud & AI

Two-stream: appearance + motion (TS-RCN)

	LADCF [34]	MFT [35]	SiamRPN [18]	DRT [36]	RCO [12]	UPDT [37]	ECO [38]	SiamFC [4]	ATOM [3]	SiamFC++ [17]	DaSiamRPN [39]	SiamMask [40]	SiamRPN++ [16]	DIMP-50 [5]	TS-RCN ours
EAO ↑	0.389	0.385	0.383	0.356	0.376	0.378	0.280	0.187	0.401	0.426	0.326	0.387	0.414	**0.422**	**0.459**
Accuracy ↑	0.505	0.508	0.587	0.519	0.507	0.536	0.487	0.505	0.590	0.587	0.569	**0.642**	0.600	**0.602**	0.579
Robustness ↓	0.159	**0.140**	0.276	0.201	0.155	0.184	0.276	0.585	0.204	0.183	0.337	0.295	0.234	0.162	**0.139**

Experiment

Ablation study

JD Cloud & AI

- Backbone Depth: residual network-18, -50, -101, -152

ResNet Depth	18	50	101	152
EAO ↑	0.345	**0.419**	**0.383**	0.233
Params (millions)	11.69	25.56	44.56	60.19

- Backbone Architecture: ResNet, ResNeXt, WRNs

	TS-RCN ResNet-50	TS-RCN ResNeXt-50	TS-RCN WRNs-50
EAO ↑	0.419	**0.459**	0.390
Accuracy ↑	0.571	**0.579**	0.568
Robustness ↓	0.168	**0.139**	0.195

- Training Datasets: GOT-10K, LaSOT, ImgNetVid

DBs	GOT-10k	GOT-10k + LaSOT	GOT-10k + LaSOT + ImgNetVid
EAO ↑	**0.383**	**0.459**	0.378

Video Demo

JD Cloud & AI

Video demo 1: confused background, occlusion, and deformed object

Video Demo

JD Cloud & AI

Video demo 2: sudden motion change, occlusion, deformed object

Conclusion

JD Cloud & AI

- TS-RCN strategically combines the RGB appearance and the optical flow motion inputs.
- TS-RCN exploits a “wider” residual network ResNeXt as its feature extraction back bone to further improve the tracking performance.
- TS-RCN was evaluated at benchmark datasets with better performance,
    - including: VOT2018, VOT2019, and GOT-10K.
- We have successfully demonstrated that our two-stream model can outperform the appearance-based tracker with real-time FPS.

JD Cloud & AI

Thanks