Learning Multi-Domain Convolutional Neural Networks for Visual Tracking

Hyeonseob Nam Bohyung Han

Computer Vision Lab.

Dept. Computer Science and Engineering

POSTECH

Our VOT2015 Submission

<Average over all sequences>

Accuracy = 0.60, Failures = 0.77

<ball2>

Accuracy=0.80, Failures=0.00

<soldier>

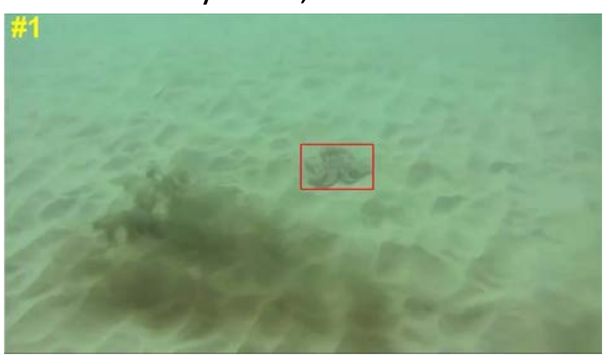
Accuracy=0.51, Failures=0.07

<sphere>

Accuracy=0.74, Failures=0.00

<octopus>

Accuracy=0.62, Failures=0.00



<godfather>

Accuracy=0.52, Failures=0.13

<wiper>

Accuracy=0.69, Failures=0.13

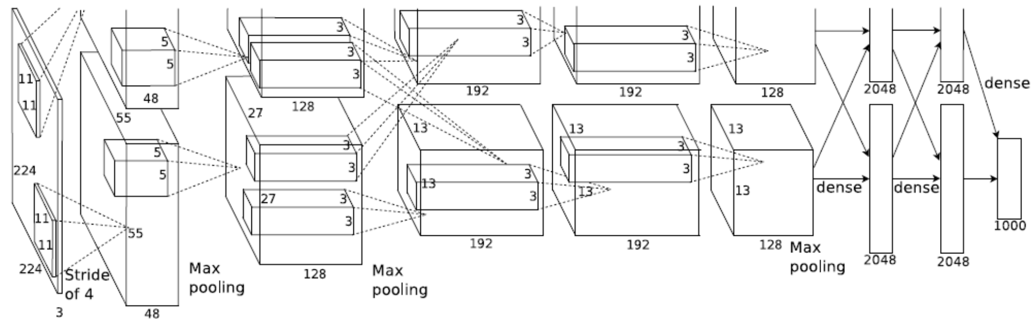
Ground-truth **Our 15 repetitions**

Key Idea

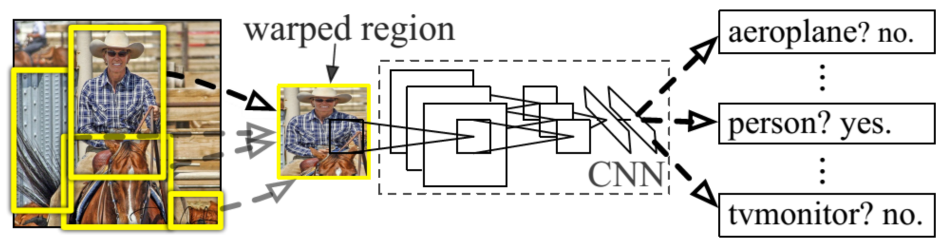
A Deep Convolutional Neural Network
trained on large amounts of visual tracking data

Convolutional Neural Networks (CNNs)

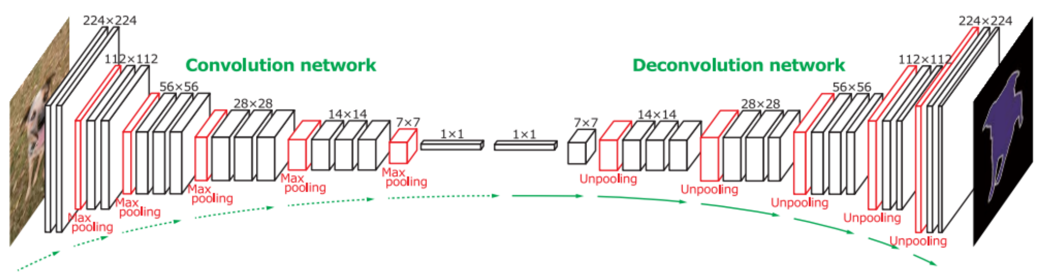
- Image Classification
  [Krizhevsky et al. NIPS'12]
  [Szegedy et al. CVPR'15]
  [Simonyan et al. ICLR'15]
  …



- Object Detection
  [Sermanet et al. ICLR'14]
  [Girshick et al. CVPR'14]
  [He et al. ECCV'14]
  …



- Semantic Segmentation
  [Chen et al. ICLR'15]
  [Long et al. CVPR'15]
  [Noh et al. ICCV'15]
  …



- Face Recognition, Image Captioning, Question Answering, …

POSTECH

Top-Performing Trackers from VOT2014

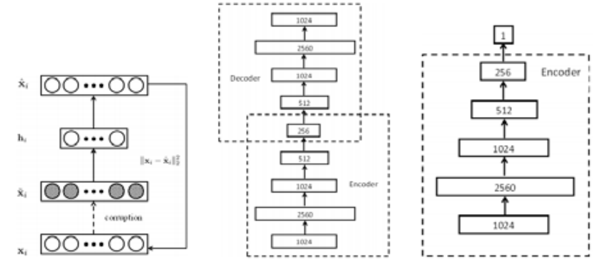
Low-Level Features

Tracker	Features	Scale	Visual model
DSST*	HoG+intensity	Yes	Discr. correl. Filtr
SAMF	HoG+colornames	Yes	Discr. correl. Filtr
KCF	HoG	Yes	Discr. correl. Filtr
DGT	Superpixels + color	Yes	Part-based
PLT_{14}	Color, intensity, derivs.	Yes	Discr. Regression
PLT_{13}	Color, intensity, derivs.	No	Discr. Regression

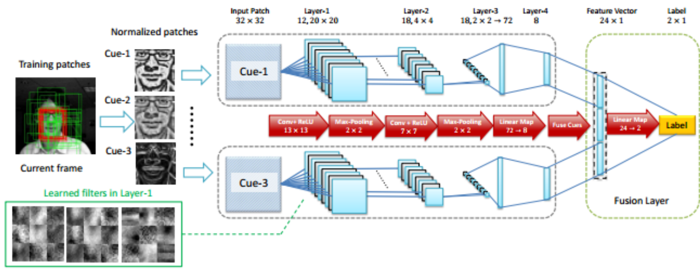
[Kristan et al. ECCVW'14]

Deep Learning for Visual Tracking

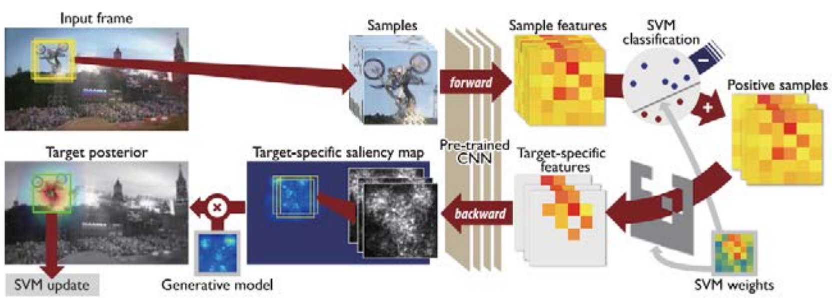
- Stacked Denoising Autoencoder
  [Wang et al. NIPS’13]



- Pool of CNNs
  [Li et al. BMVC’14]



- CNN + Online SVM
  [Hong et al. ICML’15]



- Structured output CNN
  [Wang et al. Arxiv’15]

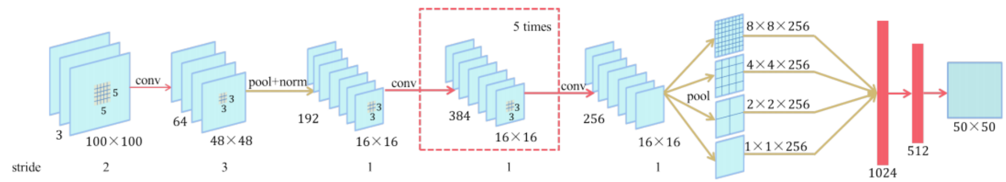
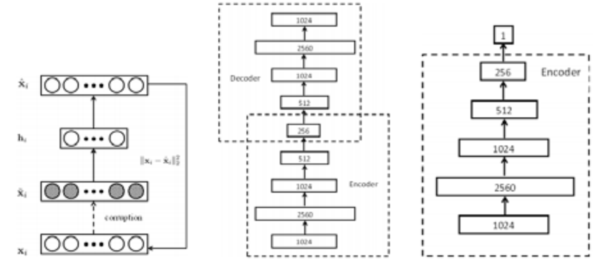


Figure 2. Architecture of the proposed structured output CNN.

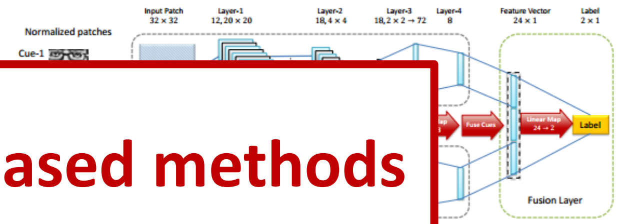
POSTECH

Deep Learning for Visual Tracking

- Stacked Denoising Autoencoder
  [Wang et al. NIPS'13]

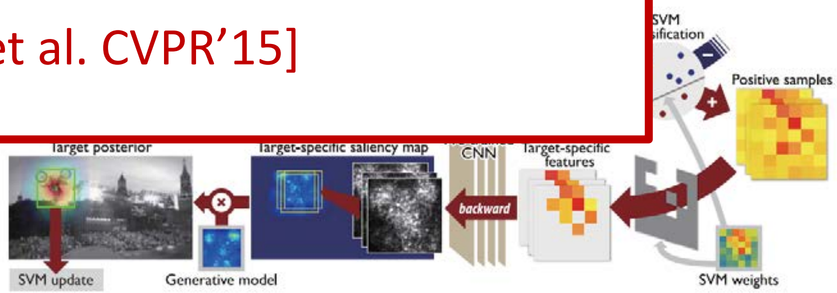


- Pool of CNNs
  [Li et a



Defeated by low-level feature based methods
- MUSTer (HOG, color, SIFT) [Hong et al. CVPR'15]
- LCT (HOG, intensity) [Ma et al. CVPR'15]

- CNN
  [Hong



- Structured output CNN
  [Wang et al. Arxiv'15]

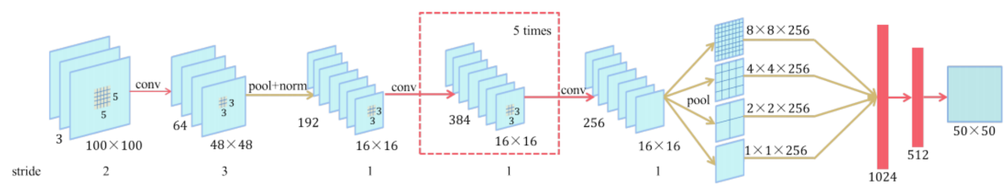


Figure 2. Architecture of the proposed structured output CNN.

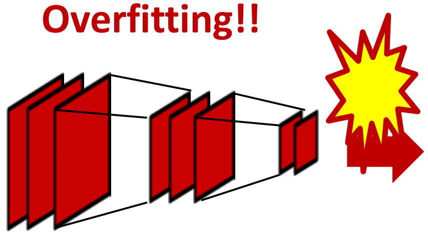
POSTECH

Issues with CNNs for Visual Tracking

- Lack of training data

1 labeled frame

Overfitting!!



Large-scale classification dataset

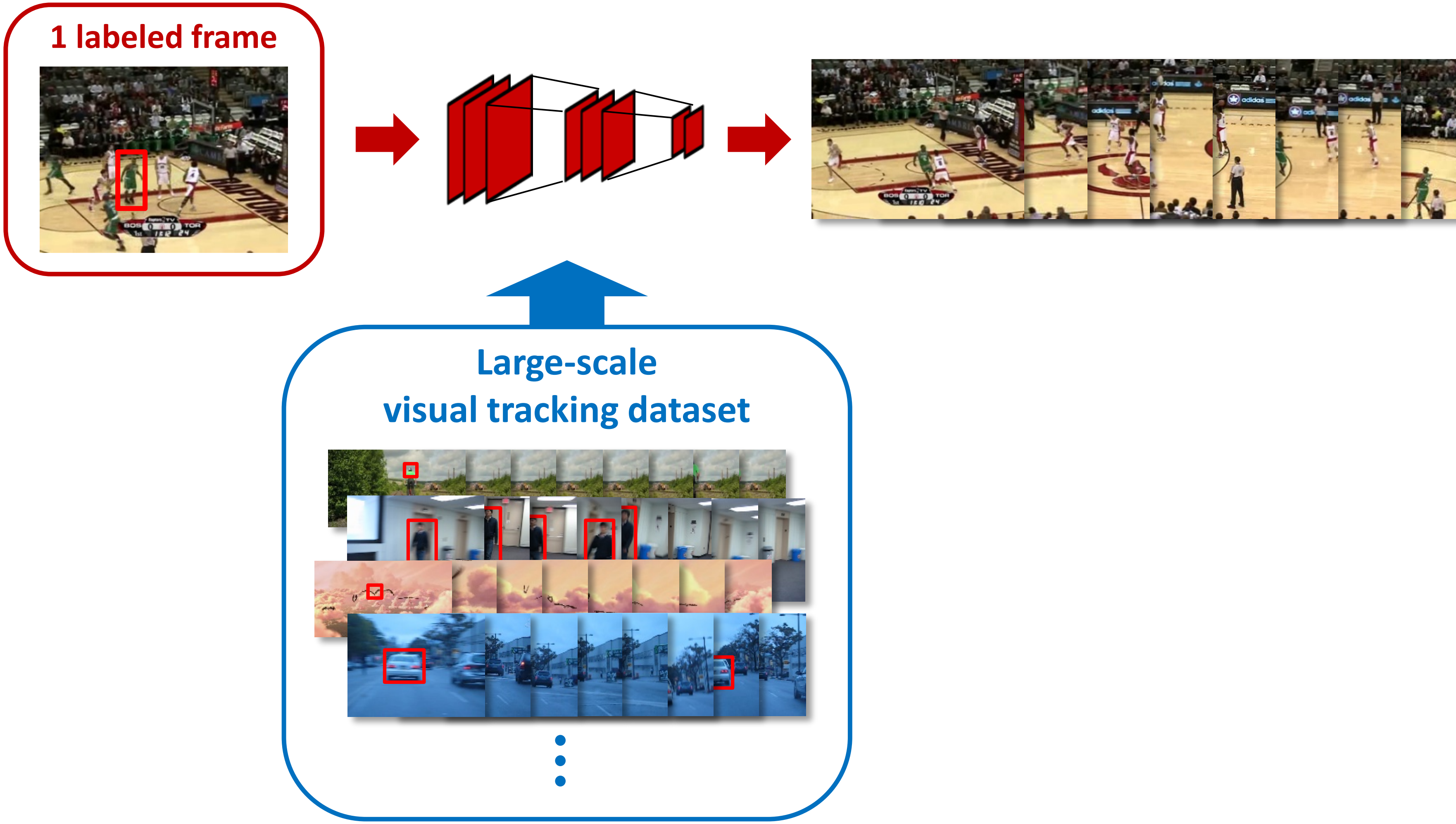
Pretrain using ImageNet?

[Hong et al. ICML’15]
[Wang et al. Arxiv’15]

Limited effect due to the fundamental difference between different tasks!

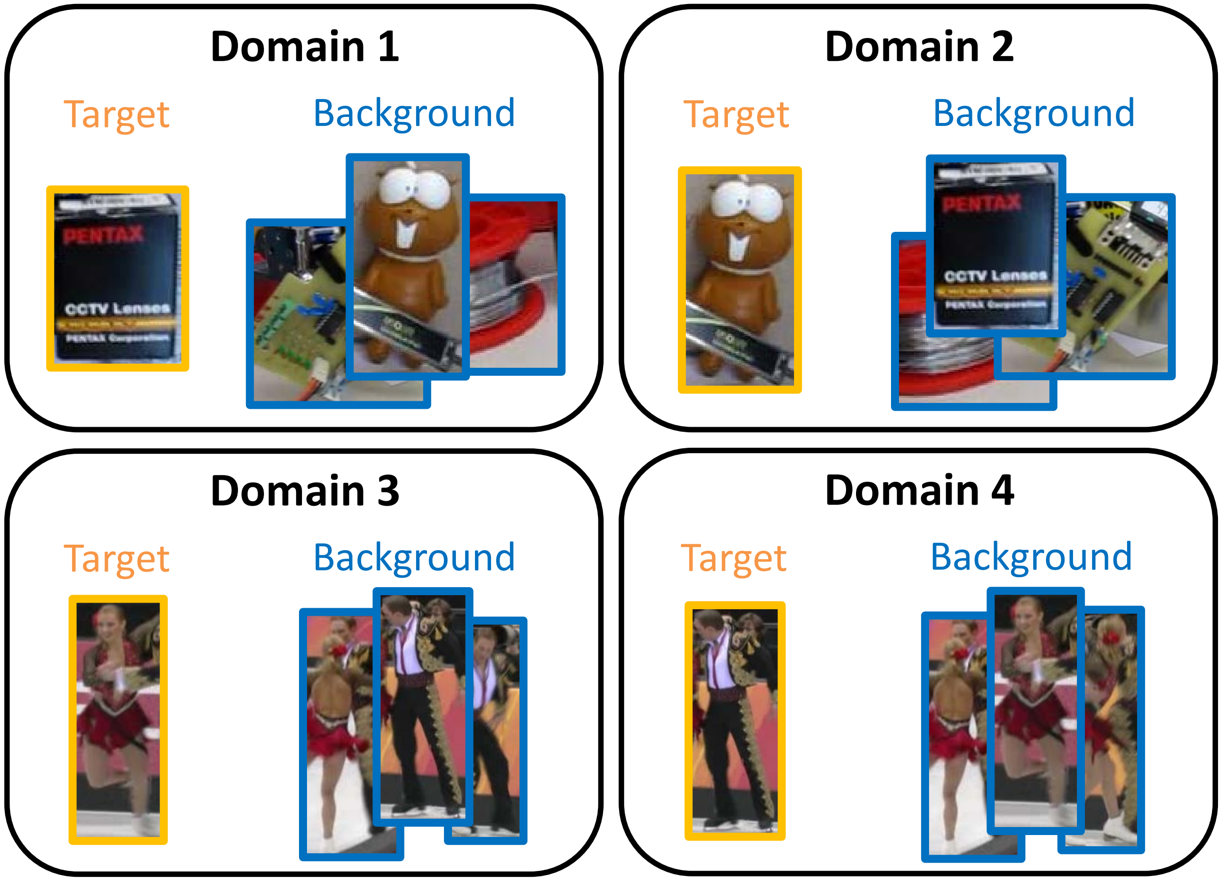
Goal

- **Exploit external tracking data to train CNN features for tracking!**



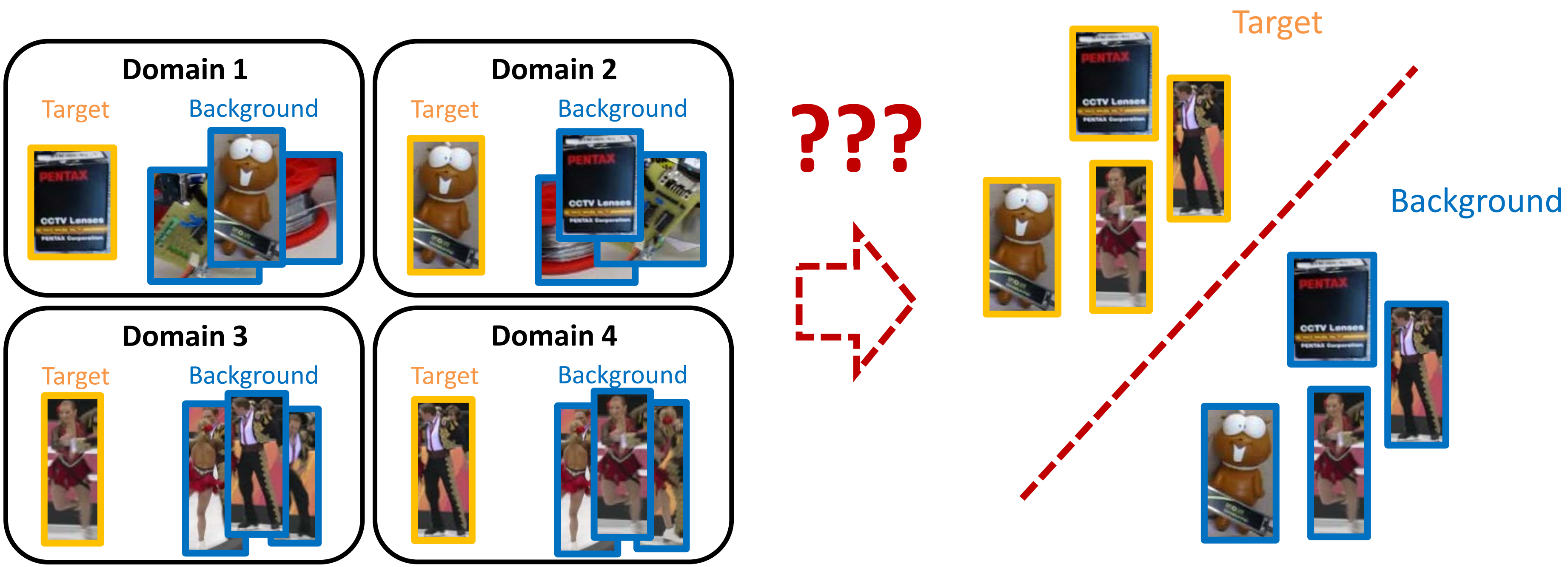
Challenge

- **Inconsistent training data** across tracking sequences (domains).



Challenge

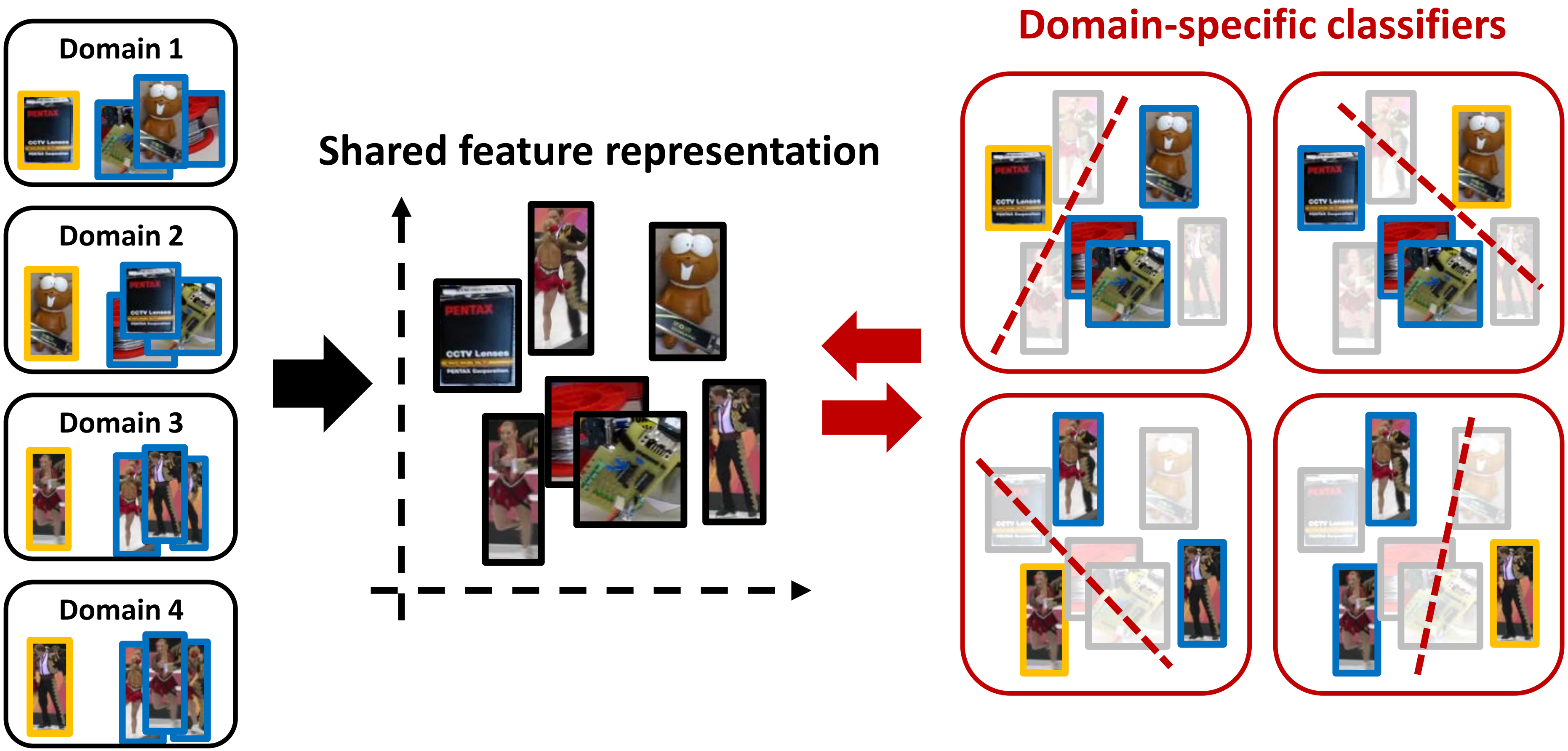
- **Inconsistent training data** across tracking sequences (domains).



POSTECH

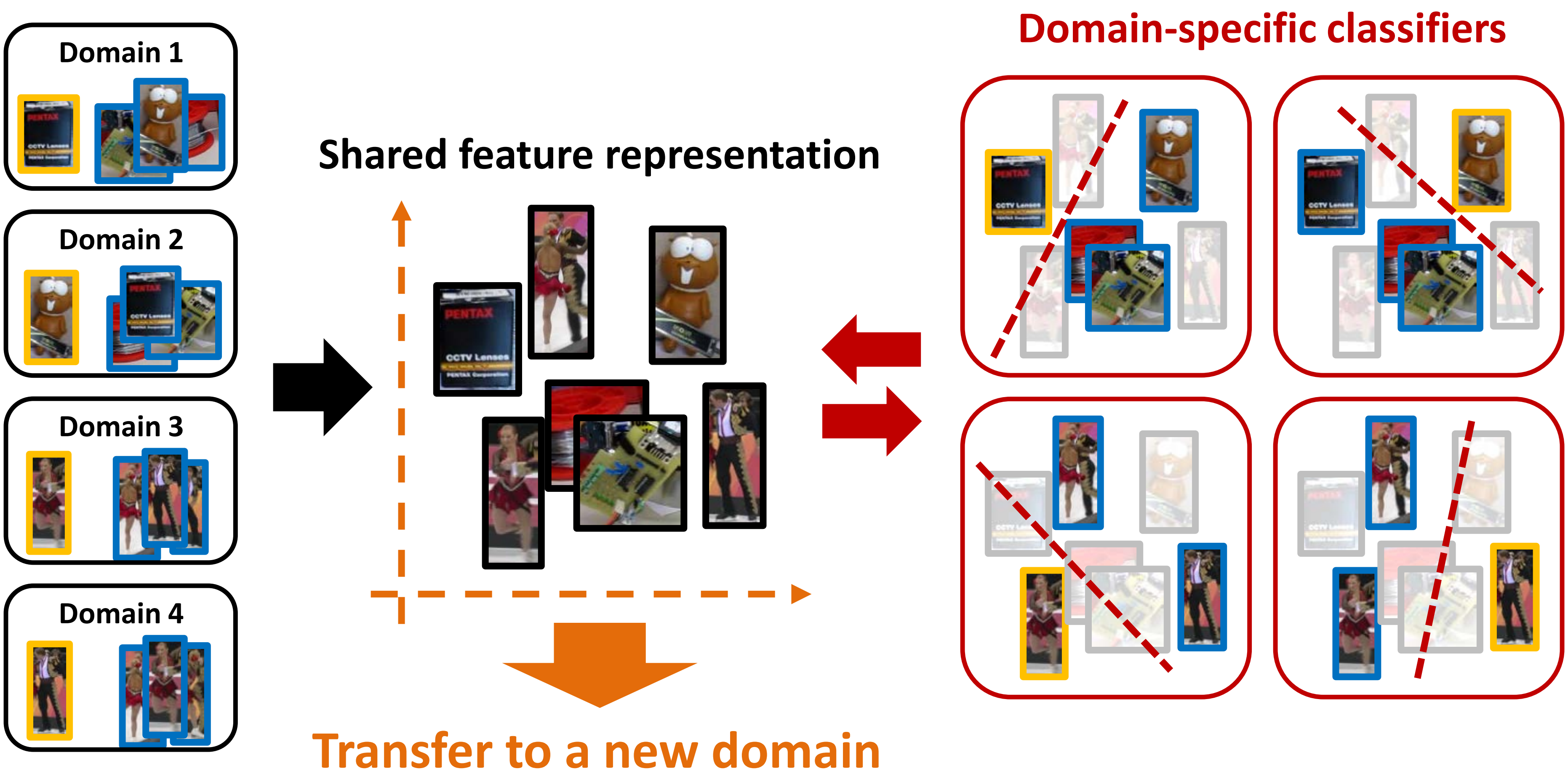
Our Approach

- Training **shared features** and **domain-specific classifiers** jointly.

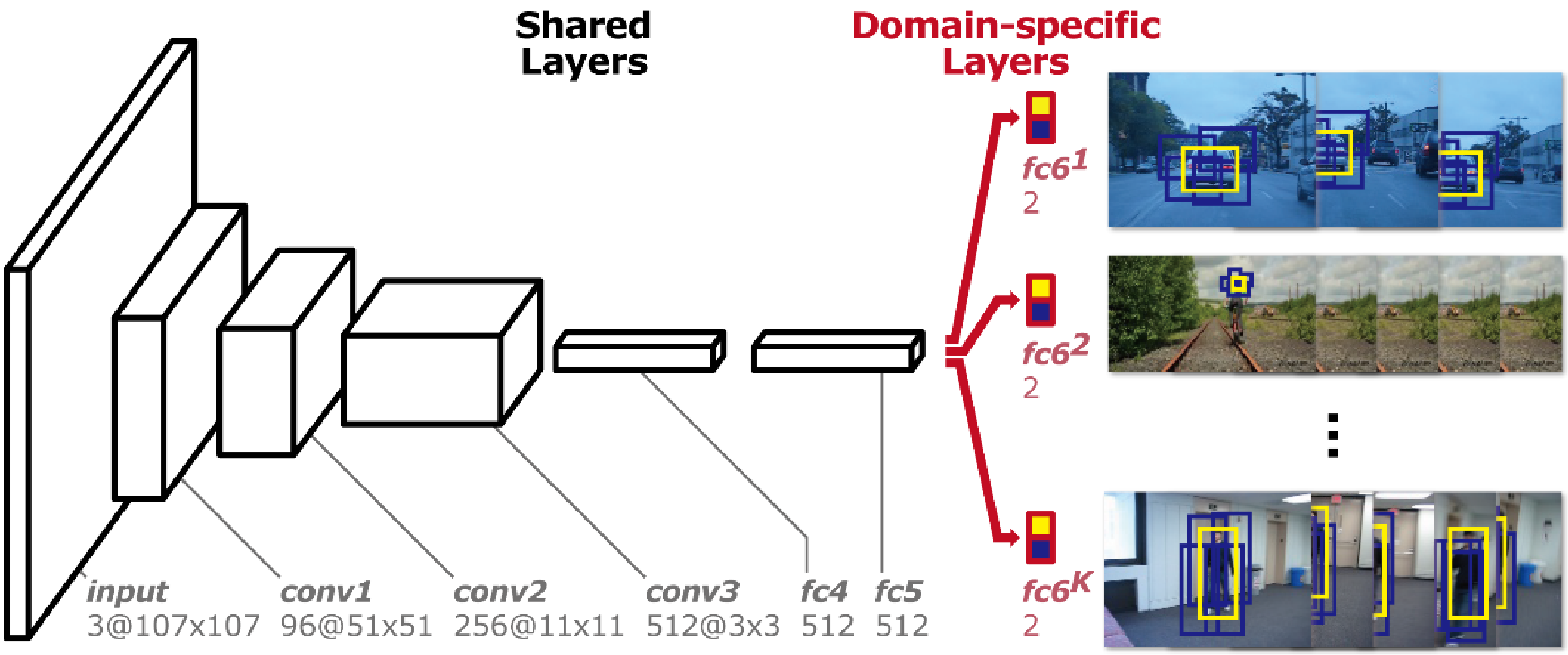


Our Approach

- Training **shared features** and **domain-specific classifiers** jointly.

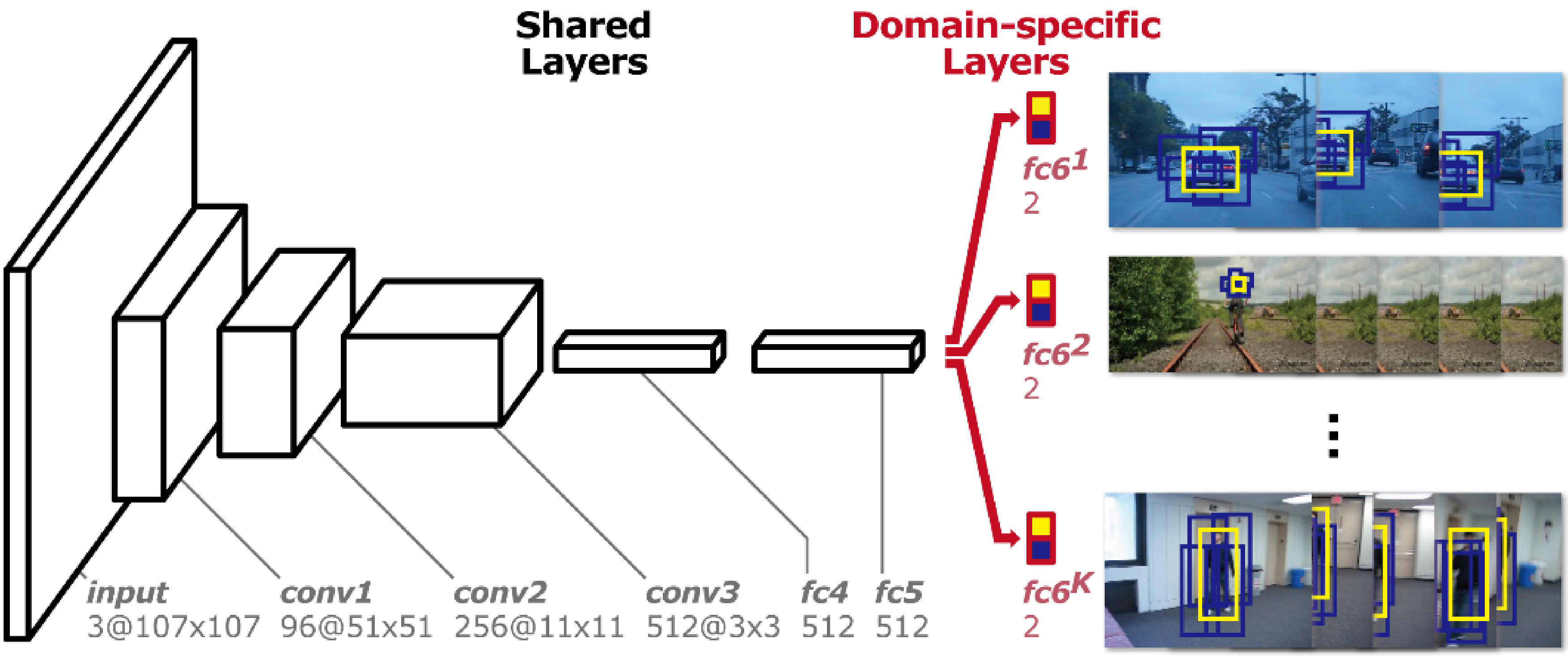


MDNet: Multi-Domain Network



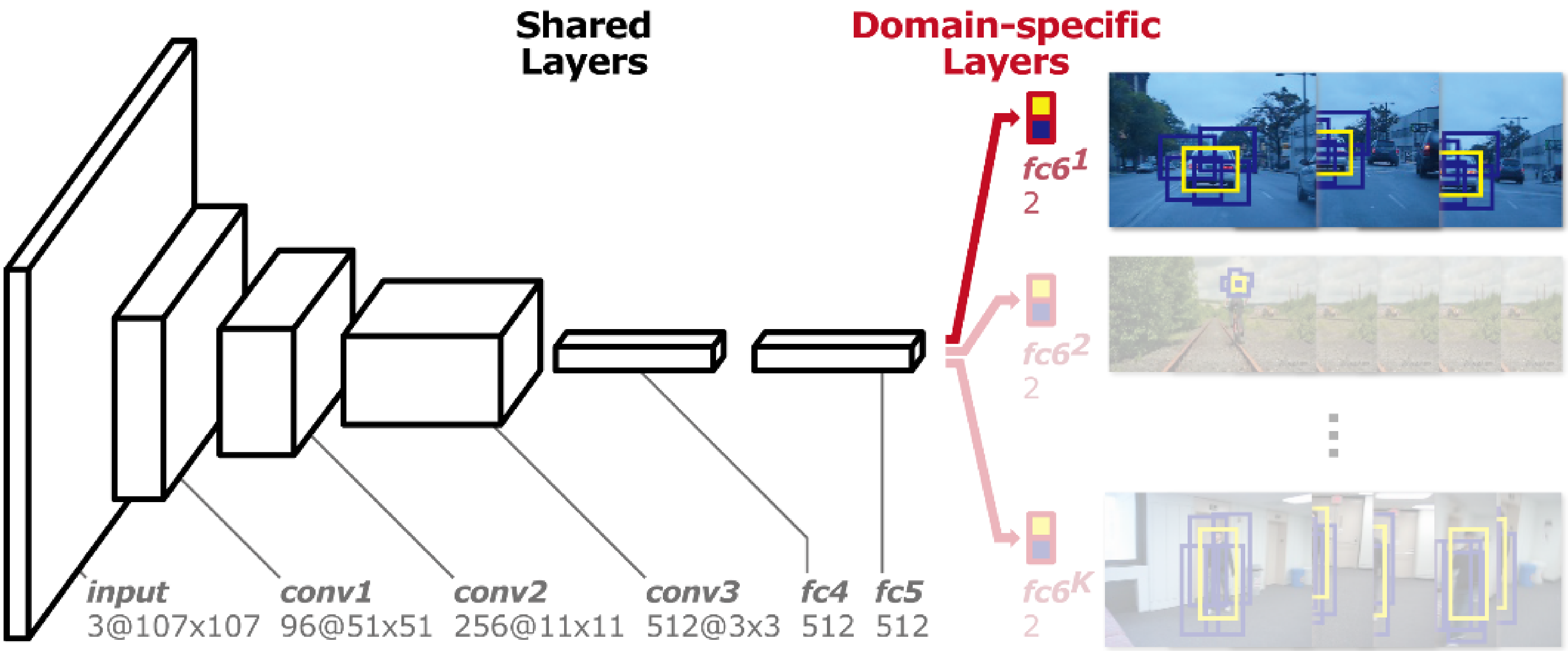
MDNet: Learning Algorithm

- Train the network for each domain iteratively.



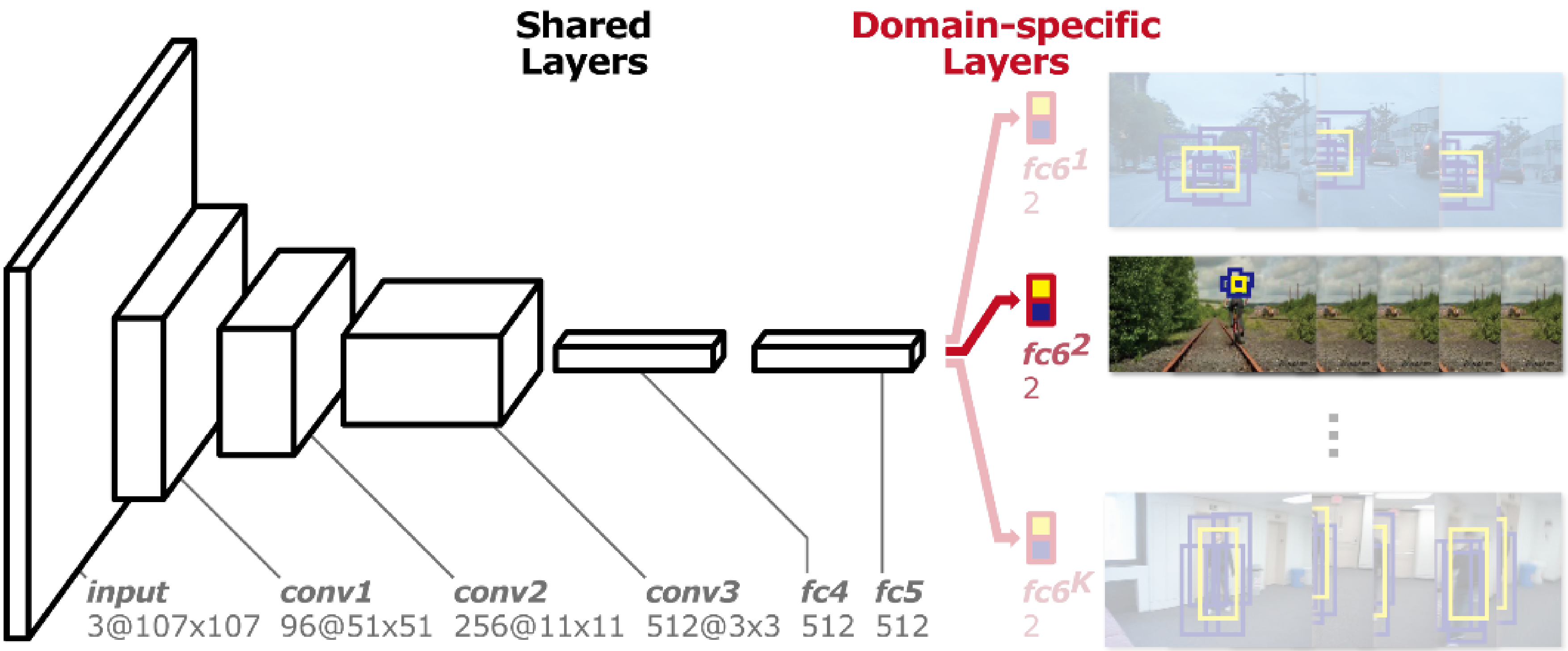
MDNet: Learning Algorithm

- Iteration #nK+1



MDNet: Learning Algorithm

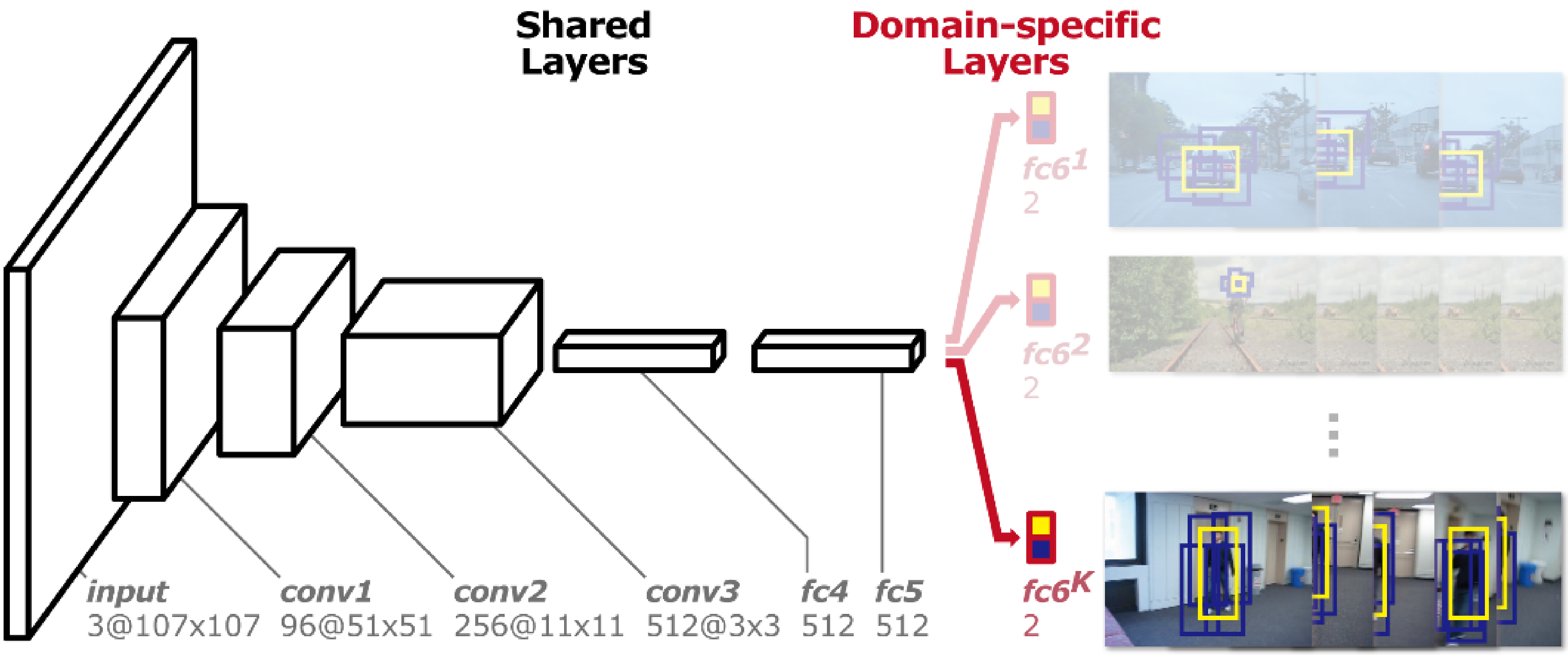
- Iteration #nK+2



POSTECH

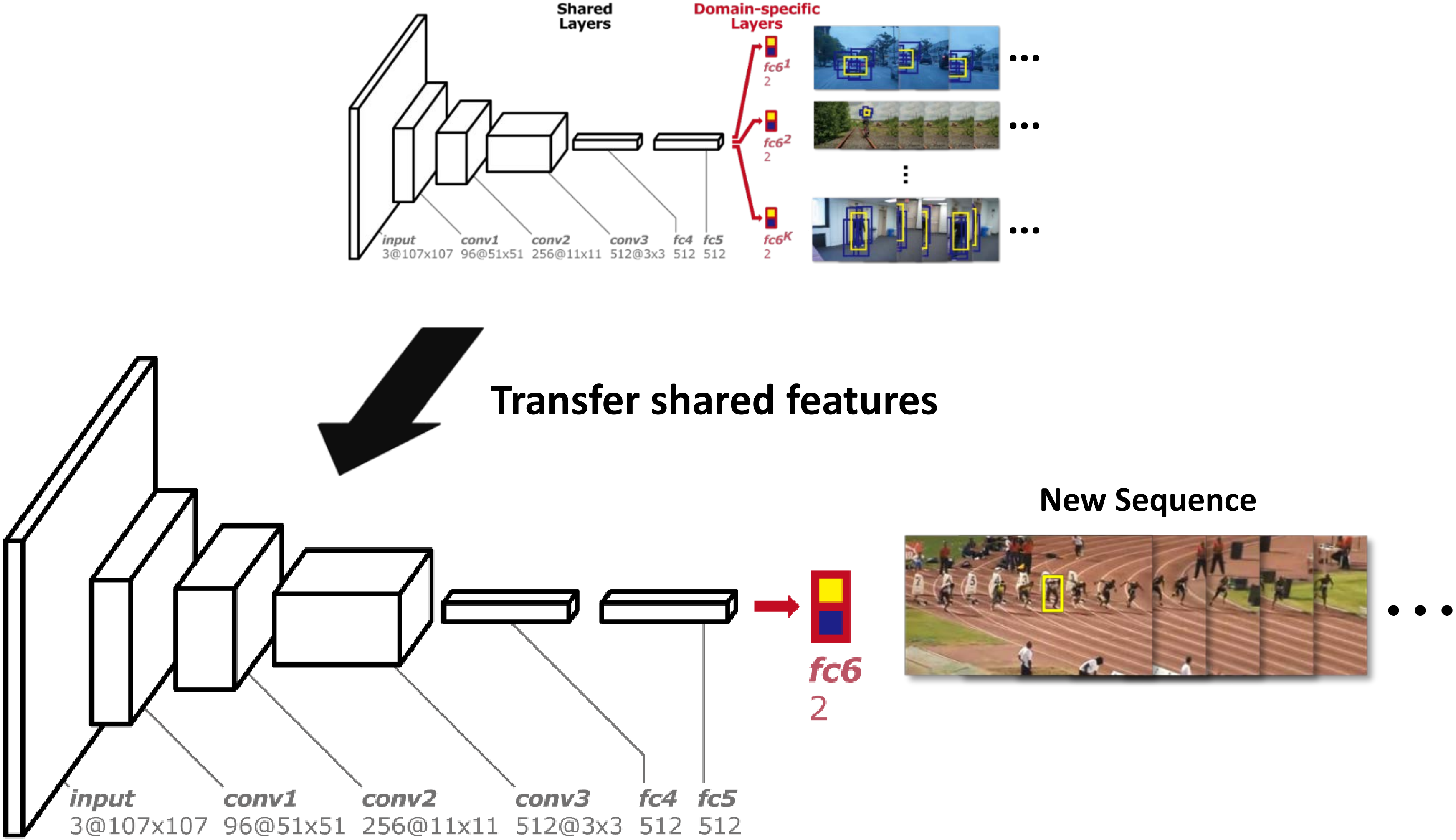
MDNet: Learning Algorithm

- Iteration #nK

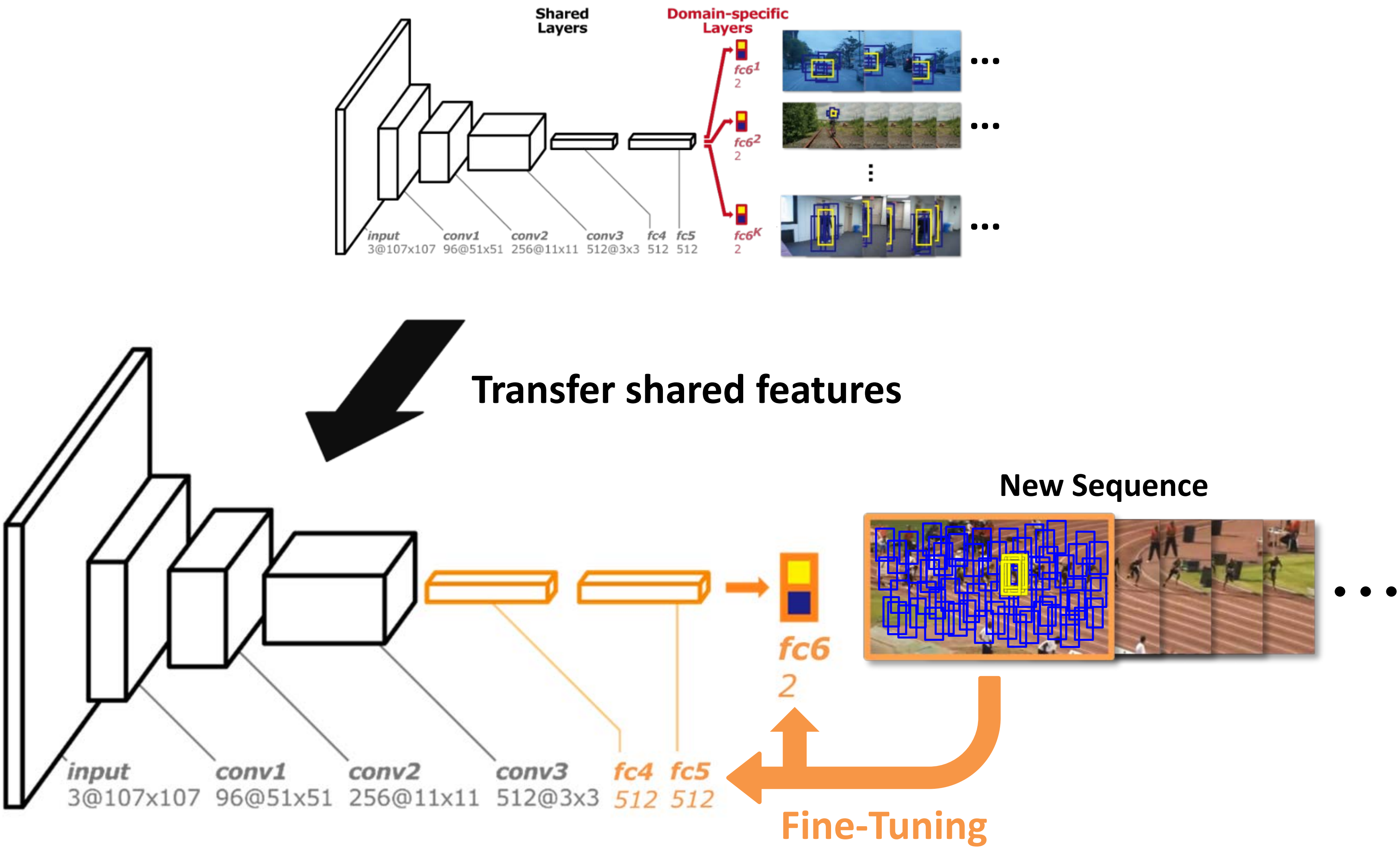


POSTECH

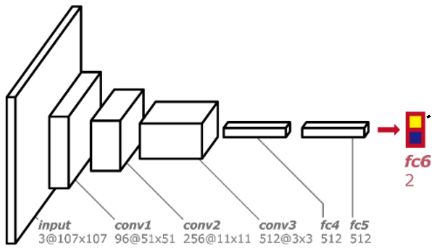
Online Tracking using MDNet Features



Online Tracking using MDNet Features



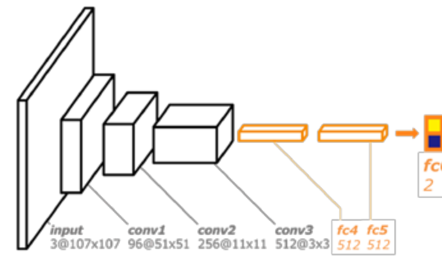
POSTECH



$f^+(\cdot)$: positive score

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

Frame $t \geq 2$



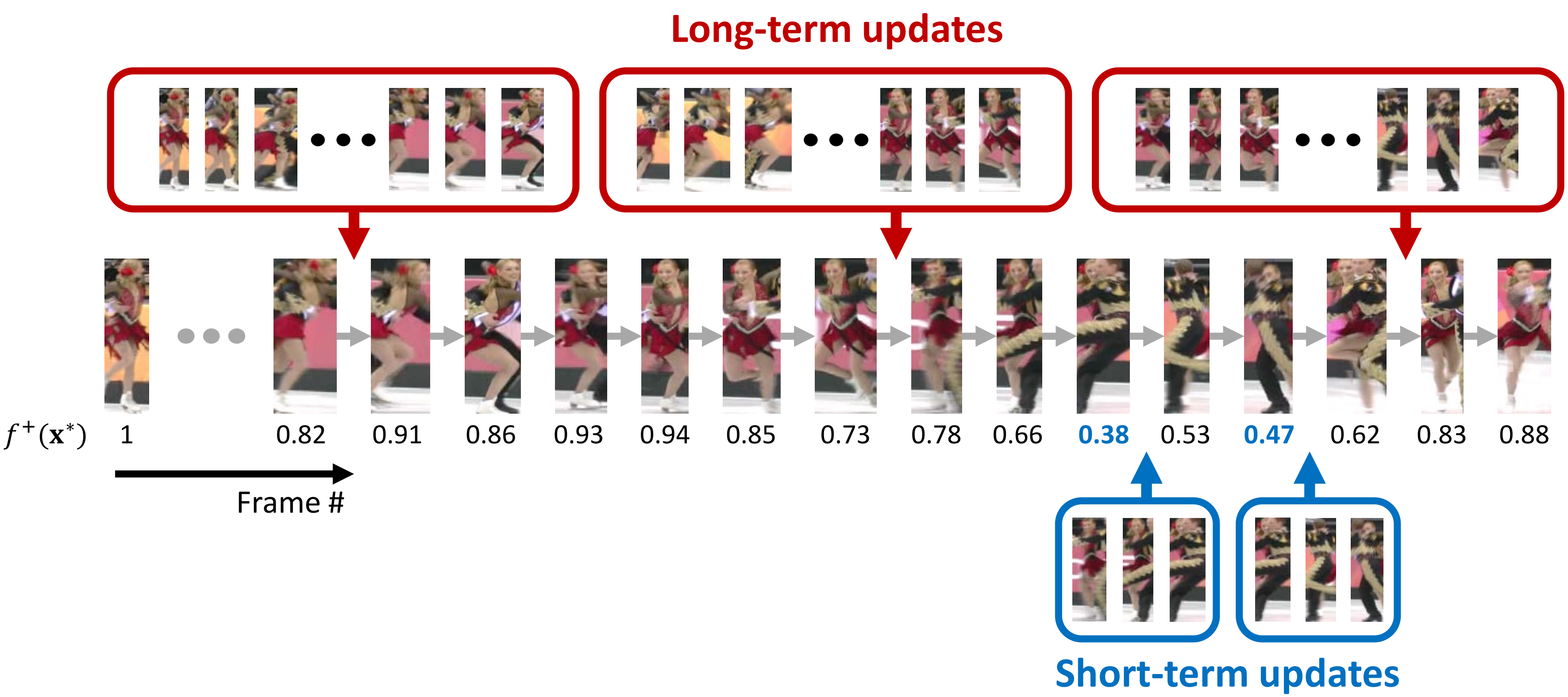
Draw target candidates → **Find the optimal state** → **Collect training samples** → **Update the CNN if needed**

Repeat for the next frame

Online Network Updates

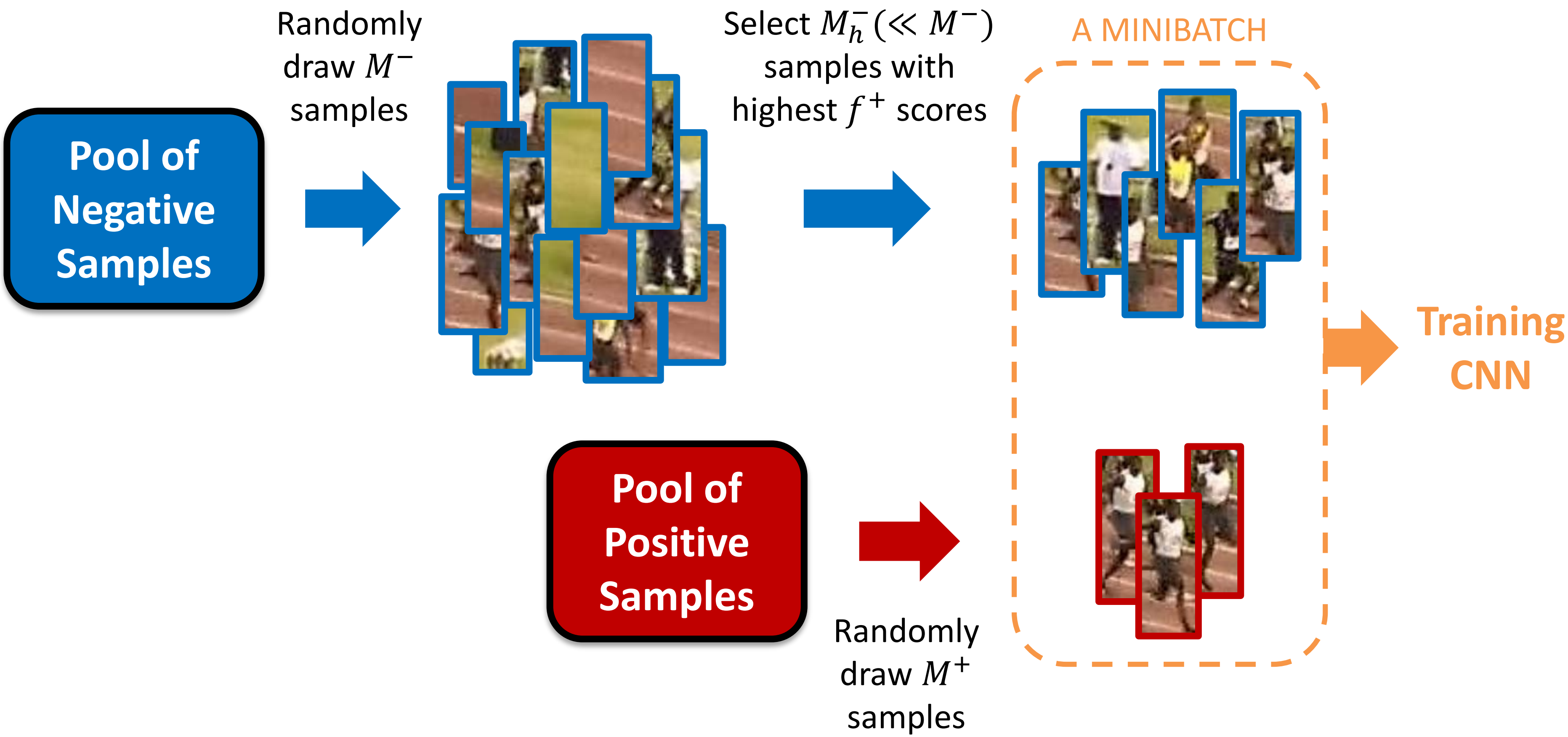
- **Long-Term Updates**
  - performed at regular intervals
  - using long-term training samples
  - **For Robustness**

- **Short-Term Updates**
  - performed at abrupt appearance changes ($f^+(\mathbf{x}^*) < 0.5$)
  - using short-term training samples
  - **For Adaptiveness**



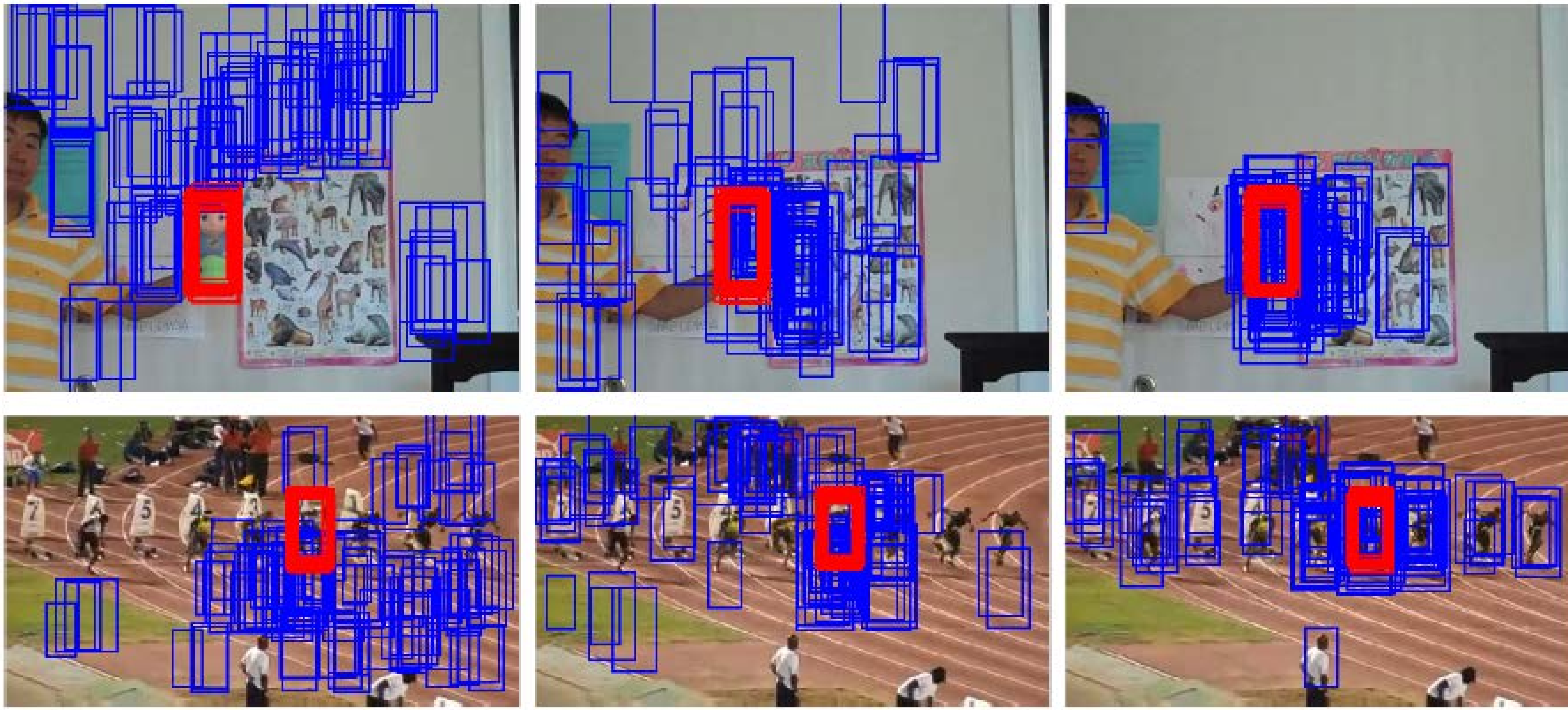
Hard Negative Mining

- Provide a “hard” minibatch in each training iteration.



Hard Negative Mining

Positive sample　　Negative sample



1st minibatch　　5th minibatch　　30th minibatch

Training iteration

Bounding Box Regression

- Improve the localization quality.
  - DPM [Felzenszwalb et al. PAMI'10]
  - R-CNN [Girshick et al. CVPR'14]

Frame 1

Train a bounding box regression model.

•••

Frame $t \geq 2$

Adjust the tracking result by bounding box regression.

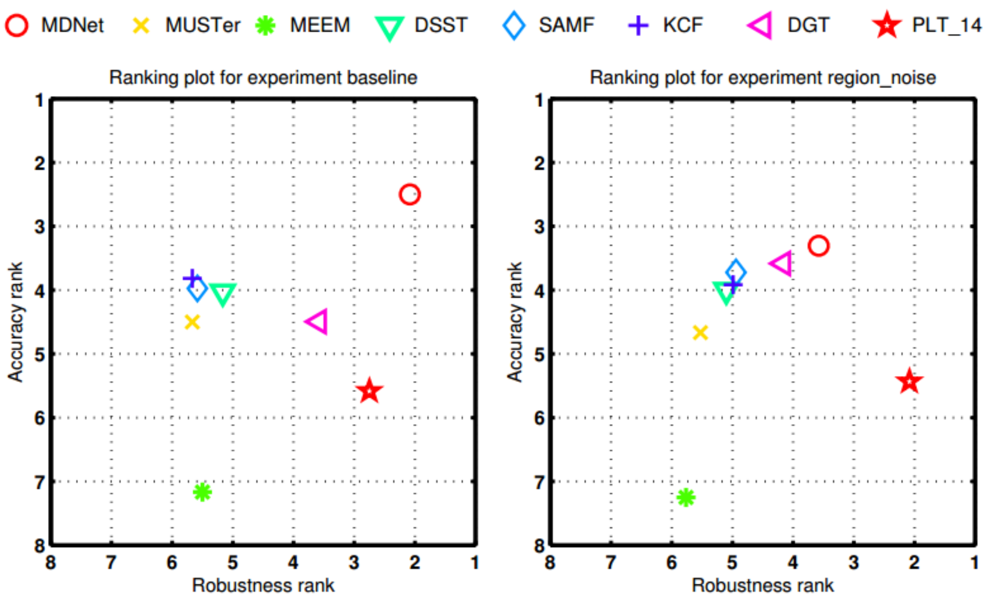
•••

Experimental Results

- Result on VOT2014 [Kristan et al. ECCVW'14]
- Result on OTB50 [Wu et al. CVPR'13]
- Result on OTB100 [Wu et al. PAMI'15]
- Component Analysis

Result on VOT2014 [Kristan et al. ECCVW'14]

- MDNet is trained with 89 sequences from {OTB100} excluding {VOT2014}
- Accuracy and robustness by baseline and region-noise experiments



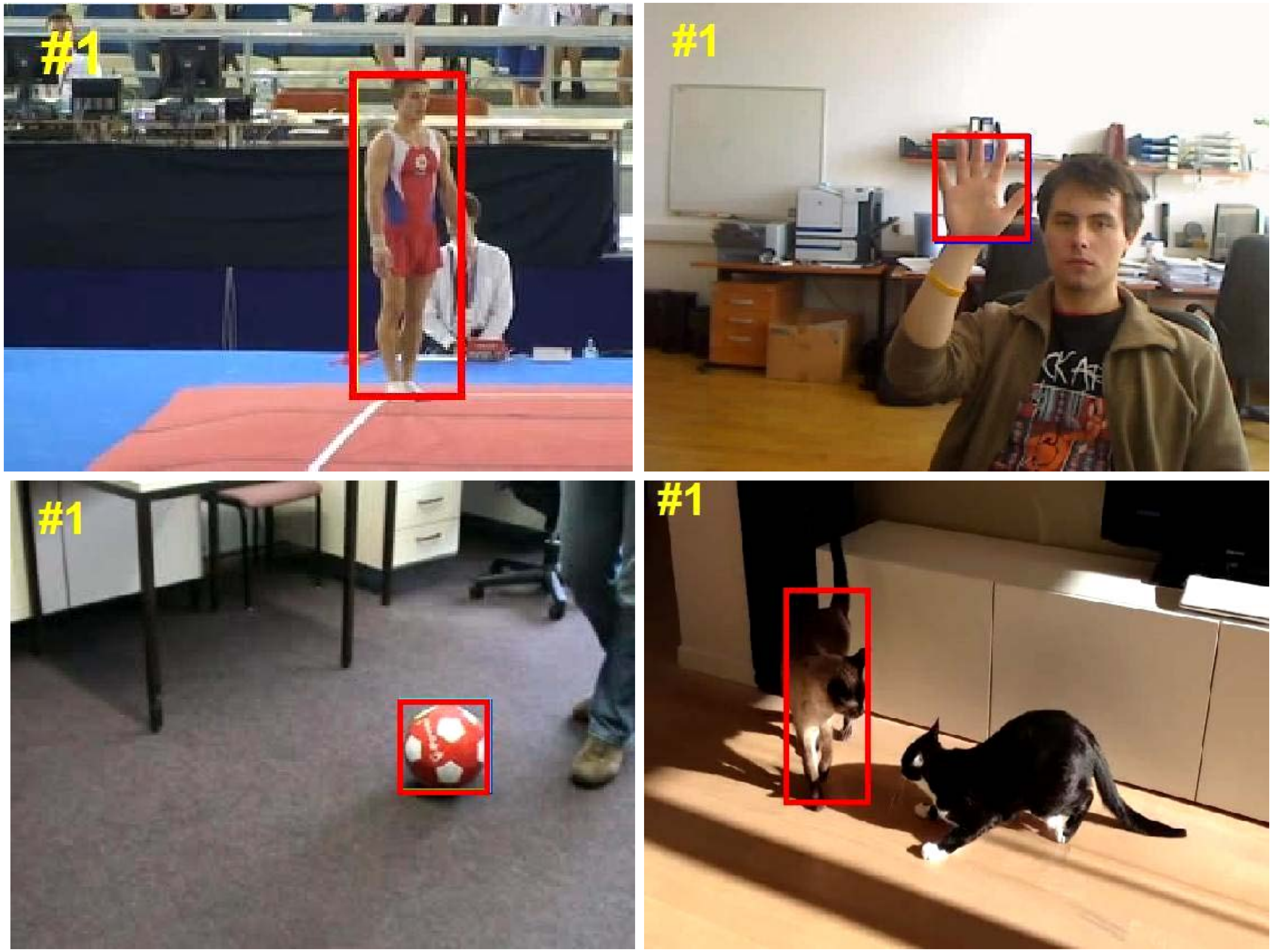
Tracker	Accuracy Score	Accuracy Rank	Robustness Score	Robustness Rank	Combined Rank
MUSTer	0.58	4.50	0.99	5.67	5.09
MEEM	0.48	7.17	0.71	5.50	6.34
DSST	0.60	4.03	0.68	5.17	4.60
SAMF	0.60	3.97	0.77	5.58	4.78
KCF	0.61	3.82	0.79	5.67	4.75
DGT	0.53	4.49	0.55	3.58	4.04
PLT_14	0.53	5.58	0.14	2.75	4.17
MDNet	0.63	2.50	0.16	2.08	2.29

(a) Baseline result

Tracker	Accuracy Score	Accuracy Rank	Robustness Score	Robustness Rank	Combined Rank
MUSTer	0.55	4.67	0.94	5.53	5.10
MEEM	0.48	7.25	0.74	5.76	6.51
DSST	0.58	4.00	0.76	5.10	4.55
SAMF	0.57	3.72	0.81	4.94	4.33
KCF	0.58	3.92	0.87	4.99	4.46
DGT	0.54	3.58	0.67	4.17	3.88
PLT_14	0.51	5.43	0.16	2.08	3.76
MDNet	0.60	3.31	0.30	3.58	3.45

(b) Region_noise result

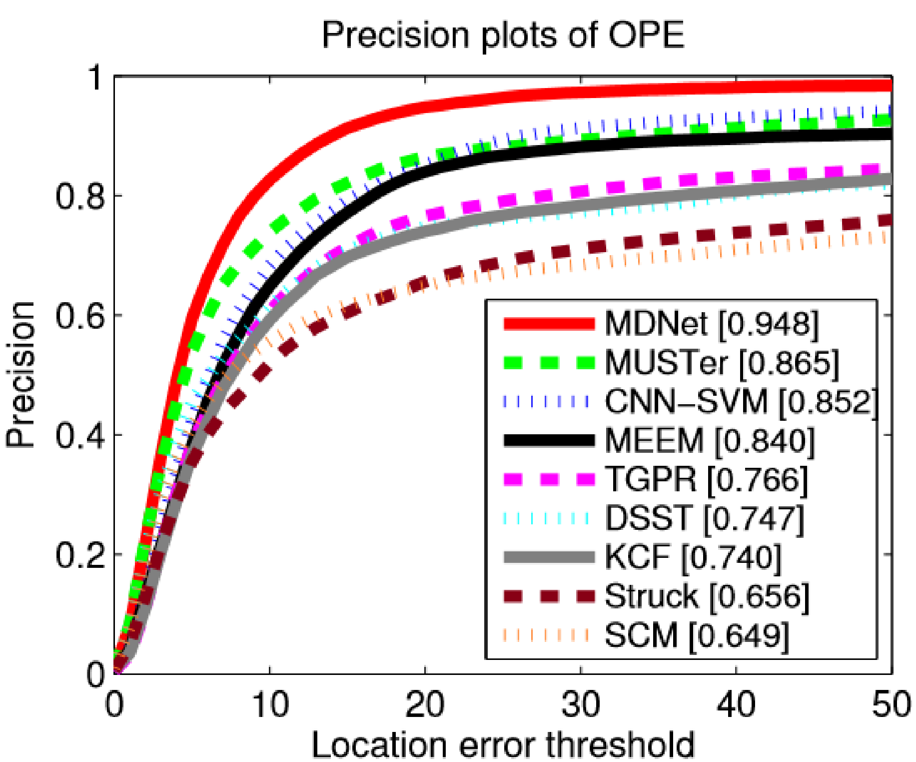
Qualitative Results on VOT2014 (w/o re-initialization)

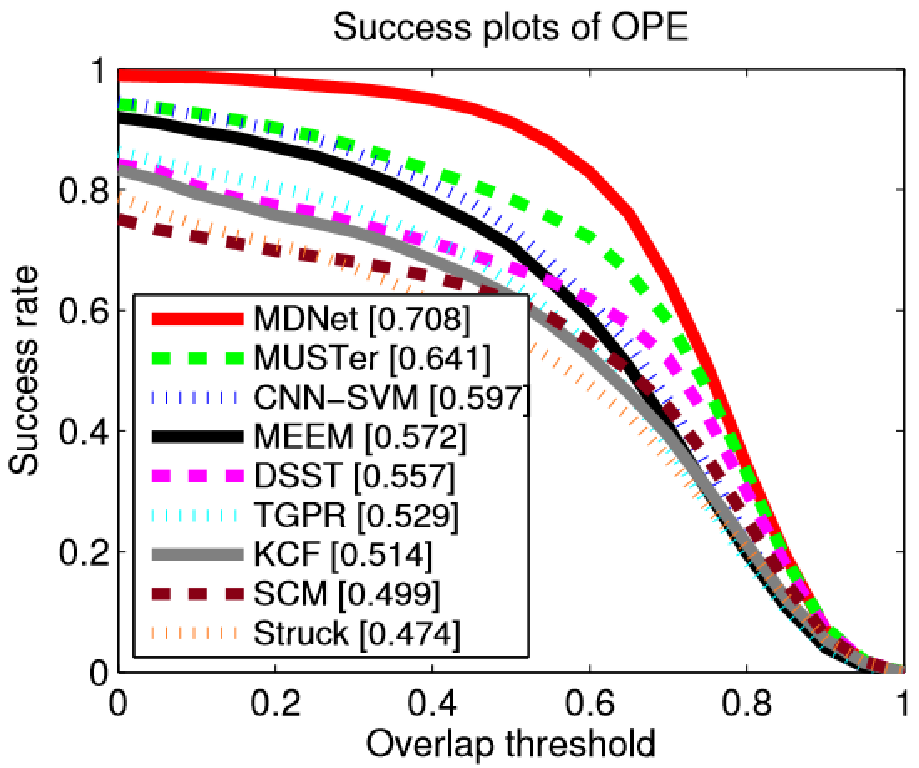


MDNet (Ours) MUSTer MEEM DSST KCF

Result on OTB50 [Wu et al. CVPR'13]

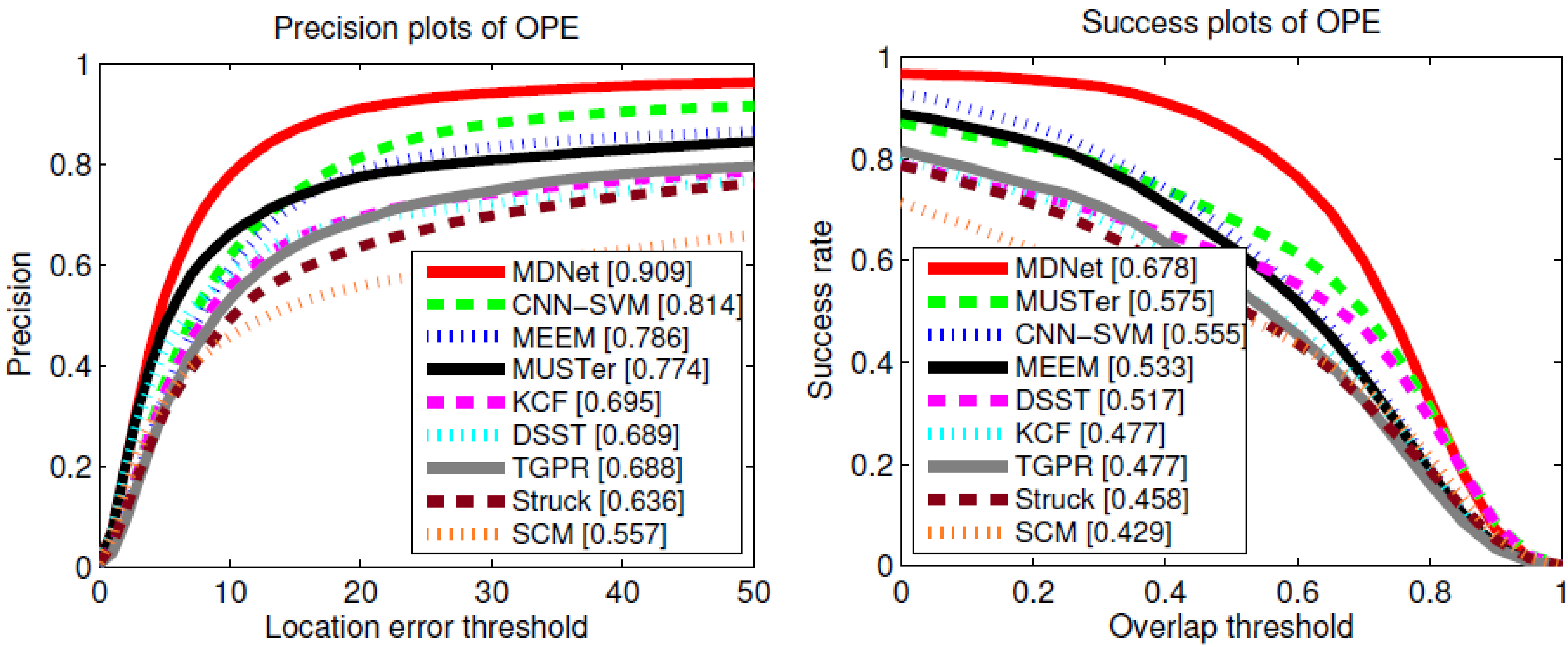
- MDNet is trained with 58 sequences from {VOT'13,'14,'15} excluding {OTB100}
- Distance precision and overlap success rate by One-Pass-Evaluation (OPE)



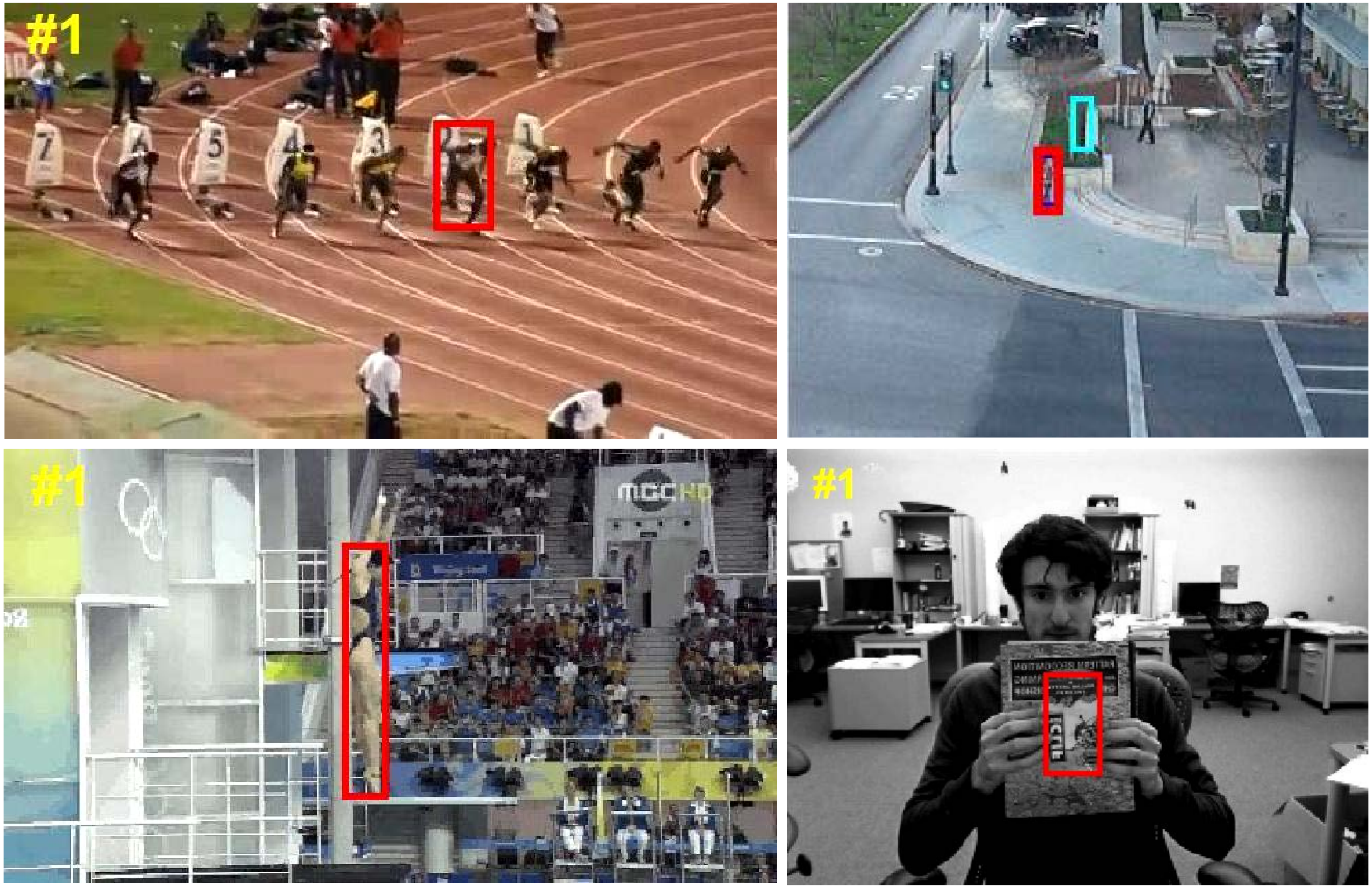


Result on OTB100 [Wu et al. TPAMI'15]

- MDNet is trained with 58 sequences from {VOT'13,'14,'15} excluding {OTB100}
- Distance precision and overlap success rate by One-Pass-Evaluation (OPE)



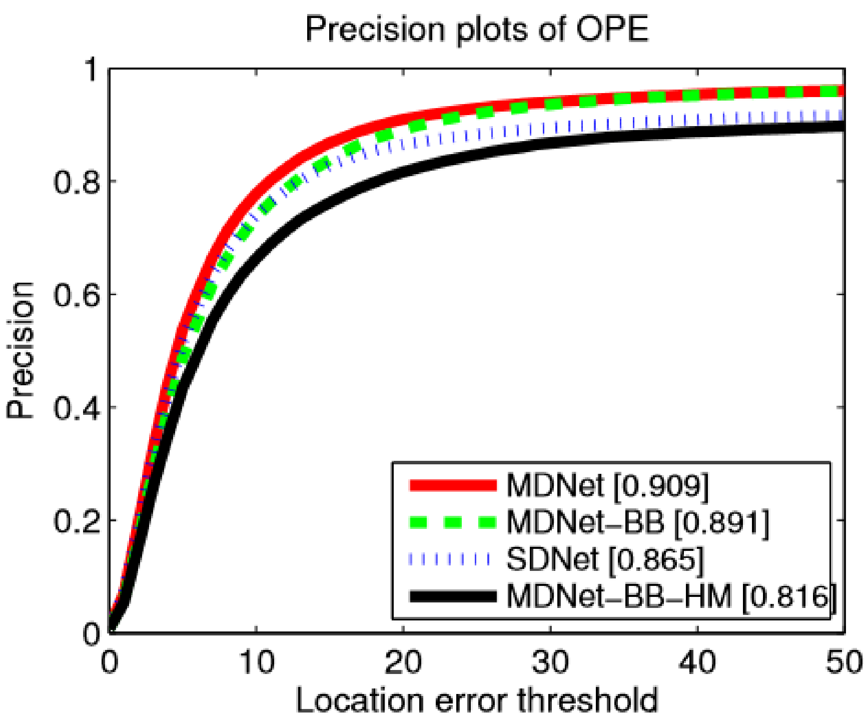
Qualitative Results on OTB100

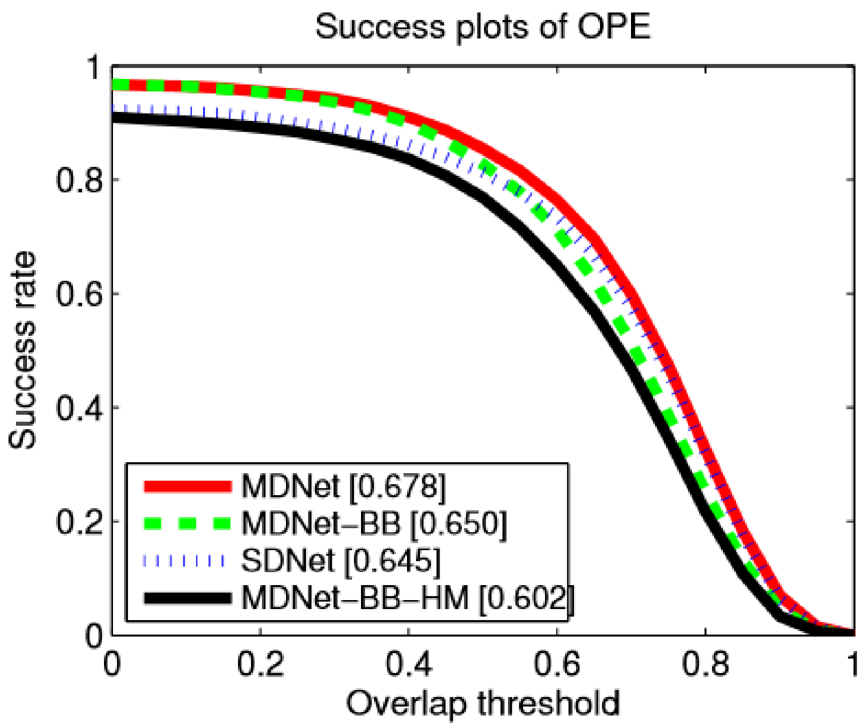


MDNet (Ours) — red; MUSTer — green; MEEM — yellow; DSST — cyan; CNN-SVM — blue

Component Analysis (OTB100)

- Our method (**MDNet**) is compared with
  - **SDNet**: pretrained by a single-domain network
  - **MDNet-BB**: MDNet w/o bounding box regression
  - **MDNet-BB-HM**: MDNet w/o bounding box regression & hard minibatch mining





Summary

- MDNet for learning generic features for visual tracking

- Online tracking algorithm by transferring MDNet features
  - Complementary network update
  - Hard negative mining
  - Bounding box regression

- Outstanding Performance in VOT2014, OTB50 and OTB100

- The Best Submitted Tracker on VOT2015 Challenge!

For More Details…

- Please refer to our arXiv paper.

arXiv.org > cs > arXiv:1510.07945

Search or Article-id (Help | Advanced search)

All papers Go!

Computer Science > Computer Vision and Pattern Recognition

Learning Multi-Domain Convolutional Neural Networks for Visual Tracking

Hyeonseob Nam, Bohyung Han

(Submitted on 27 Oct 2015)

We propose a novel visual tracking algorithm based on the representations from a discriminatively trained Convolutional Neural Network (CNN). Our algorithm pretrains a CNN using a large set of videos with tracking ground-truths to obtain a generic target representation. Our network is composed of shared layers and multiple branches of domain-specific layers, where domains correspond to individual training sequences and each branch is responsible for binary classification to identify target in each domain. We train each domain in the network iteratively to obtain generic target representations in the shared layers. When tracking a target in a new sequence, we construct a new network by combining the shared layers in the pretrained CNN with a new binary classification layer, which is updated online. Online tracking is performed by evaluating the candidate windows randomly sampled around the previous target state. The proposed algorithm illustrates outstanding performance in existing tracking benchmarks.

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**
Cite as: arXiv:1510.07945 **[cs.CV]**
(or arXiv:1510.07945v1 **[cs.CV]** for this version)

Submission history

From: Hyeonseob Nam [view email]
[v1] Tue, 27 Oct 2015 15:53:00 GMT (3024kb,D)

Download:

- PDF
- Other formats

(license)

Current browse context:
cs.CV
< prev | next >
new | recent | 1510

Change to browse by:
cs

References & Citations
- NASA ADS

DBLP - CS Bibliography
listing | bibtex
Hyeonseob Nam
Bohyung Han

Bookmark (what is this?)

- Code and results will be uploaded soon.
  - http://cvlab.postech.ac.kr