MCMOT: Multi-Class Multi-Object Tracking using Changing Point Detection

ILSVRC 2016 Object Detection from Video

Byungjae Lee[1], Songguo Jin[1], Enkhbayar Erdenee[1], Mi Young Nam[2], Young Gui Jung[2], Phill Kyu Rhee[1]

Inha University[1], **NaeulTech**[2]

ECCV'16 EUROPEAN CONFERENCE ON COMPUTER VISION

ITLAB Intelligence Technology Laboratory

Results with additional training data

- Object Detection from Video (VID)
  2nd place (mAP: 73.15%)

- Object Detection/Tracking from Video (VID)
  2nd place (mAP: 49.09%)

Overview

I. Faster R-CNN Object Detector

+ Context region
+ Larger feature map
+ Ensemble
+ Data configuration

II. MCMOT: Multi-Class Multi-Object Tracking

- Tracking by Detection
- Detection: Ensemble of CNNs
- Tracking: MCMOT using CPD

S. Ren, K. He, R. Girshick, & J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". TPAMI 2016.

B. Lee, E. Erdenee, S. Jin, & P. Rhee. "Multi-Class Multi-Object Tracking using Changing Point Detection". arXiv 2016.

I. Faster R-CNN Object Detector

S. Ren, K. He, R. Girshick, & J. Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. TPAMI 2016.

Object Detection from Video

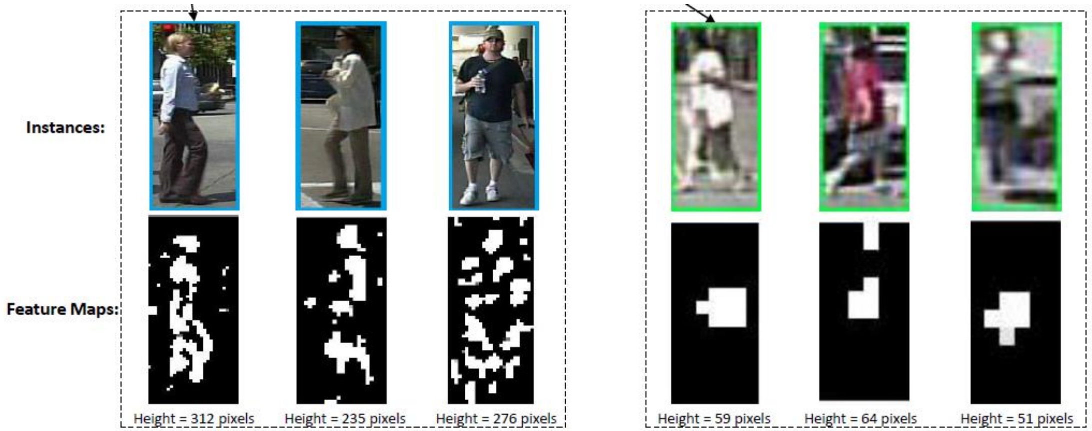
Challenge I: Small Object

* Figure from J. Li et al. "Scale-aware Fast R-CNN for Pedestrian Detection". arXiv 2015.

Object Detection from Video

Challenge I: Small Object

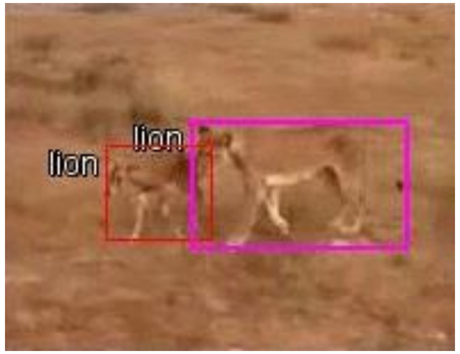
Solution – Larger feature map

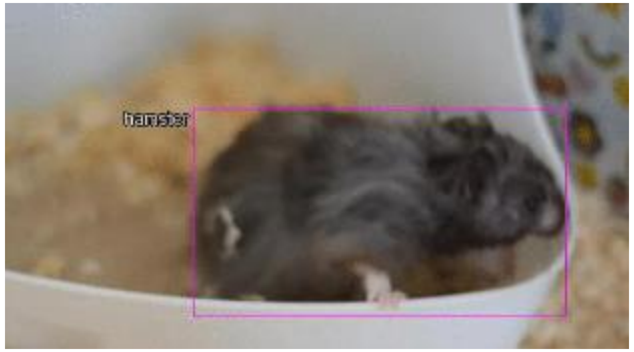


* Figure from J. Li et al. “Scale-aware Fast R-CNN for Pedestrian Detection”. arXiv 2015.

Object Detection from Video

Challenge II: Blurred Object

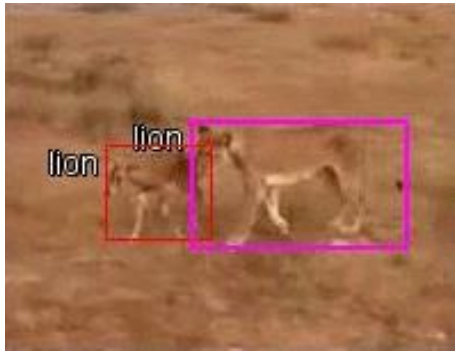


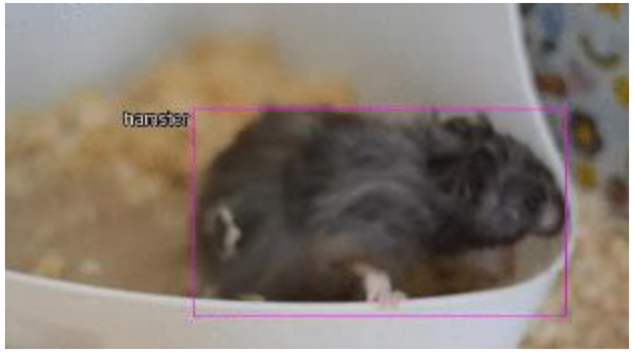


* Figure from Y. Zhu et al. "segDeepM: Exploiting Segmentation and Context in Deep Neural Networks for Object Detection". CVPR 2015.

Object Detection from Video

Challenge II: Blurred Object



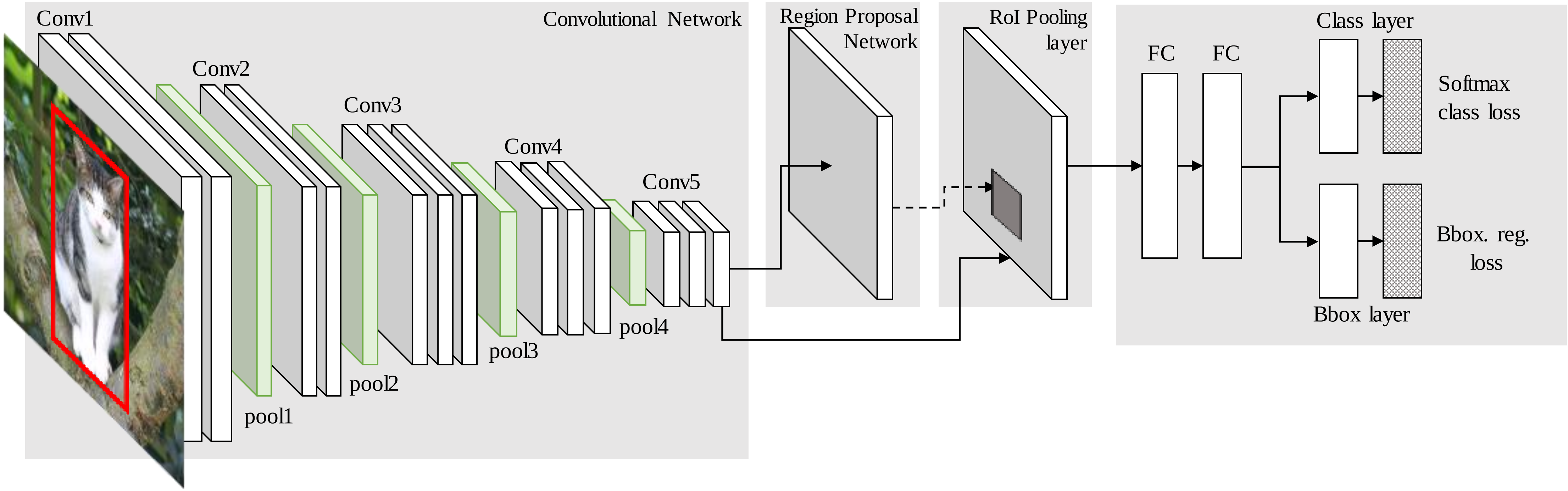


Solution – Context Region

* Figure from Y. Zhu et al. "segDeepM: Exploiting Segmentation and Context in Deep Neural Networks for Object Detection". CVPR 2015.

Network Architecture

I. Faster R-CNN with VGG16

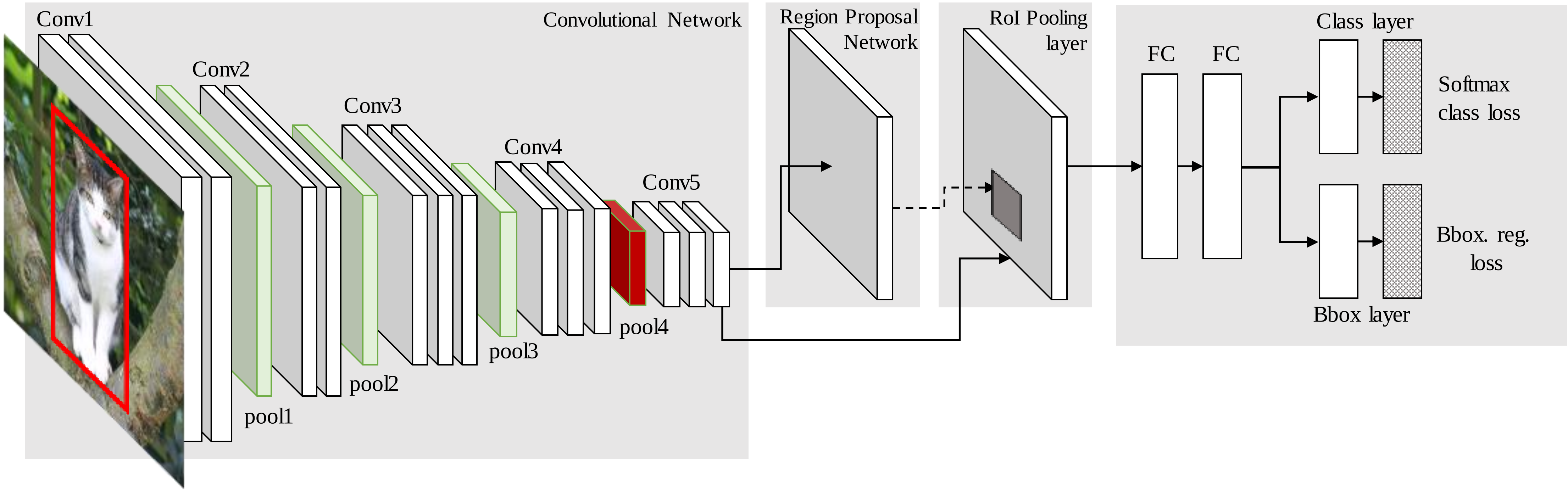


Karen Simonyan, & Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". arXiv 2015.

S. Ren, K. He, R. Girshick, & J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". TPAMI 2016.

Network Architecture

I. Faster R-CNN with VGG16



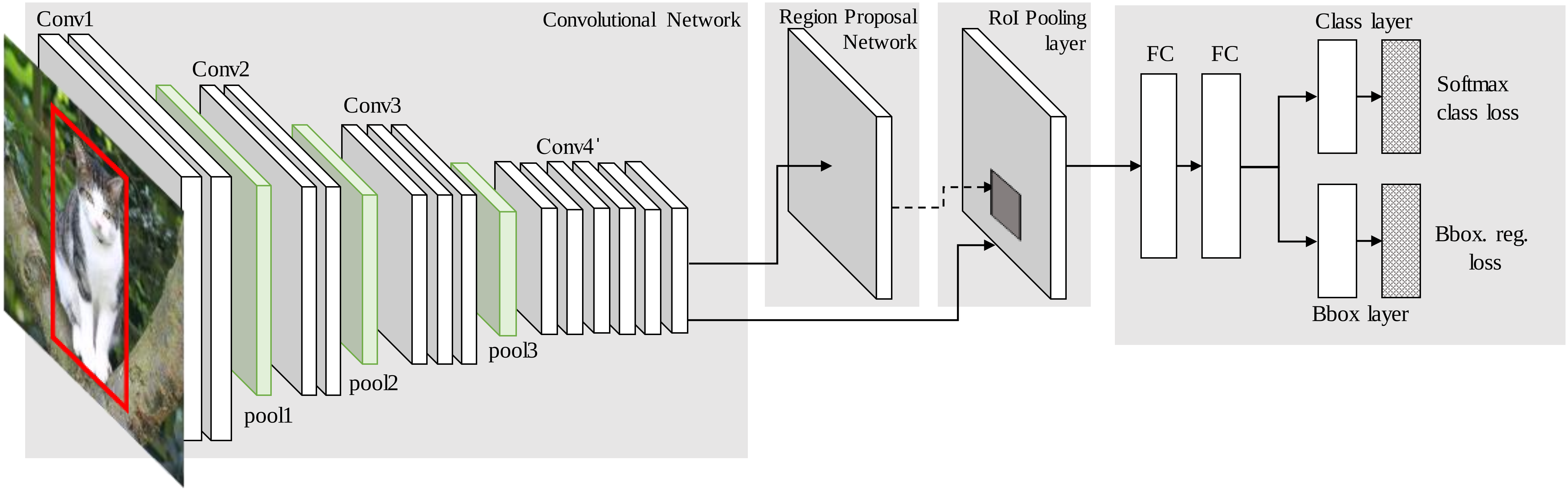
Karen Simonyan, & Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. arXiv 2015.

S. Ren, K. He, R. Girshick, & J. Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. TPAMI 2016.

Network Architecture

I. Faster R-CNN with VGG16

+ Larger feature map (remove 'pool4' layer) (+2.9% mAP)



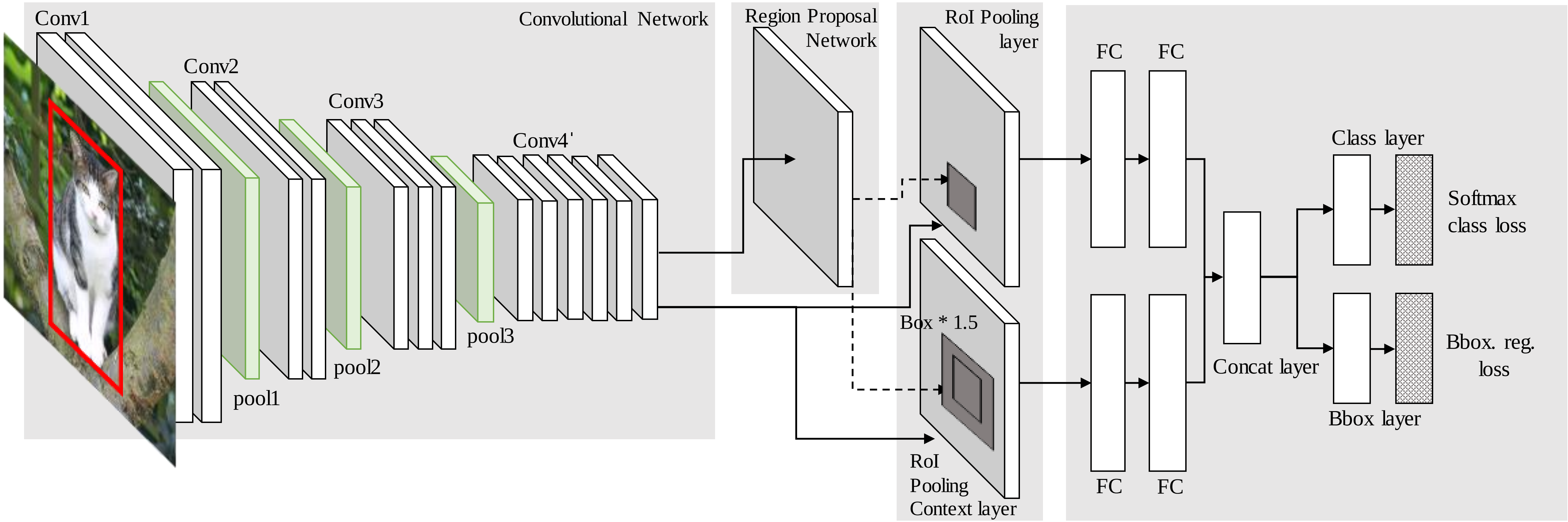
J. Li, X. Liang, S. Shen, T. Xu, & S. Yan. "Scale-aware Fast R-CNN for Pedestrian Detection". arXiv 2015.

S. Ren, K. He, R. Girshick, & J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". TPAMI 2016.

Network Architecture

I. Faster R-CNN with VGG16

- \+ Larger feature map (remove 'pool4' layer) (+2.9% mAP)
- \+ Context region (+2.6% mAP)



S. Gidaris, & N. Komodakis. "Object detection via a multi-region & semantic segmentation-aware CNN model". CVPR 2015.

S. Ren, K. He, R. Girshick, & J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". TPAMI 2016.

Training Data Configuration

Overview of ILSVRC VID Dataset

ILSVRC VID	Training	Validation
Images	1122397	176126
Snippets	3862	555

* Images are from the ILSVRC VID dataset

Training Data Configuration

Overview of ILSVRC VID Dataset

ILSVRC VID	Training	Validation
Images	1122397	176126
Snippets	3862	555

- **Redundant images** within each snippet
- **Diversity is too low** to train CNN

* Images are from the ILSVRC VID dataset

Training Data Configuration

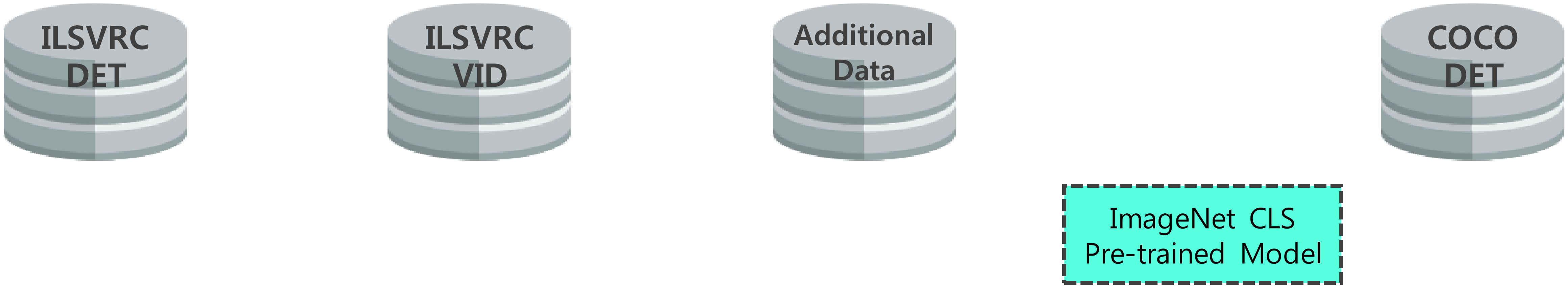
Overview of ILSVRC VID Dataset

ILSVRC VID	Training	Validation
Images	1122397	176126
Snippets	3862	555

- **Redundant images** within each snippet
- **Diversity is too low** to train CNN
- **We need more data**

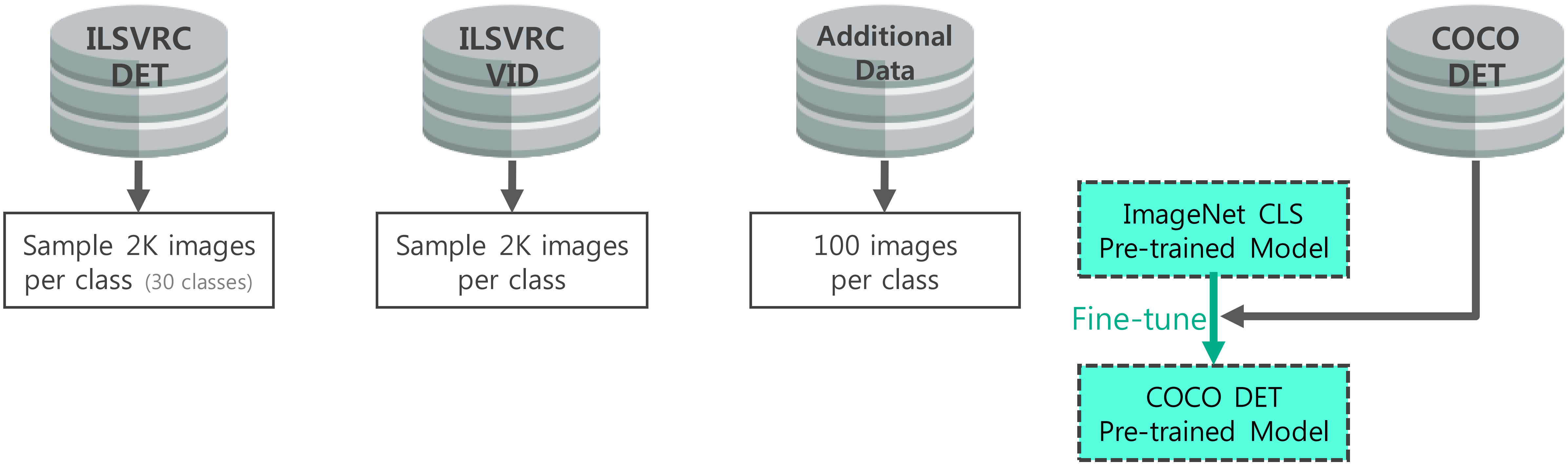
* Images are from the ILSVRC VID dataset

Training Data Configuration

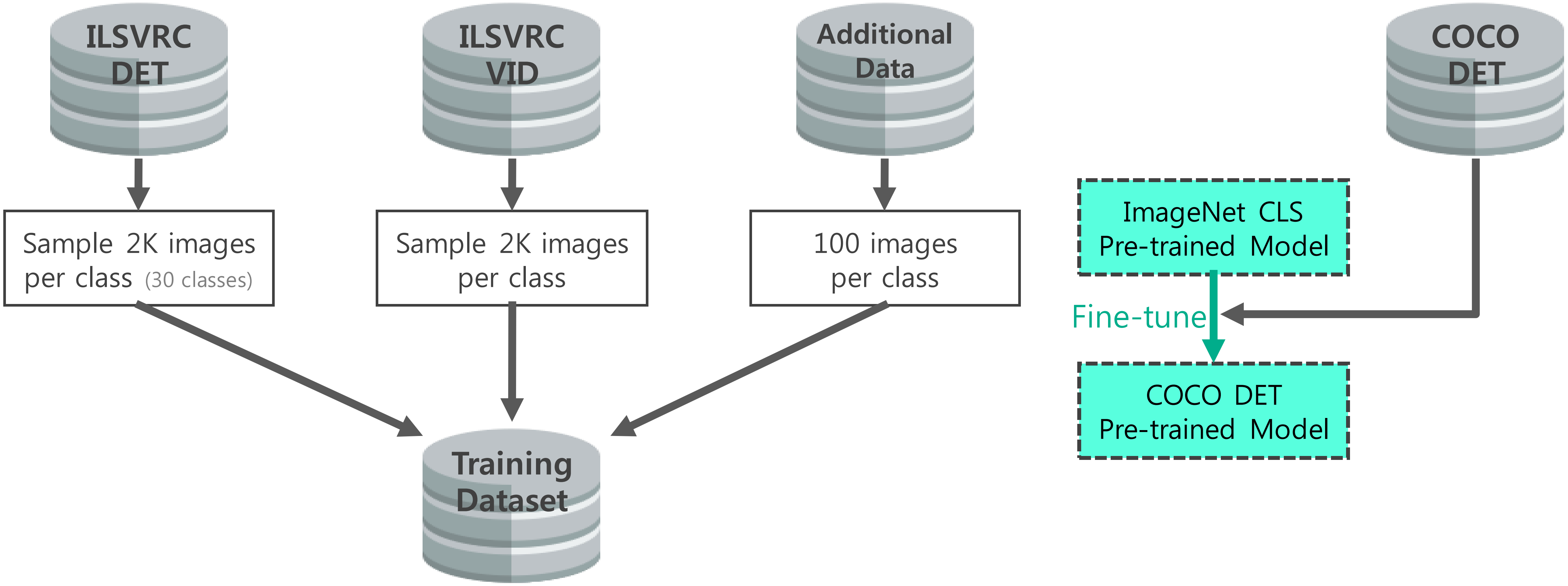


Training Data Configuration

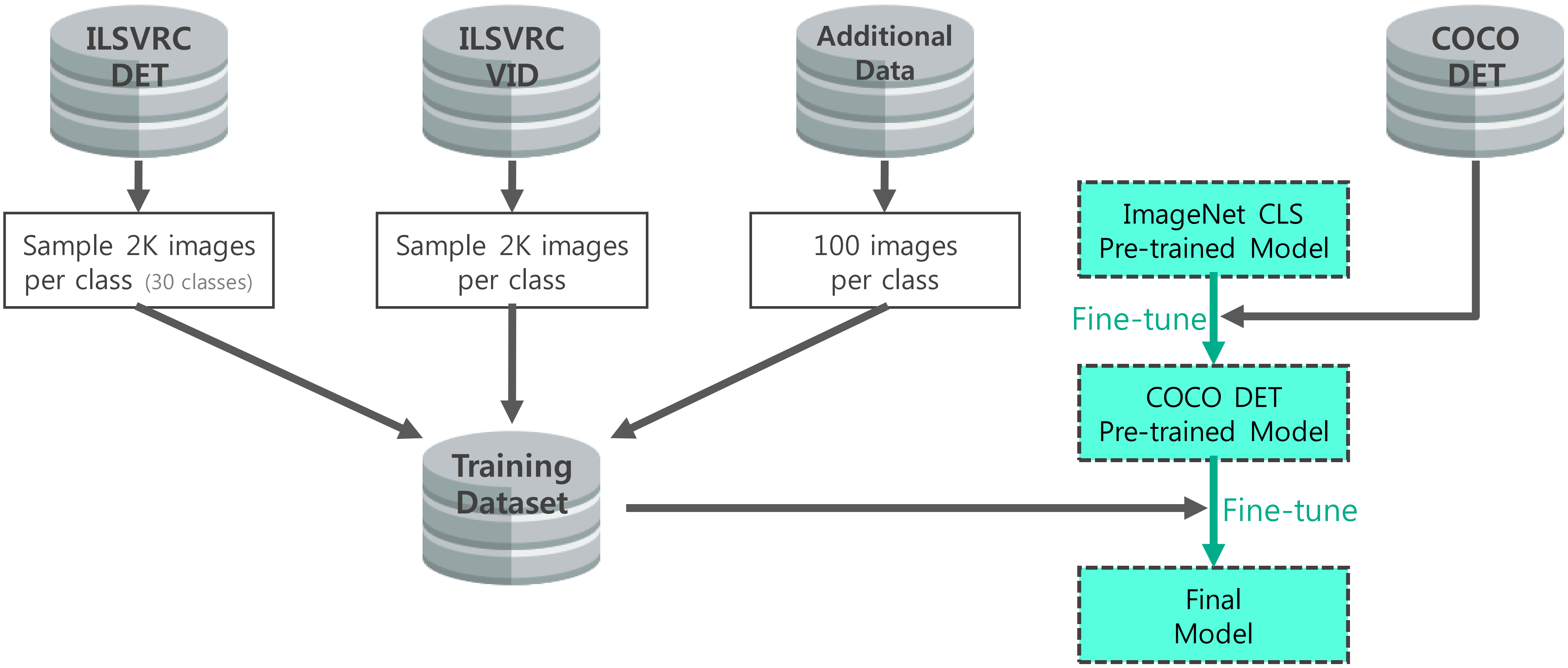
Training Data Configuration



Training Data Configuration



Training Data Configuration



Detection Components

VGG 16	mAP(%)
Baseline	70.7%

ResNet-101	mAP(%)
Baseline	78.8%

* We used 12 anchors for RPN

* Performances are evaluated on VID minival (only used initial 30 frames per each snippet, total 16,524 images)

Detection Components

VGG 16	mAP(%)
Baseline	70.7%
+ Larger feature map	73.6%
+ Context layer	76.2%

ResNet-101	mAP(%)
Baseline	78.8%

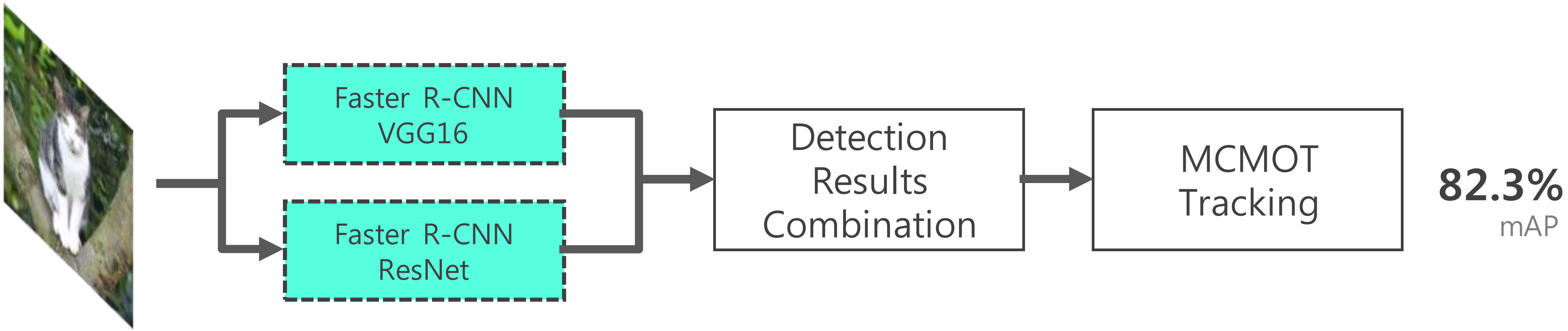
* We used 12 anchors for RPN

* Performances are evaluated on VID minival (only used initial 30 frames per each snippet, total 16,524 images)

Detection Components

VGG 16	mAP(%)
Baseline	70.7%
+ Larger feature map	73.6%
+ Context layer	76.2%

ResNet-101	mAP(%)
Baseline	78.8%



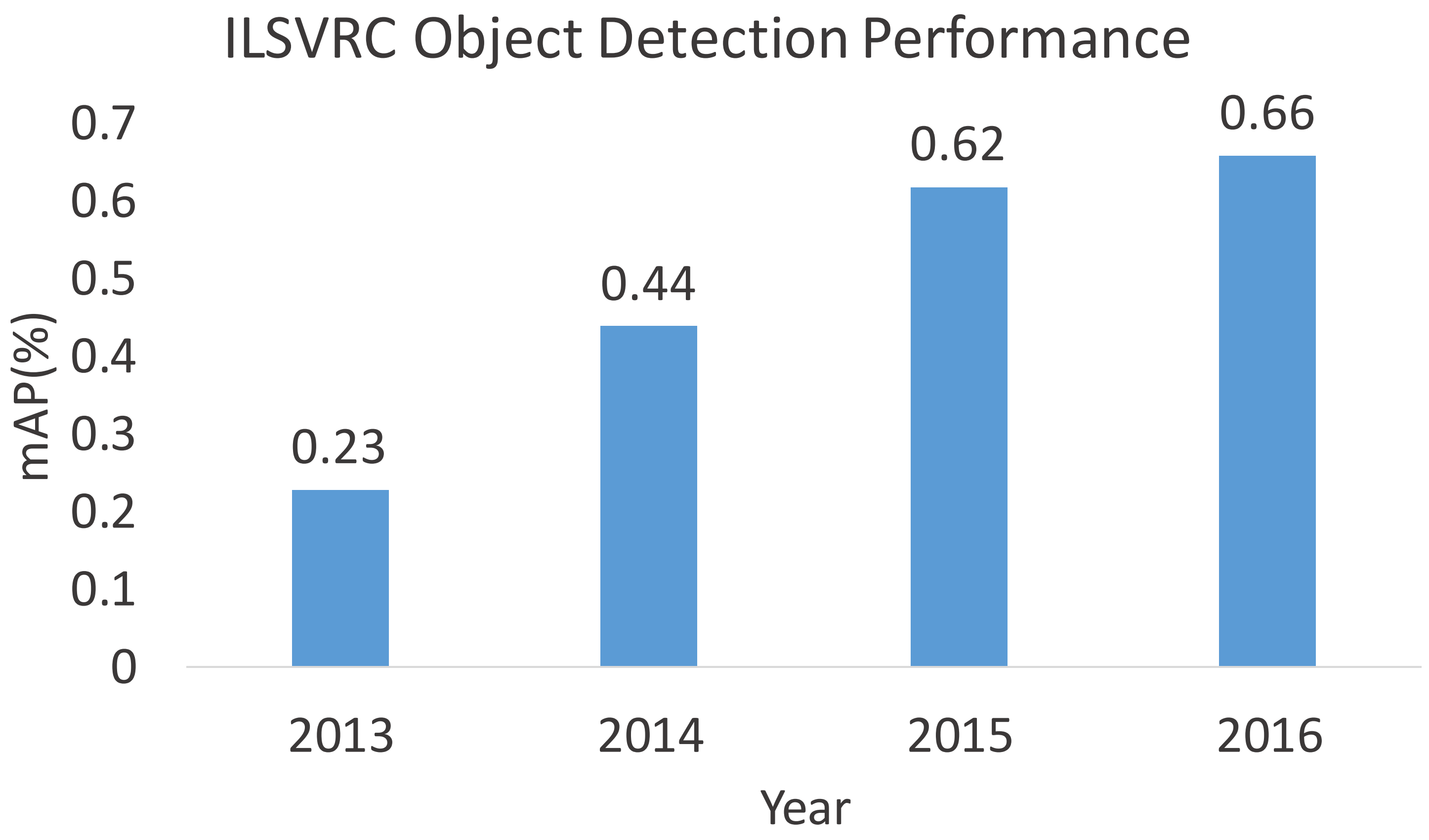
* We used 12 anchors for RPN

* Performances are evaluated on VID minival (only used initial 30 frames per each snippet, total 16,524 images)

Ⅱ. MCMOT: Multi-Class Multi-Object Tracking using Changing Point Detection

B. Lee, E. Erdenee, S. Jin, & P. Rhee. “Multi-Class Multi-Object Tracking using Changing Point Detection”. arXiv 2016.

Motivation



- Object detector becomes robust
- Should we use **complex** multi-object tracking algorithm?

Motivation

Based on high performance detection,
simple & fast MOT algorithm
can achieve competitive result

Results

- ILSVRC2016 Object Detection/Tracking from Video (VID) with additional training data
  2nd place (mAP: 49.09%)

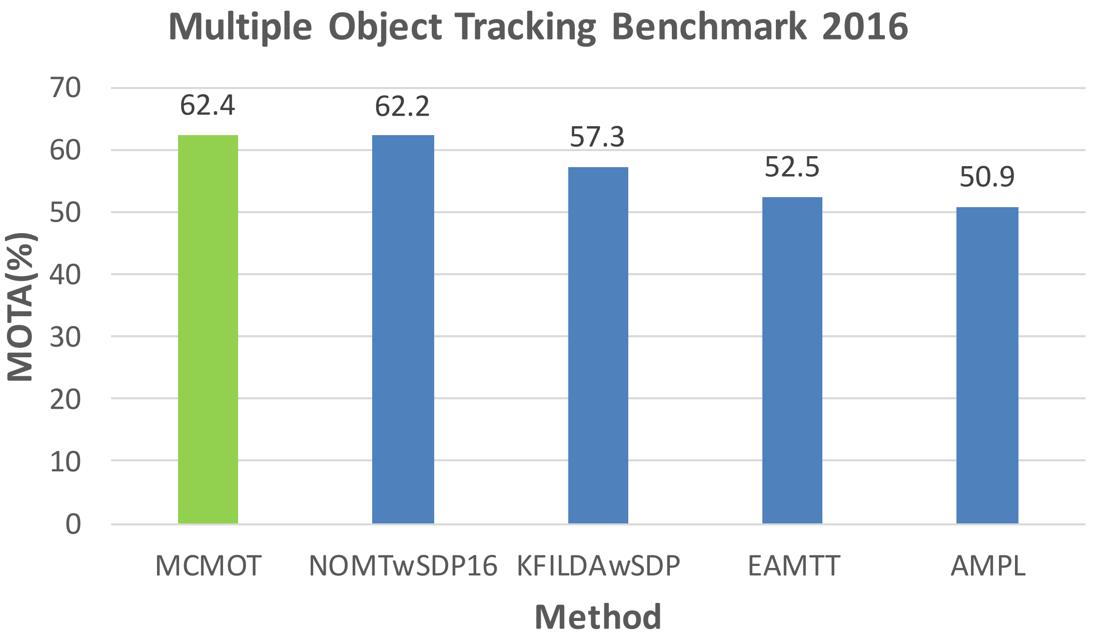
https://motchallenge.net/results/MOT16/

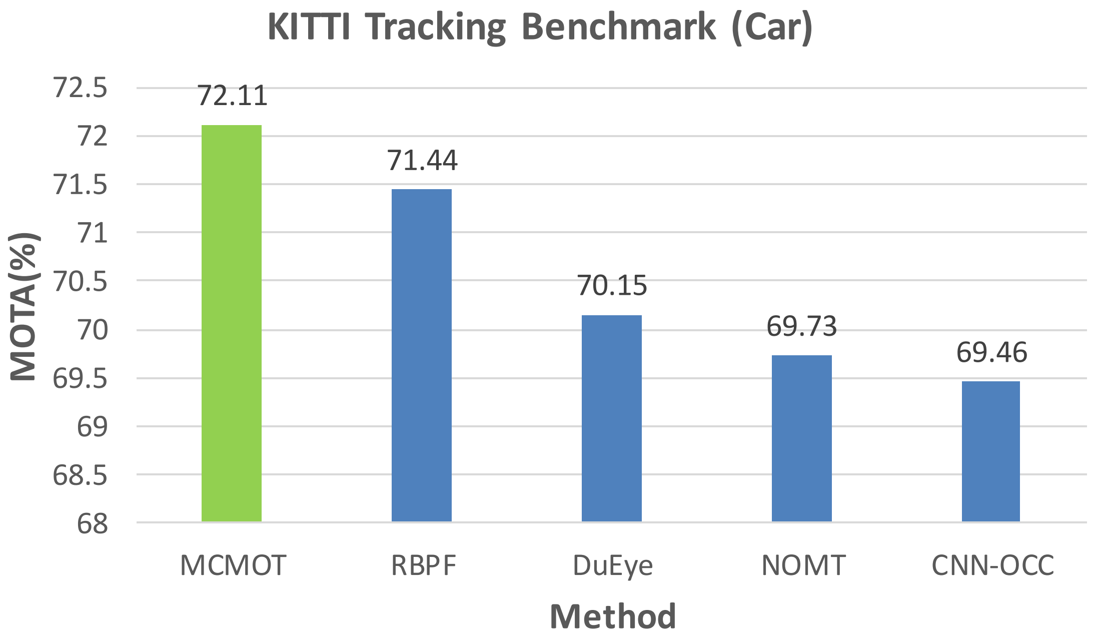
http://www.cvlibs.net/datasets/kitti/eval_tracking.php

B. Lee, E. Erdenee, S. Jin, & P. Rhee. "Multi-Class Multi-Object Tracking using Changing Point Detection". arXiv 2016.

Results

- ILSVRC2016 Object Detection/Tracking from Video (VID) with additional training data
  2nd place (mAP: 49.09%)
- Our MCMOT also achieves state-of-the-art results in different MOT datasets

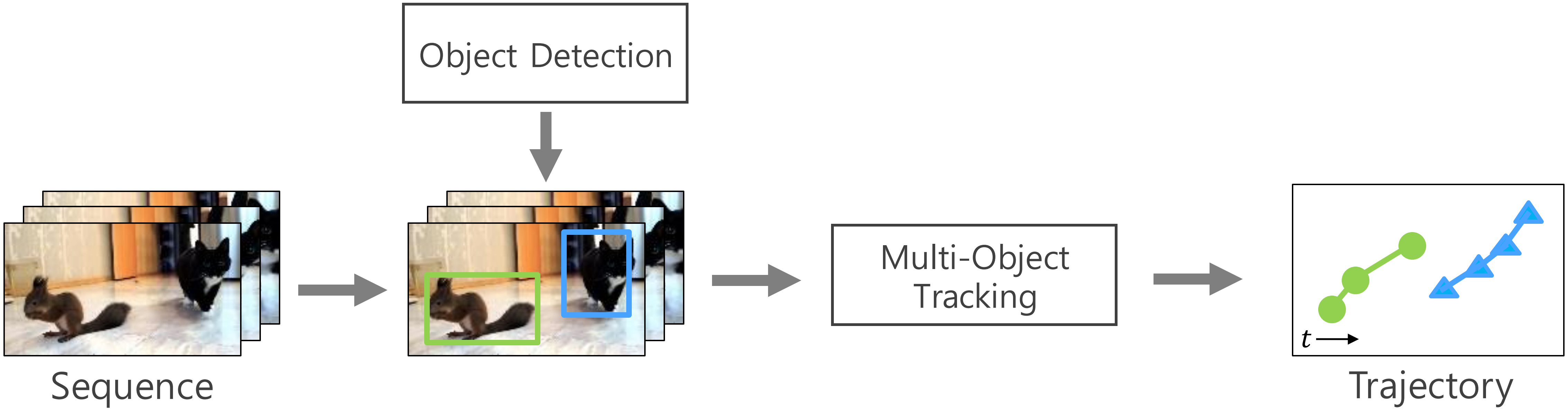




https://motchallenge.net/results/MOT16/
http://www.cvlibs.net/datasets/kitti/eval_tracking.php

B. Lee, E. Erdenee, S. Jin, & P. Rhee. "Multi-Class Multi-Object Tracking using Changing Point Detection". arXiv 2016.

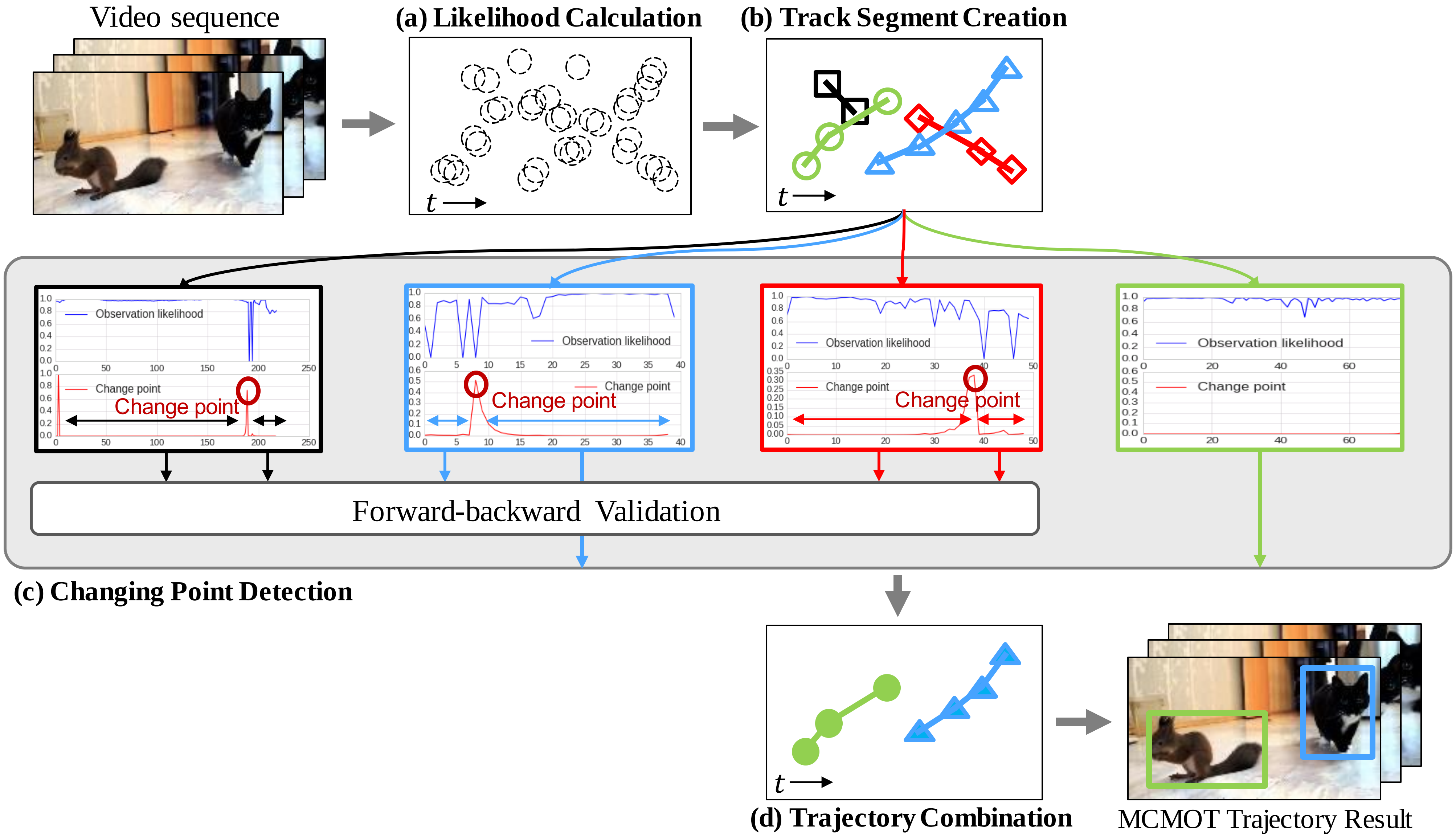
Tracking by Detection



- Object Detection: Ensemble of CNNs
- Multi-Object Tracking: MCMOT using CPD

B. Lee, E. Erdenee, S. Jin, & P. Rhee. “Multi-Class Multi-Object Tracking using Changing Point Detection”. arXiv 2016.

MCMOT using CPD



B. Lee, E. Erdenee, S. Jin, & P. Rhee. "Multi-Class Multi-Object Tracking using Changing Point Detection". arXiv 2016.

Track Segment Creation

- MCMC-based MOT approach

  Changing number of moving objects are challenging, which require **high computation overheads** due to a **high-dimensional state space**

- Separating motion dynamics

  The method separates the motion dynamic model of Bayesian filter into the **entity transitions** and **motion moves**
  No dimension variation in the iteration loop by separating the moves of birth and death

H. Sakaino. "Video-based tracking, learning, and recognition method for multiple moving objects". IEEE trans. On circuits and systems for video technology. 2013.

B. Lee, E. Erdenee, S. Jin, & P. Rhee. "Multi-Class Multi-Object Tracking using Changing Point Detection". arXiv 2016.

Track Segment Creation

- Estimation of entity state transition (Birth, Death)

  The entity transitions are modeled as the birth and death events
  We estimate the entity prior by data-driven approach,
  instead of the inside of MCMC loop

H. Sakaino. "Video-based tracking, learning, and recognition method for multiple moving objects". IEEE trans. On circuits and systems for video technology. 2013.

B. Lee, E. Erdenee, S. Jin, & P. Rhee. "Multi-Class Multi-Object Tracking using Changing Point Detection". arXiv 2016.

Track Segment Creation

- Separating motion dynamics

Pros

Since the Markov chain has no dimension variation in the iteration loop,
it can reach to stationary states with **less computation overhead**

Cons

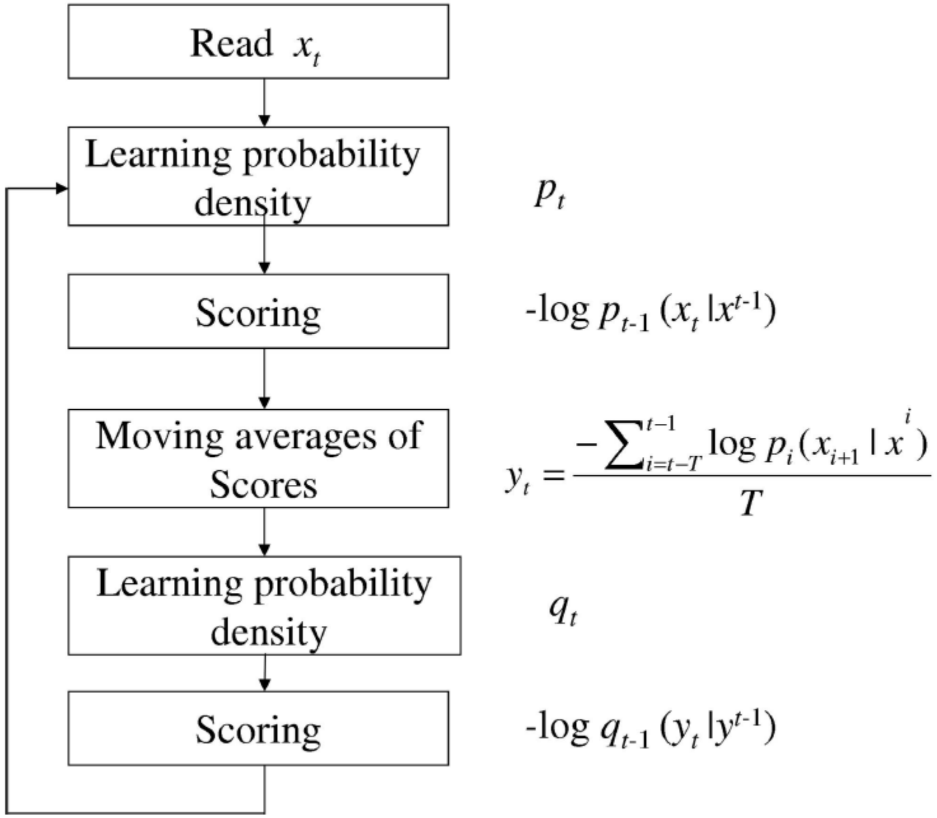
such a simple approach cannot deal with complex situations that occur in MOT
Many of them are **suffered from track drifts** due to appearance variations

Drift problem is attacked by a CPD algorithm

H. Sakaino. "Video-based tracking, learning, and recognition method for multiple moving objects". IEEE trans. On circuits and systems for video technology. 2013.
B. Lee, E. Erdenee, S. Jin, & P. Rhee. "Multi-Class Multi-Object Tracking using Changing Point Detection". arXiv 2016.

Changing Point Detection

- Two-Stage Learning for Changing Point Detection

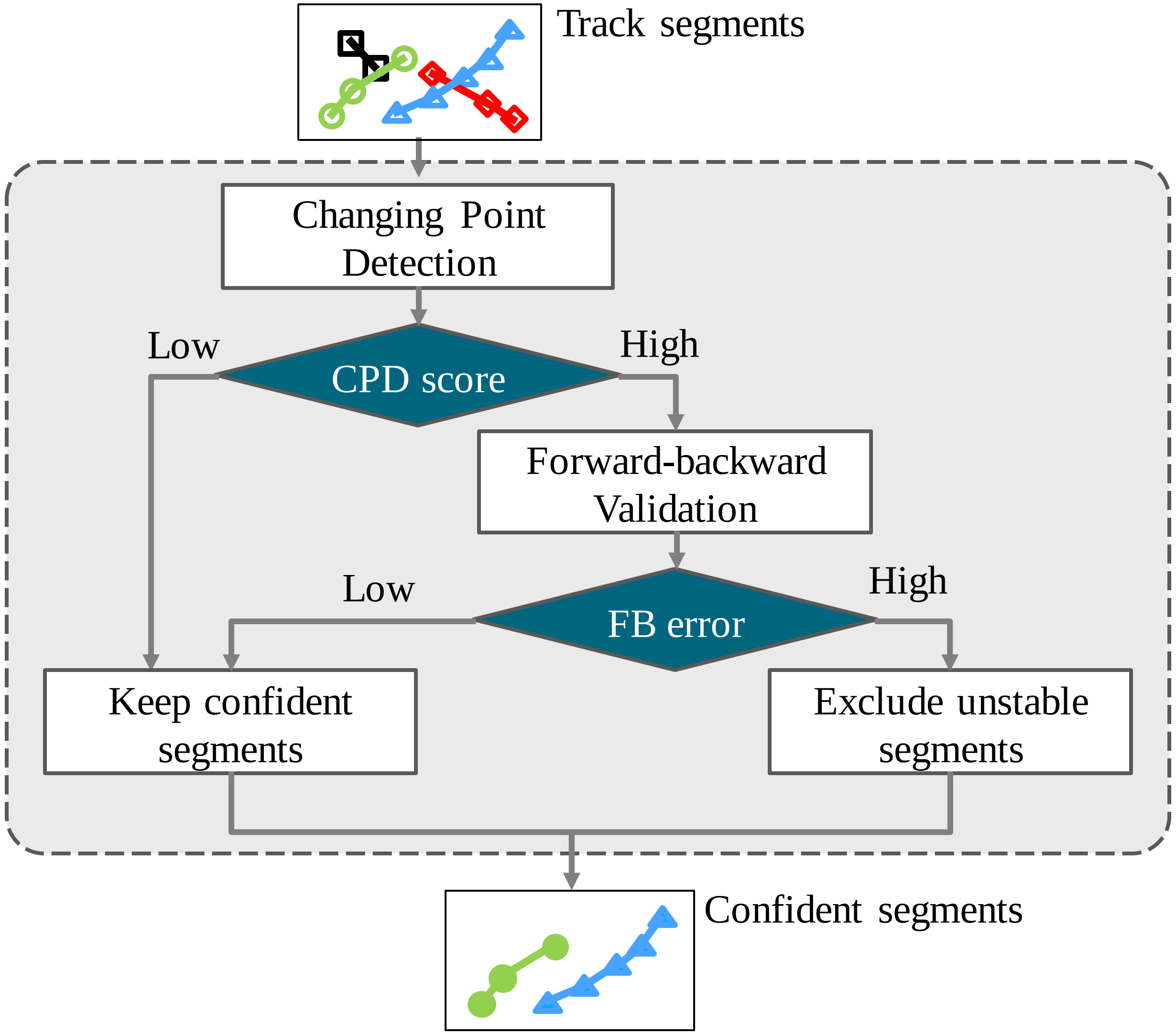


- Drifts in MCMOT are investigated by detection such **abrupt change points** between stationary time series that represent track segment
- **A possible track drift is determined** by a changing point detection
- 2nd level time series is built using the scanned average responses to reduce outliers in the time series

* Figure from J. Takeuchi, & K. Yamanishi. "A Unifying Framework for Detecting Outliers and Change Points from Time Series". TKDE 2006.

B. Lee, E. Erdenee, S. Jin, & P. Rhee. "Multi-Class Multi-Object Tracking using Changing Point Detection". arXiv 2016.

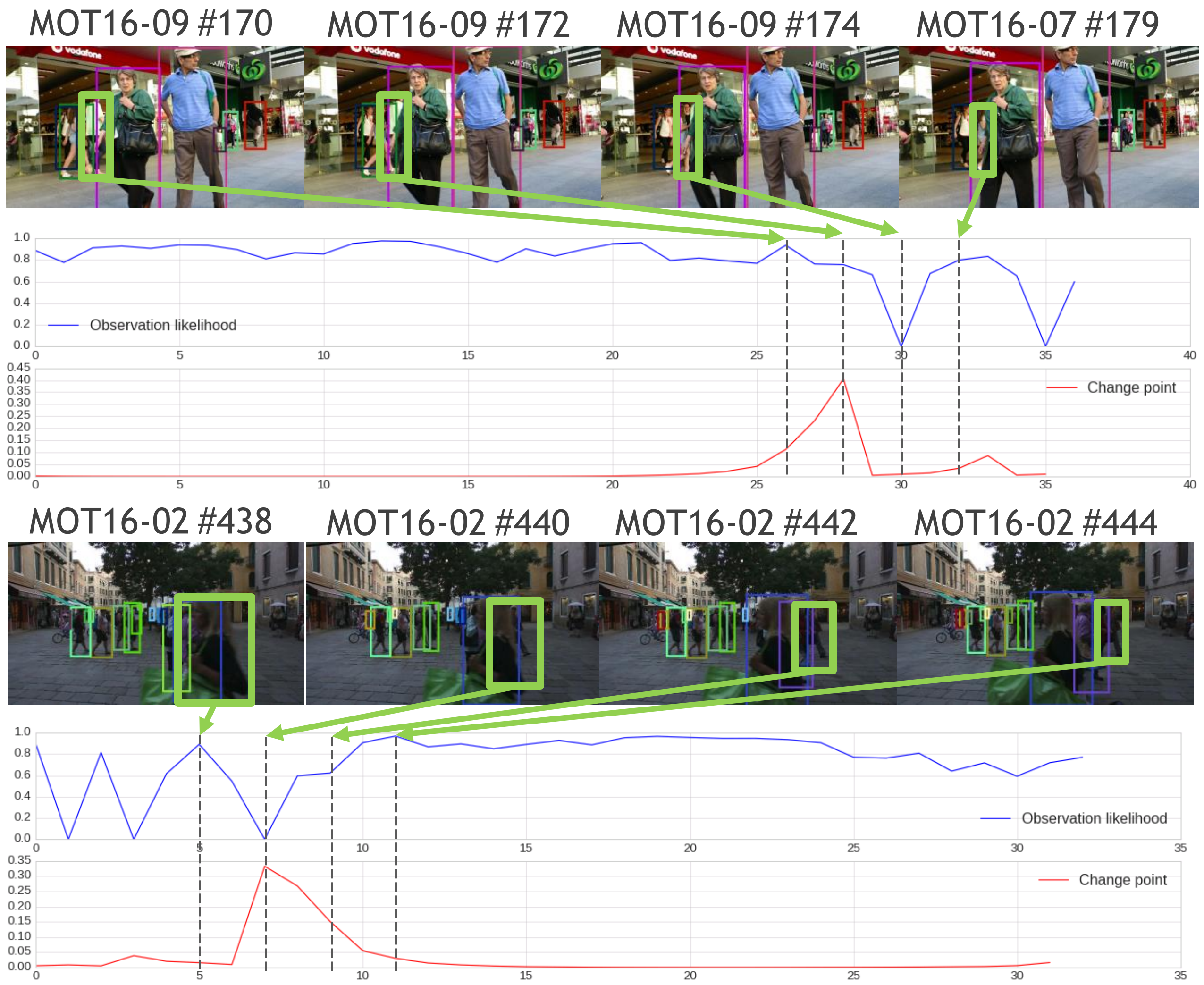
Changing Point Detection



J. Takeuchi, & K. Yamanishi. "A Unifying Framework for Detecting Outliers and Change Points from Time Series". TKDE 2006.

B. Lee, E. Erdenee, S. Jin, & P. Rhee. "Multi-Class Multi-Object Tracking using Changing Point Detection". arXiv 2016.

Changing Point Detection



* Images are from MOT 2016 benchmark

J. Takeuchi, & K. Yamanishi. "A Unifying Framework for Detecting Outliers and Change Points from Time Series". TKDE 2006.
B. Lee, E. Erdenee, S. Jin, & P. Rhee. "Multi-Class Multi-Object Tracking using Changing Point Detection". arXiv 2016.

Tracking Speed

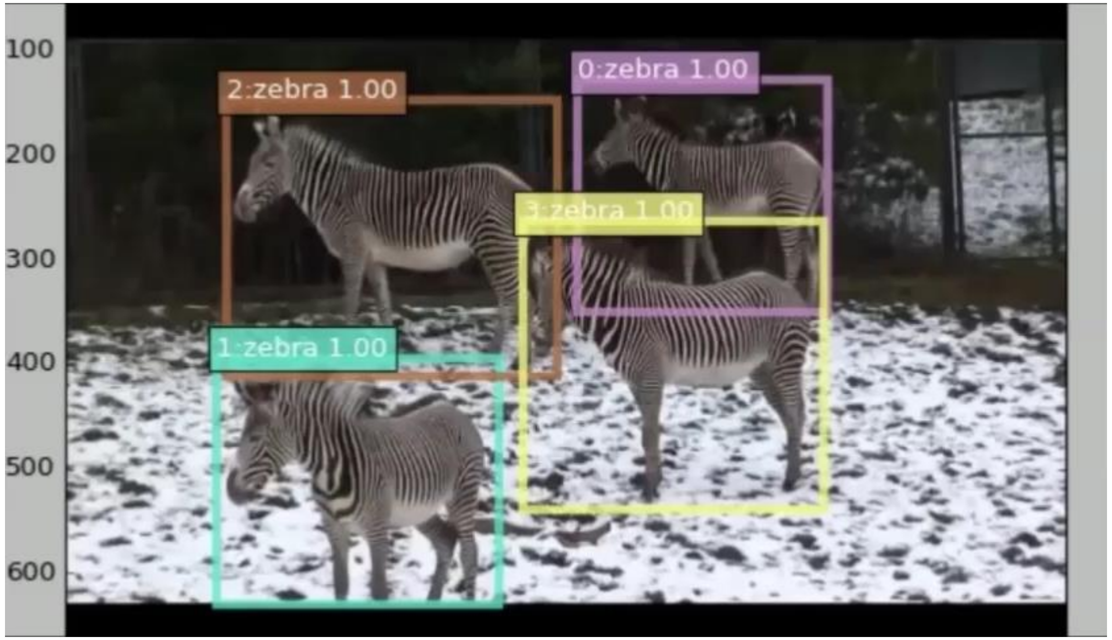
Method	MOTA↑	MOTP↑	FAF↓	MT↑	ML↓	FP↓	FN↓	ID Sw↓	Frag↓	Hz↑
GRIM	-14.5%	73.0%	10.0	9.9%	49.5%	59,040	147,908	1,869	2,454	10.0
JPDA_m	26.2%	76.3%	0.6	4.1%	67.5%	3,689	130,549	365	638	22.2
SMOT	29.7%	75.2%	2.9	5.3%	47.7%	17,426	107,552	3,108	4,483	0.2
DP_NMS	32.2%	76.4%	**0.2**	5.4%	62.1%	**1,123**	121,579	972	944	**212.6**
CEM	33.2%	75.8%	1.2	7.8%	54.4%	6,837	114,322	642	731	0.3
TBD	33.7%	76.5%	1.0	7.2%	54.2%	5,804	112,587	2,418	2,252	1.3
LINF1	41.0%	74.8%	1.3	11.6%	51.3%	7,896	99,224	430	963	1.1
olCF	43.2%	74.3%	1.1	11.3%	48.5%	6,651	96,515	381	1,404	0.4
NOMT	46.4%	76.6%	1.6	18.3%	41.4%	9,753	87,565	359	**504**	2.6
AMPL	50.9%	77.0%	0.5	16.7%	40.8%	3,229	86,123	**196**	639	1.5
NOMTwSDP16	62.2%	**79.6%**	0.9	**32.5%**	31.1%	5,119	63,352	406	642	3.1
MCMOT_HDM (Ours)	**62.4%**	78.3%	1.7	31.5%	**24.2%**	9,855	**57,257**	1,394	1,318	34.9

Tracking performances comparison on the MOT benchmark 2016

- The timing excludes detection time
- With a Titan X Maxwell GPU, the detector runs at approximately 3.5 FPS

B. Lee, E. Erdenee, S. Jin, & P. Rhee. "Multi-Class Multi-Object Tracking using Changing Point Detection". arXiv 2016.

Results



ImageNet VID

MOT Benchmark

KITTI Tracking Benchmark

Acknowledgement

Thank You

Email: byungjae89@gmail.com