Improving Context Modeling for Video Object Detection and Tracking

National University of Singapore:

Yunchao Wei, Mengdan Zhang, Jianan Li, Yunpeng Chen, Jiashi Feng

University of Illinois at Urbana-Champaign

Honghui Shi

Qihoo 360 AI Institute:

Jian Dong, Shuicheng Yan

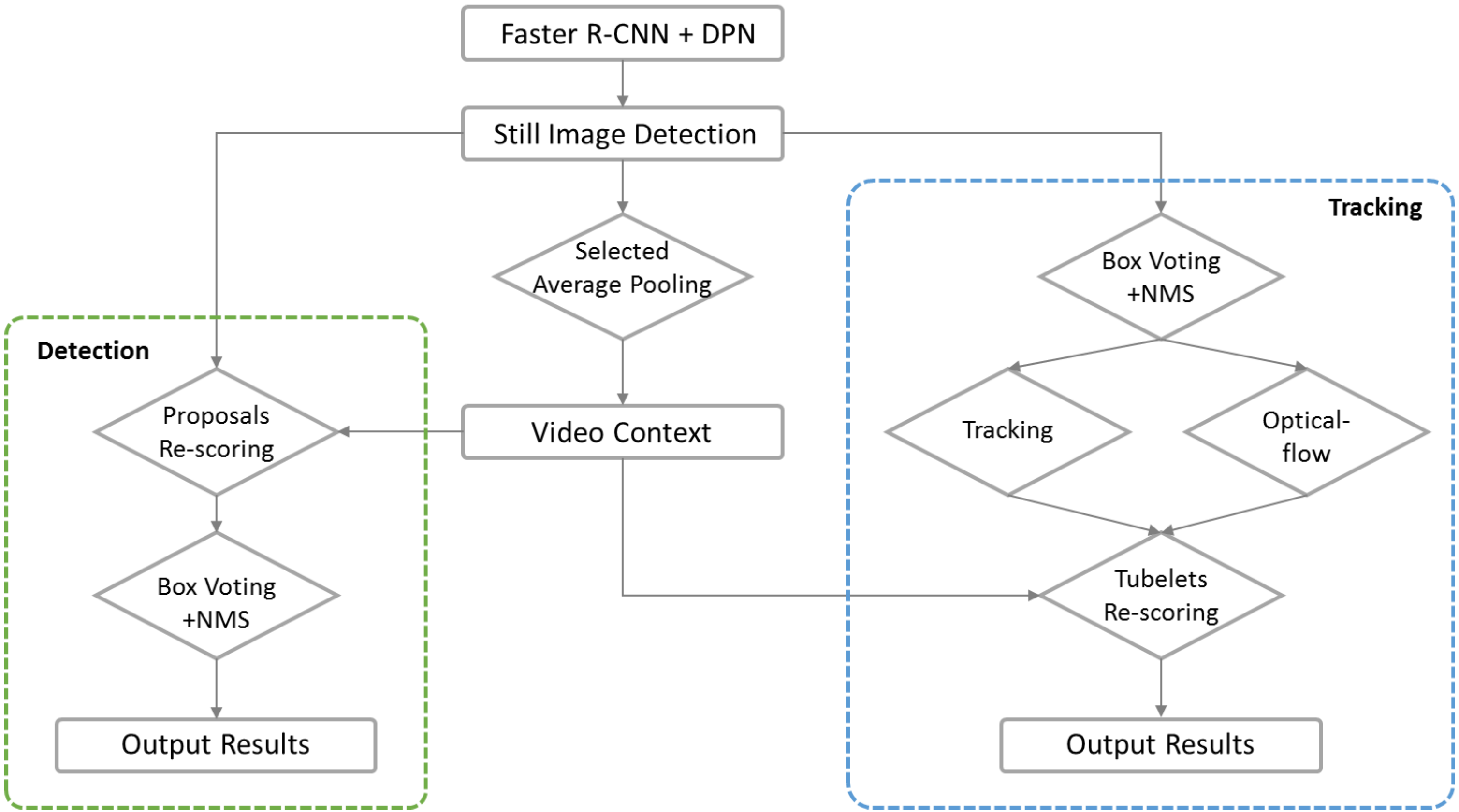
Speaker: Yunchao Wei

Thank Min Lin, Qiang Chen from Qihoo 360 for the extensive discussions.
Thank Xiaoli Liu, Ying Liu from Qihoo 360 for helping collect and annotate "external" data.

Results Overview

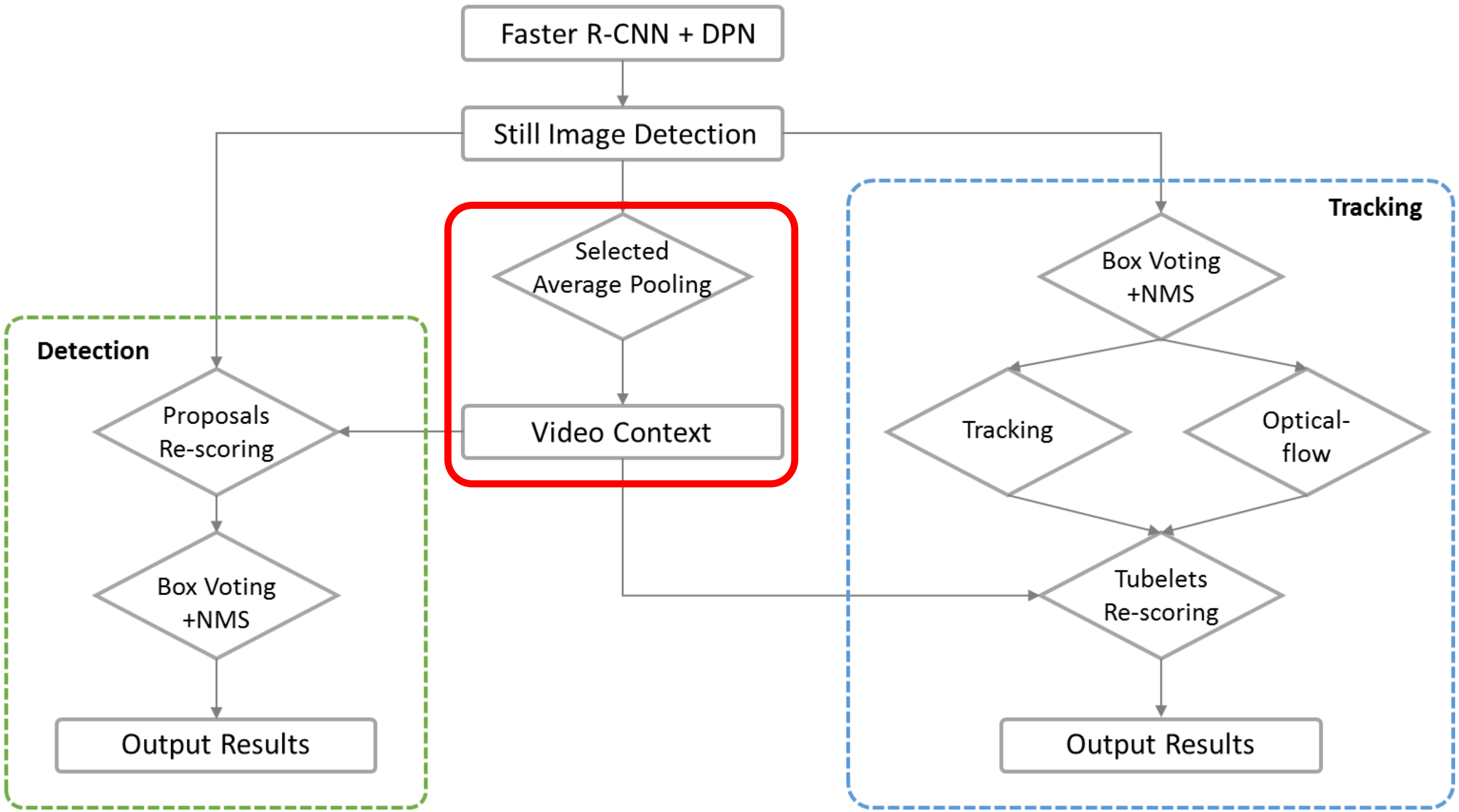
- **Objection Detection from Video**
  - a) **with "provided" data:** **2nd place** (by mAP: 75.8%)
  - b) **with "external" data:** **2nd place** (by mAP: 76.0%)
- **Object Detection/Tracking from Video**
  - a) **with "provided" data:** **2nd place** (by mAP: 54.5%)
  - b) **with "external" data:** **2nd place** (by mAP: 55.0%)

Framework



Faster R-CNN + DPN
Still Image Detection
Selected Average Pooling
Video Context
Detection
Proposals Re-scoring
Box Voting +NMS
Output Results
Tracking
Box Voting +NMS
Tracking
Optical-flow
Tubelets Re-scoring
Output Results

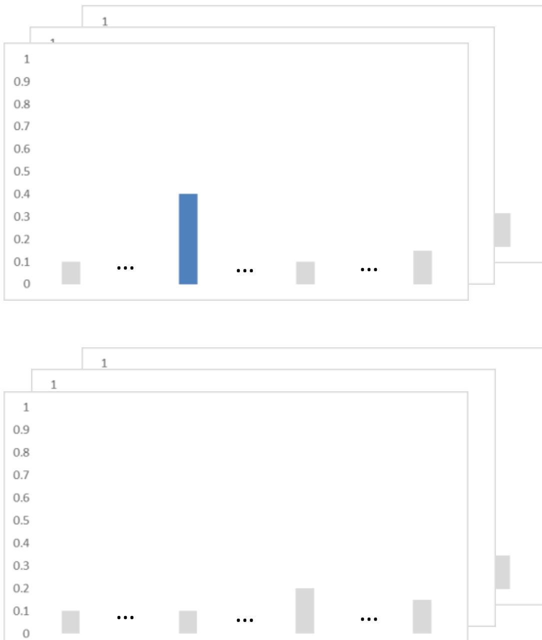
Framework

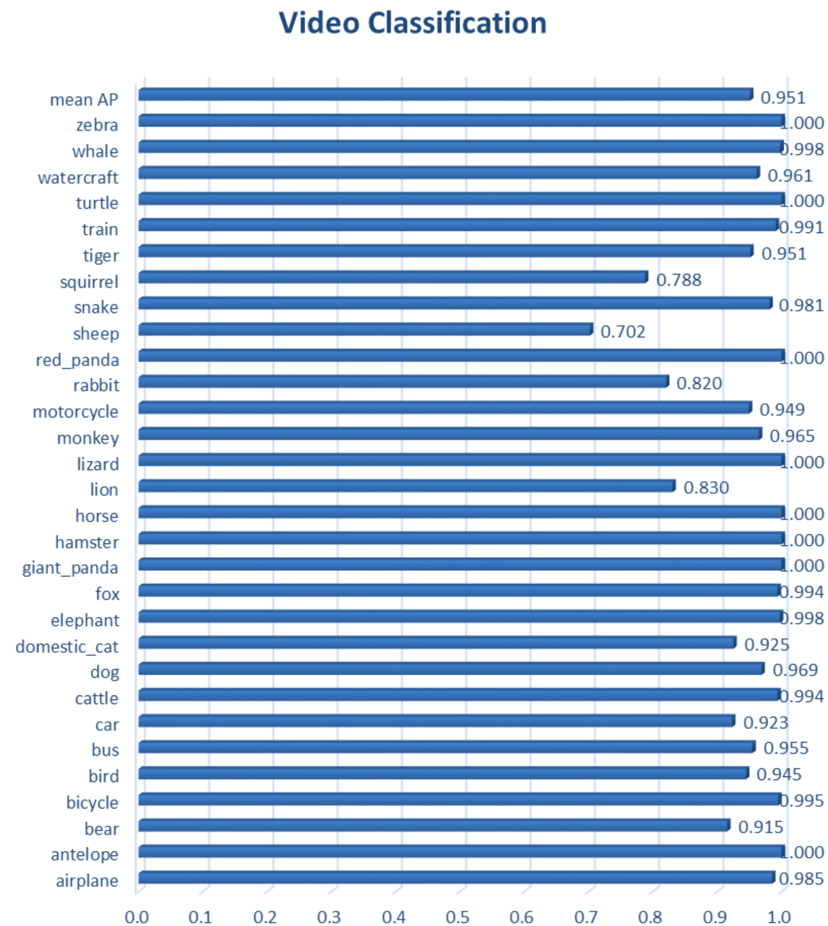


Faster R-CNN + DPN
Still Image Detection
Selected
Average Pooling
Video Context
Detection
Proposals
Re-scoring
Box Voting
+NMS
Output Results
Tracking
Box Voting
+NMS
Tracking
Optical-
flow
Tubelets
Re-scoring
Output Results

Video Context Modeling

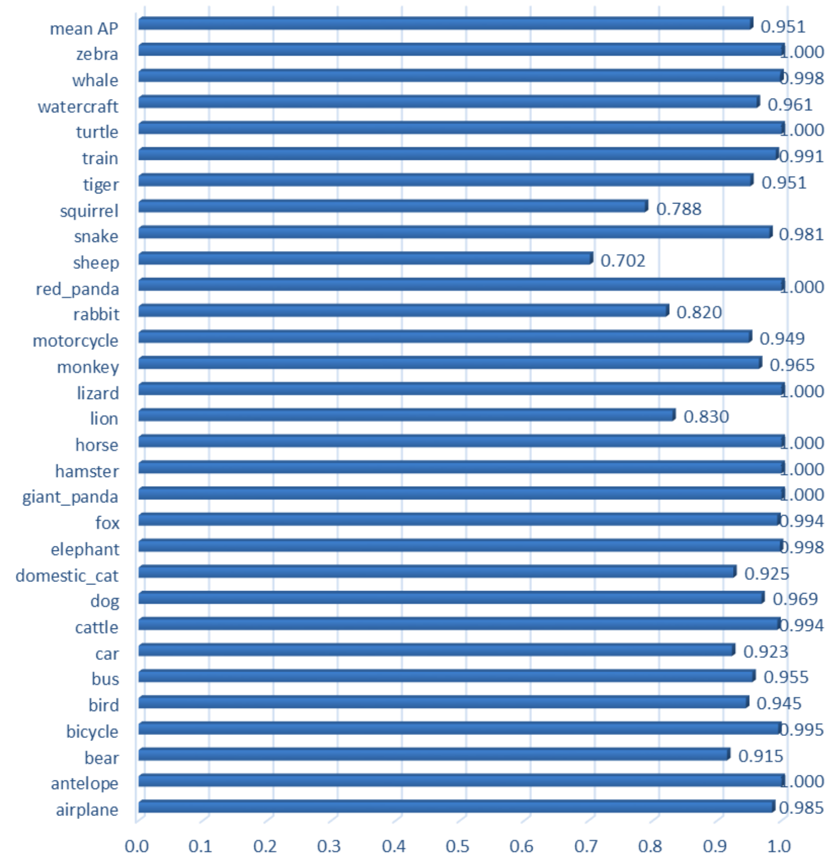
A selected-average-pooling method is proposed for modeling video-level context.





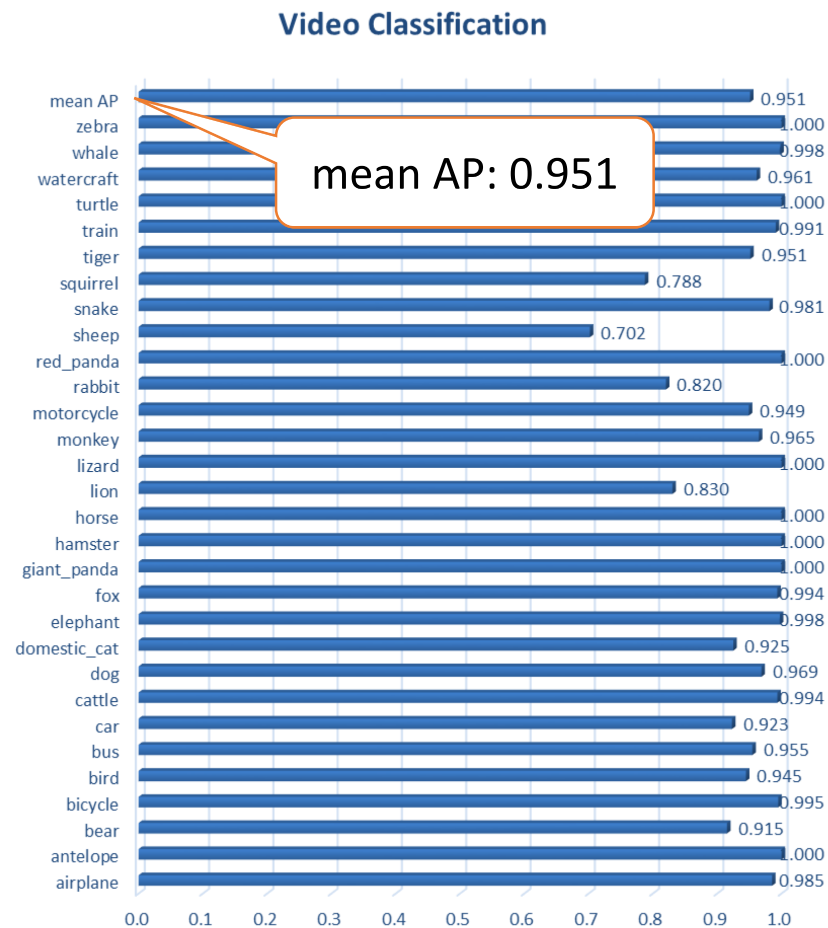
Video Context Modeling

A selected-average-pooling method is proposed for modeling video-level context.



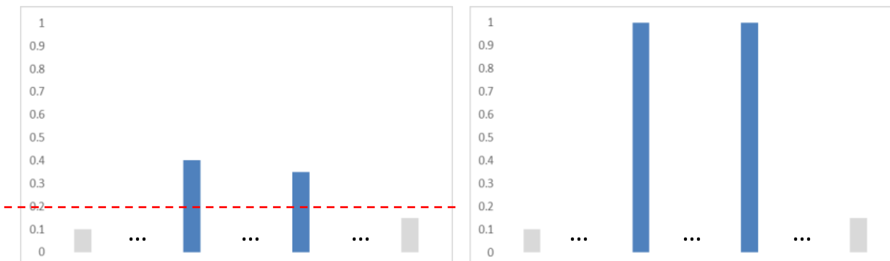
Video Context Modeling

A selected-average-pooling method is proposed for modeling video-level context.

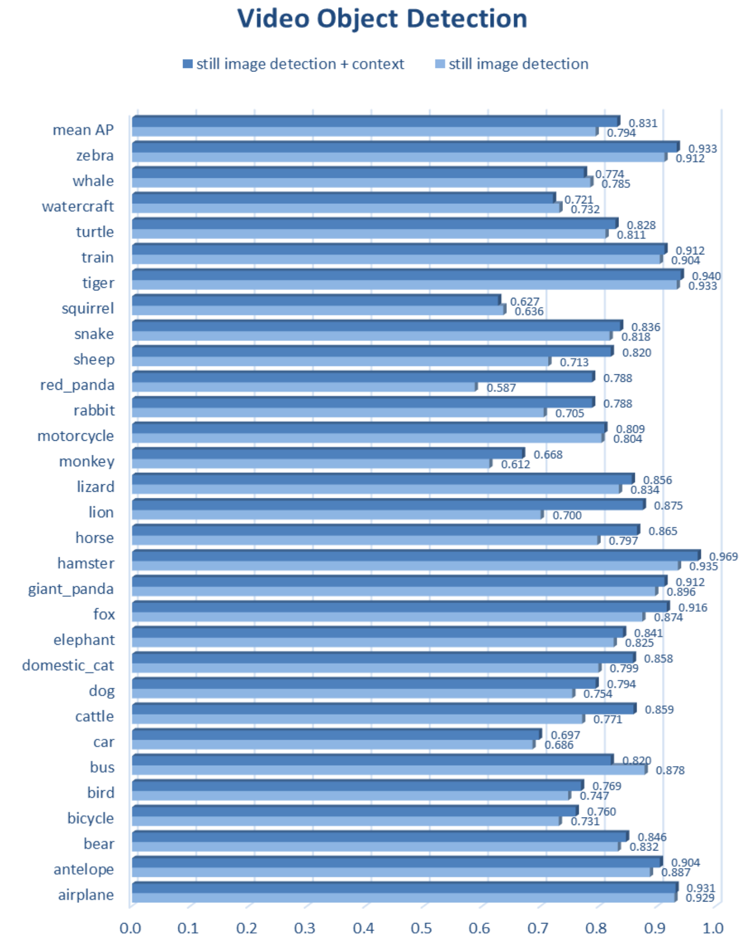


Video Object Detection

A ***larger-keep(LK)*** strategy is proposed to re-score proposal confidence scores using video context.



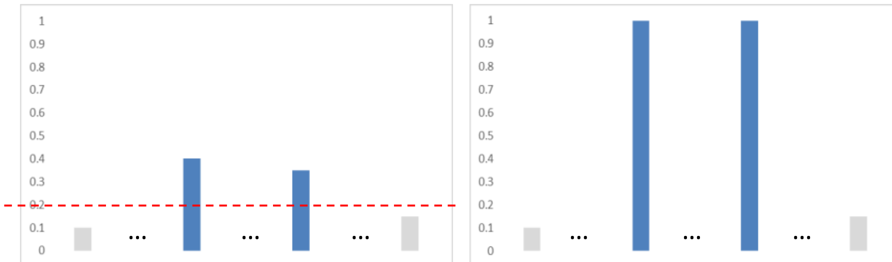
Method	mAP
Still Image Det	79.4
+Context(MCS[1])	80.6
+Context(ours w/o **LK**)	80.8
+Context(ours w/ **LK**)	83.1



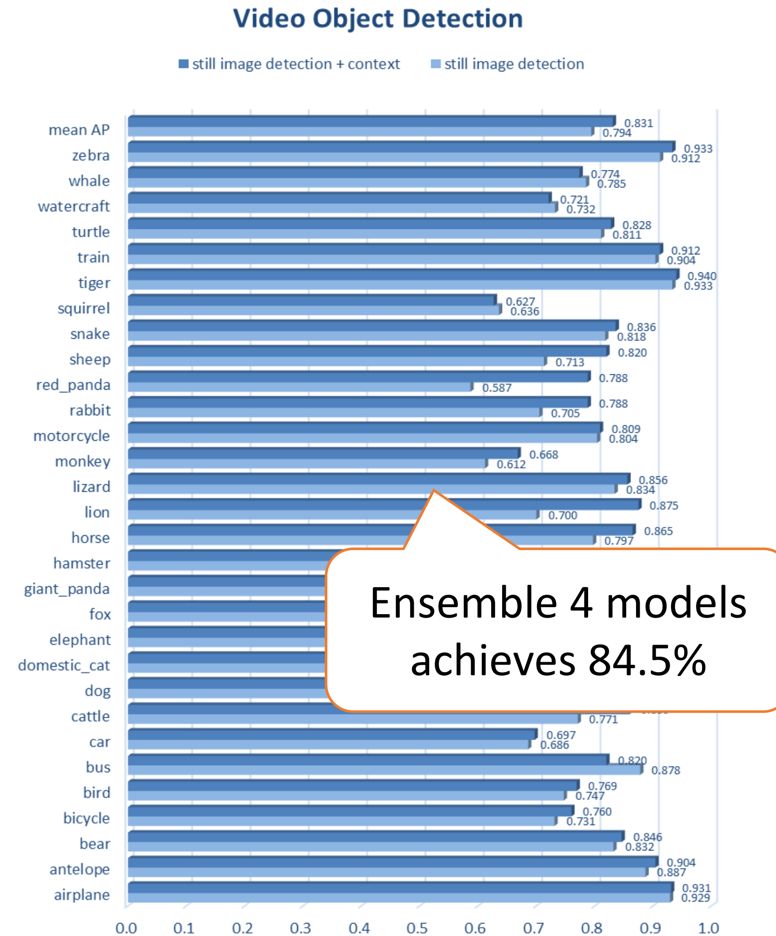
[1] K Kang et al. T-CNN: Tubelets with Convolutional Neural Networks for Object Detection from Videos. arXiv preprint 2016

Video Object Detection

A ***larger-keep(LK)*** strategy is proposed to re-score proposal confidence scores using video context.

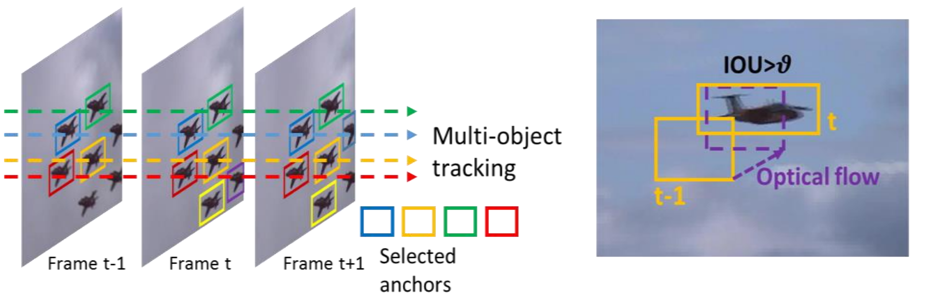


Method	mAP
Still Image Det	79.4
+Context(MCS[1])	80.6
+Context(ours w/o **LK**)	80.8
+Context(ours w/ **LK**)	83.1

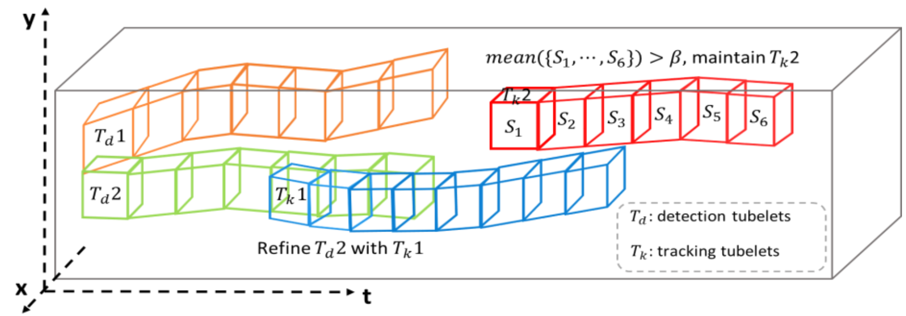


[1] K Kang et al. T-CNN: Tubelets with Convolutional Neural Networks for Object Detection from Videos. arXiv preprint 2016

Video Object Tracking



Tubelet Generation



Tubelet Fusion

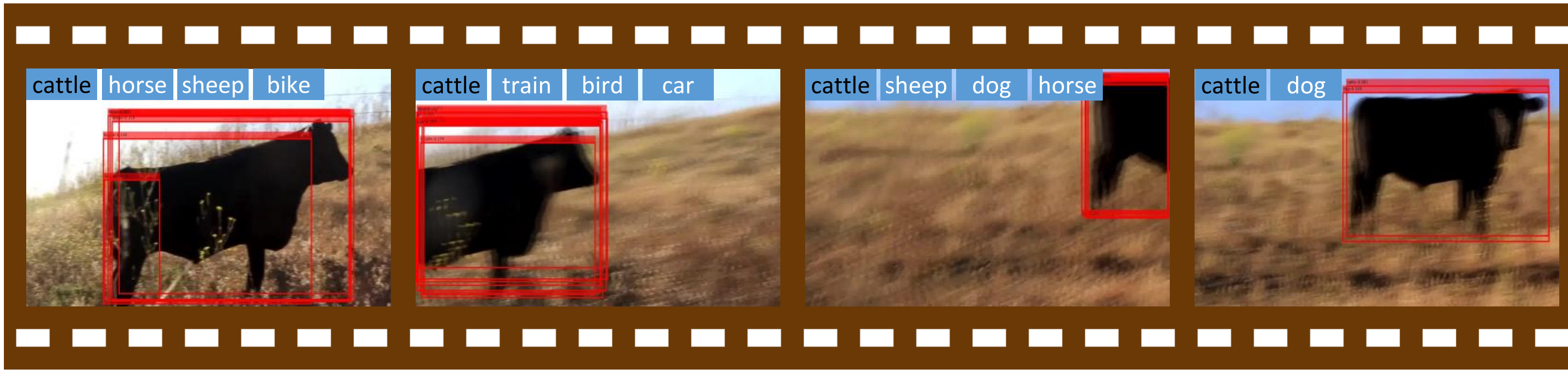
Results	Track_Det	Track_Det+MCS [1]	Track_Det+Context (Ours)
mAP@0.25	0.594	0.766	0.800
mAP@0.50	0.541	0.695	0.714
mAP@0.75	0.454	0.578	0.594
mAP	0.530	0.680	0.703

Comparison of Tracking Results

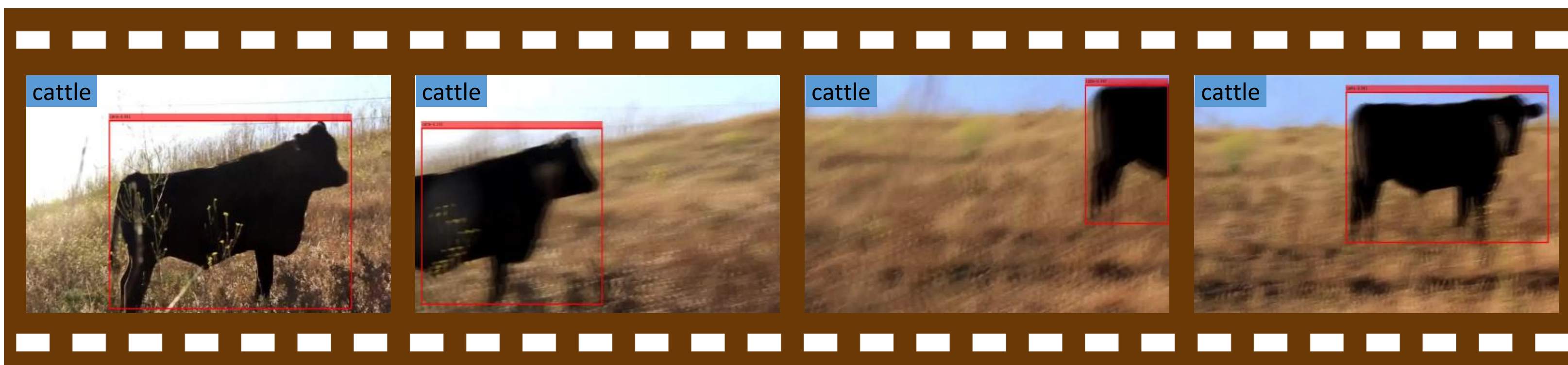
[1] K Kang et al. T-CNN: Tubelets with Convolutional Neural Networks for Object Detection from Videos. arXiv preprint 2016

Visualization

Still Image Detection



Still Image Detection + Video Context



Thank You!

National University of Singapore:

Yunchao Wei, Mengdan Zhang, Jianan Li, Yunpeng Chen, Jiashi Feng

elefjia@u.nus.edu

University of Illinois at Urbana-Champaign

Honghui Shi

shihonghui3@gmail.com

Qihoo 360 AI Institute:

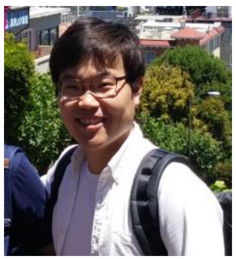
Jian Dong, Shuicheng Yan

yanshuicheng@360.cn

Thank Min Lin, Qiang Chen from Qihoo 360 for the extensive discussions.
Thank Xiaoli Liu, Ying Liu from Qihoo 360 for helping collect and annotate "external" data.

IEEE 2017 Conference on Computer Vision and Pattern Recognition

ObjectNet: Rank of Experts

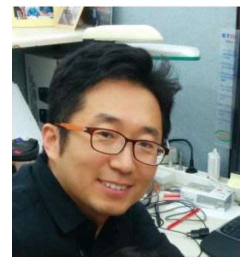


S. H. Bae

Y. J. Jo

J. W. Hwang

Y. W. Lee



Y. S. Yoon

Y. S. Bae

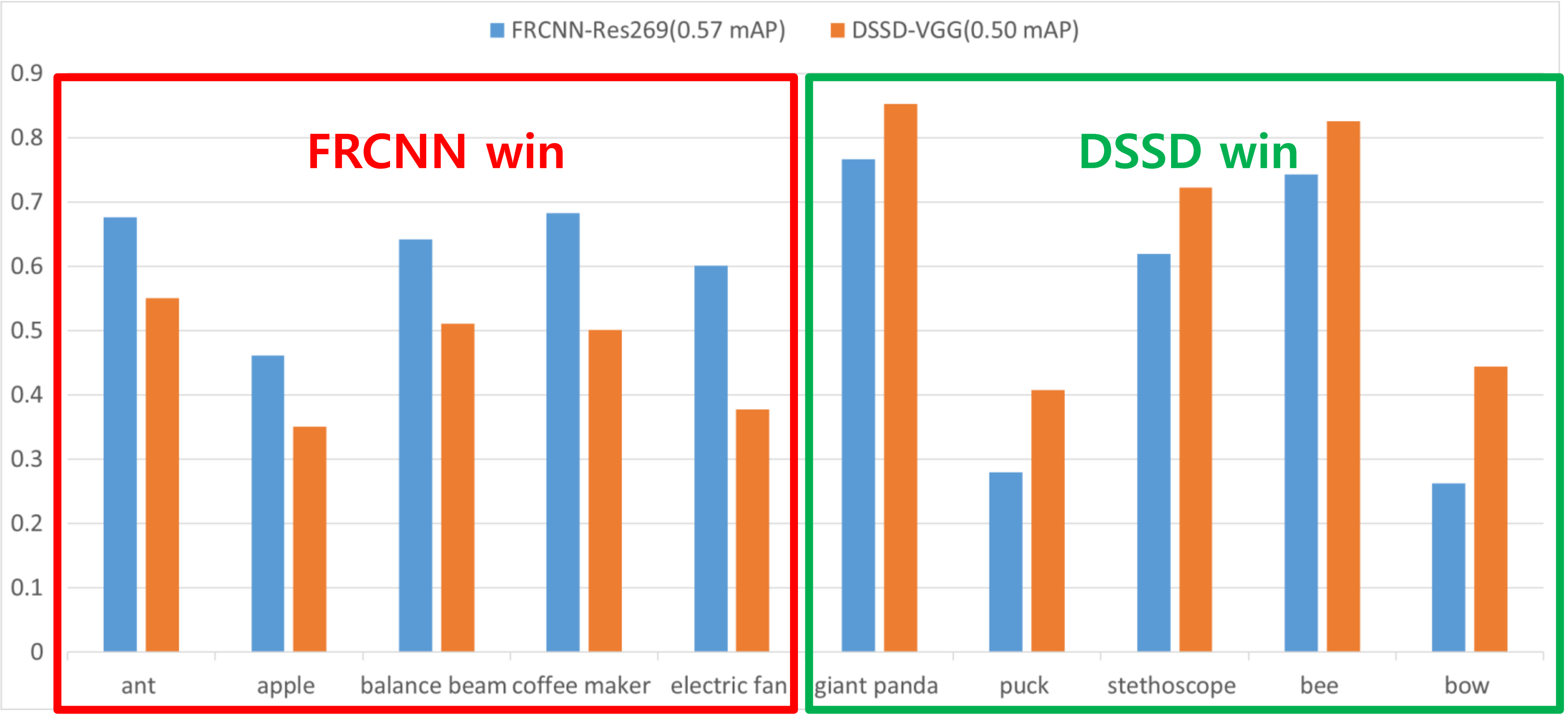
J. Y. Park

ILSVRC2017 DET results

Team	Categories won	Mean AP
BDAT	85	73.13%
DeepView (ETRI)	**10**	**59.30%**
NUS_Qihoo_DPNs	9	65.69%
KAISTNIA_ETRI	1	61.02%

Motivation

- **Difficult to train a dominant model for all classes**
  - ➢ Each model has different performance for classes
- **mAP is an indirect metric to select models for ensemble**
  - ➢ High mAP does not ensure superiority on class-wise performance



Our Approach: Detector Pool

- **Pursue Meta-Architecture Diversity**
  - Utilizing multiple feature extractor & meta-architecture pairs

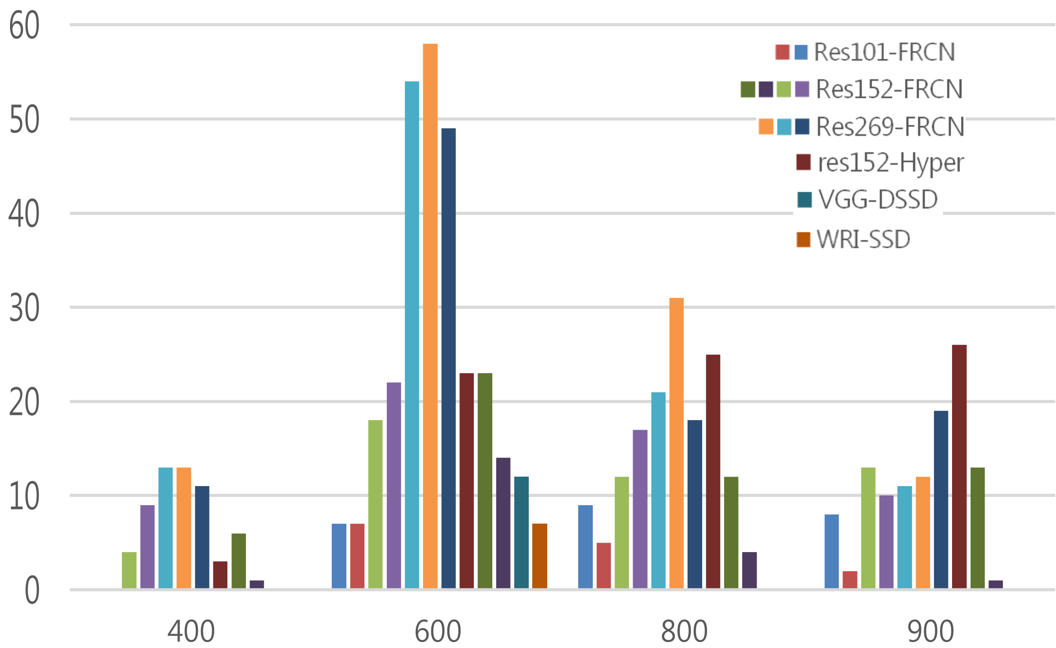
Feature Extractor	Meta-Architecture
Residual Network (101,152,269)	Faster RCNN
WR-Inception	SSD
VGG	DSSD

- **Enhance Small Object Detection**
  - Utilizing hyper feature maps
  - Multi-scale test: 400, 600, 800, 900
  - Mini-batch sampling: considering all ROI proposals (area > 0)

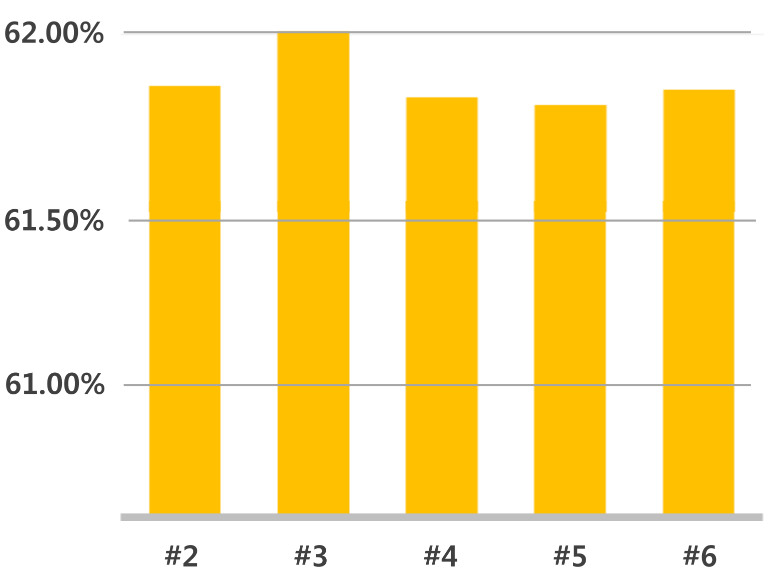
- **Solve Data Imbalance Problem**
  - Data balance: setting the positive & negative sample ratio to be equal
  - Data augmentation: generating augmented images for minority classes

Our Approach: Network Ensemble

- **Rank of Experts : Ranking & Selection**
  - ➢ Ranking models by class-wise performance → Combining results class-wise
  - ➢ Improving mean AP about 4~5% on val2 evaluation
  - ➢ Improving mean AP about 1% on the test set, but increasing number of object categories won



The number of selected models using our raking method



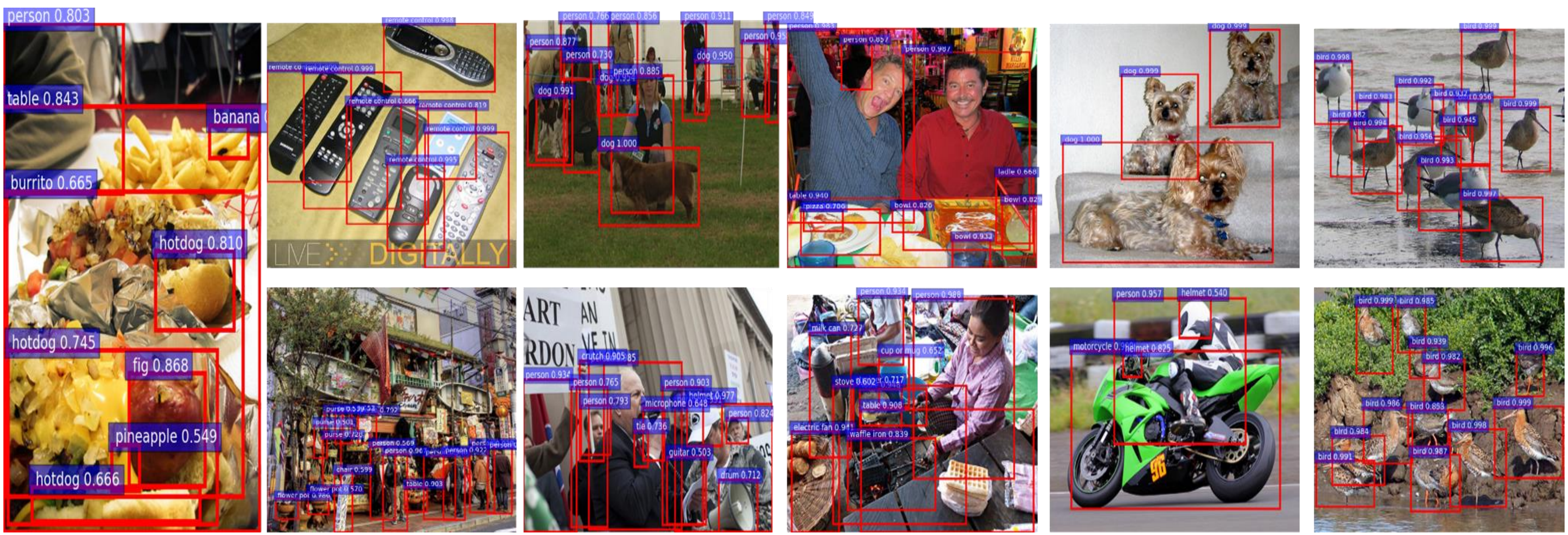
mean AP using our ensemble on val2

Experimental Results

ResNet-FRCN with different image resolutions

mean AP improvement

Methods	mean AP
Rank of experts (Ensemble)	**4~5% ↑**
Data augmentation	1~2% ↑
Multi Scale Test	~1% ↑
Soft-NMS	~1% ↑



Qualitative evaluation results using our ensemble model

MIL_UT at ILSVRC 2017

(5th Place in CLS Task)

Yuji Tokozume[1], Kosuke Arase[1], Yoshitaka Ushiku[1], Tatsuya Harada[1, 2]

[1]The University of Tokyo, [2]RIKEN

CVPR 2017 July 21-26 HONOLULU

東京大学 THE UNIVERSITY OF TOKYO

MIL Machine Intelligence Laboratory Beyond Human Intelligence

RIKEN

Core Idea

- We trained some existing networks with a novel learning method.

  (Temp. name: **TZ learning**)

Core Idea

- We trained some existing networks with a novel learning method.

(Temp. name: **TZ learning**)

TZ learning (ours):

Coming soon!

- A simple and powerful learning method for sound recognition. (Under review)

Core Idea

- We trained some existing networks with a novel learning method.

(Temp. name: **TZ learning**)

TZ learning (ours):

Coming soon!

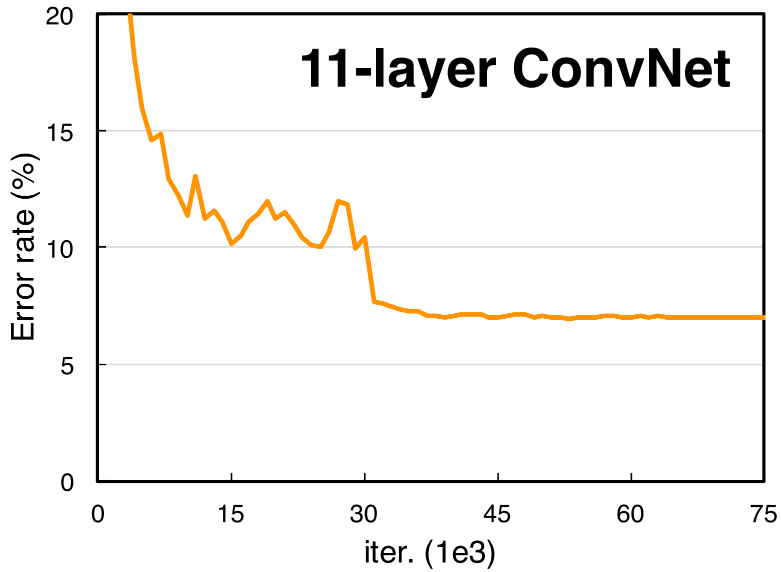
- A simple and powerful learning method for sound recognition. (Under review)
- It can boost the performance of various models without changing other settings.

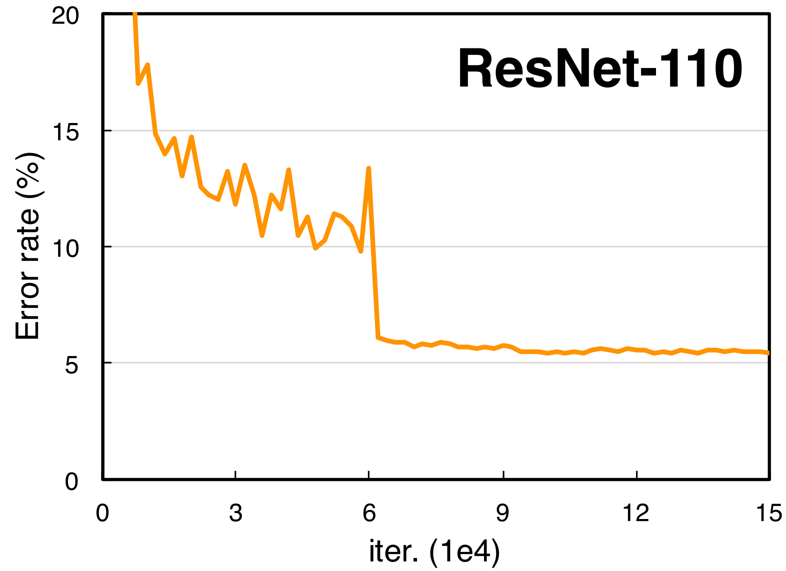
⋮

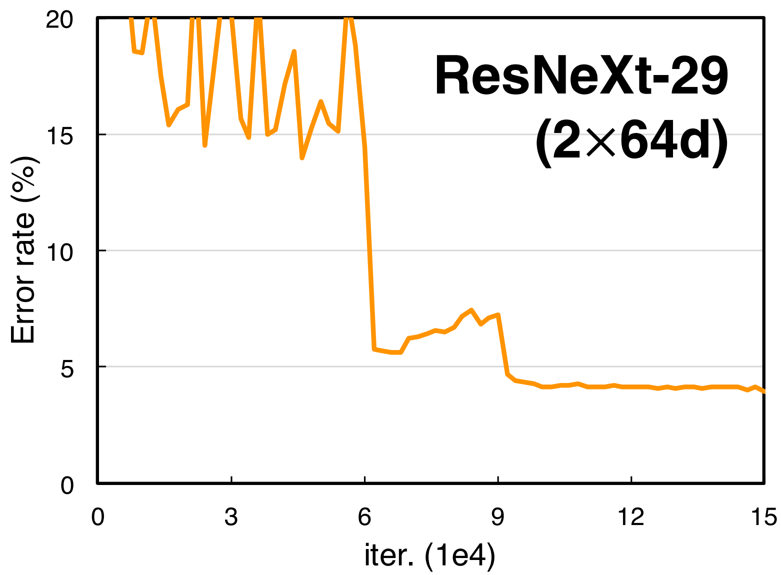
Preprocessing, Data augmentation, optimizer, etc.

CIFAR-10

Standard learning





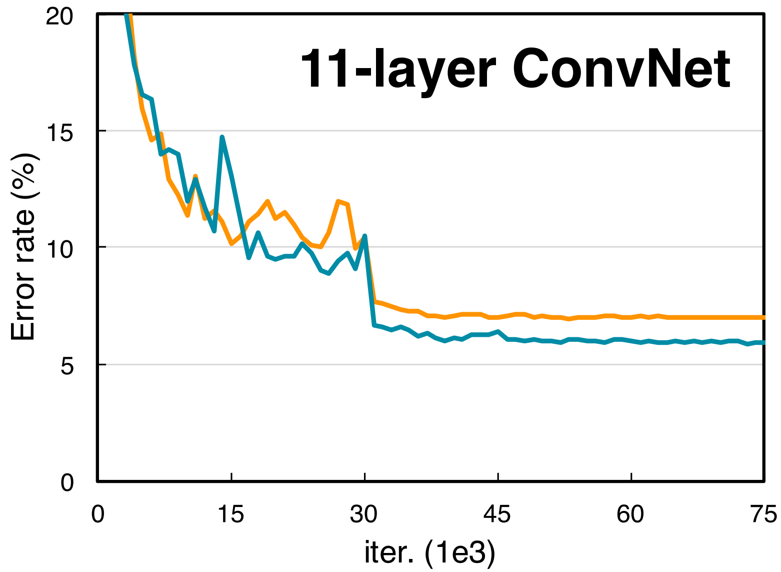


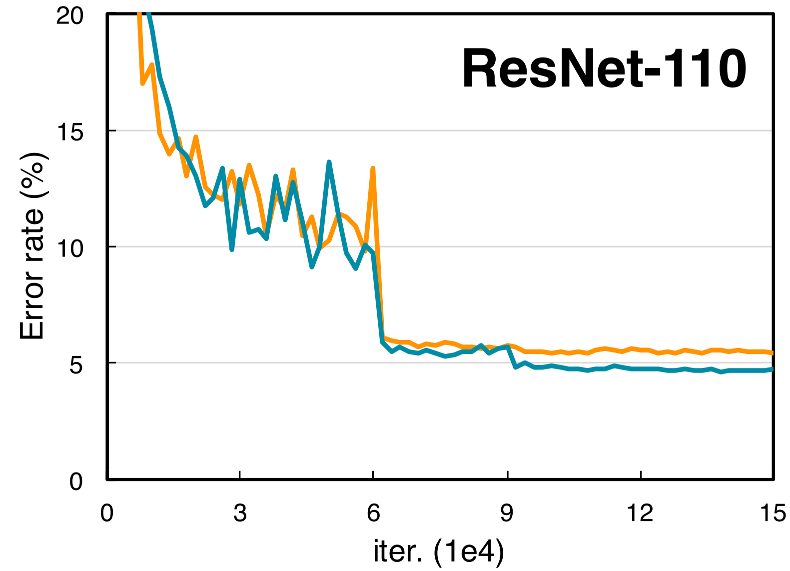
Model	Standard	TZ (ours)
11-layer ConvNet	7.20	
ResNet-110	5.69	
ResNeXt-29 (2×64d)	4.31	

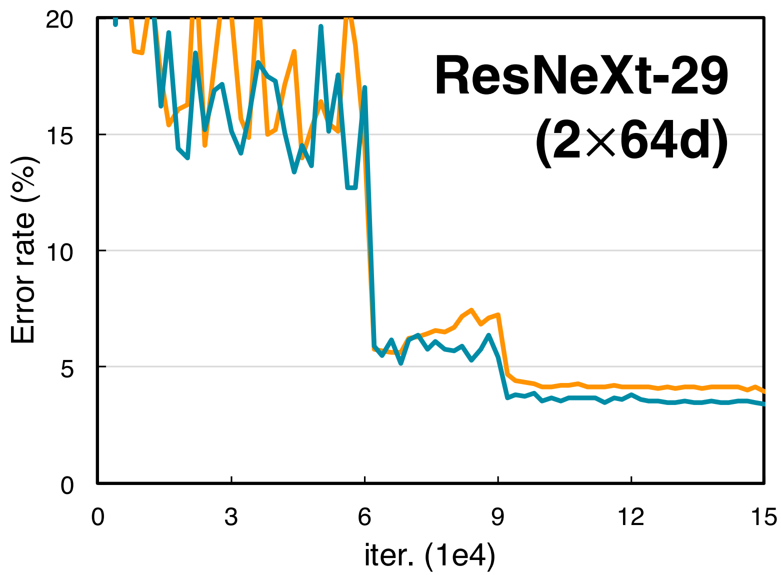
Error rate % (avg. of 5 trials)

CIFAR-10

Standard learning
TZ learning (ours)



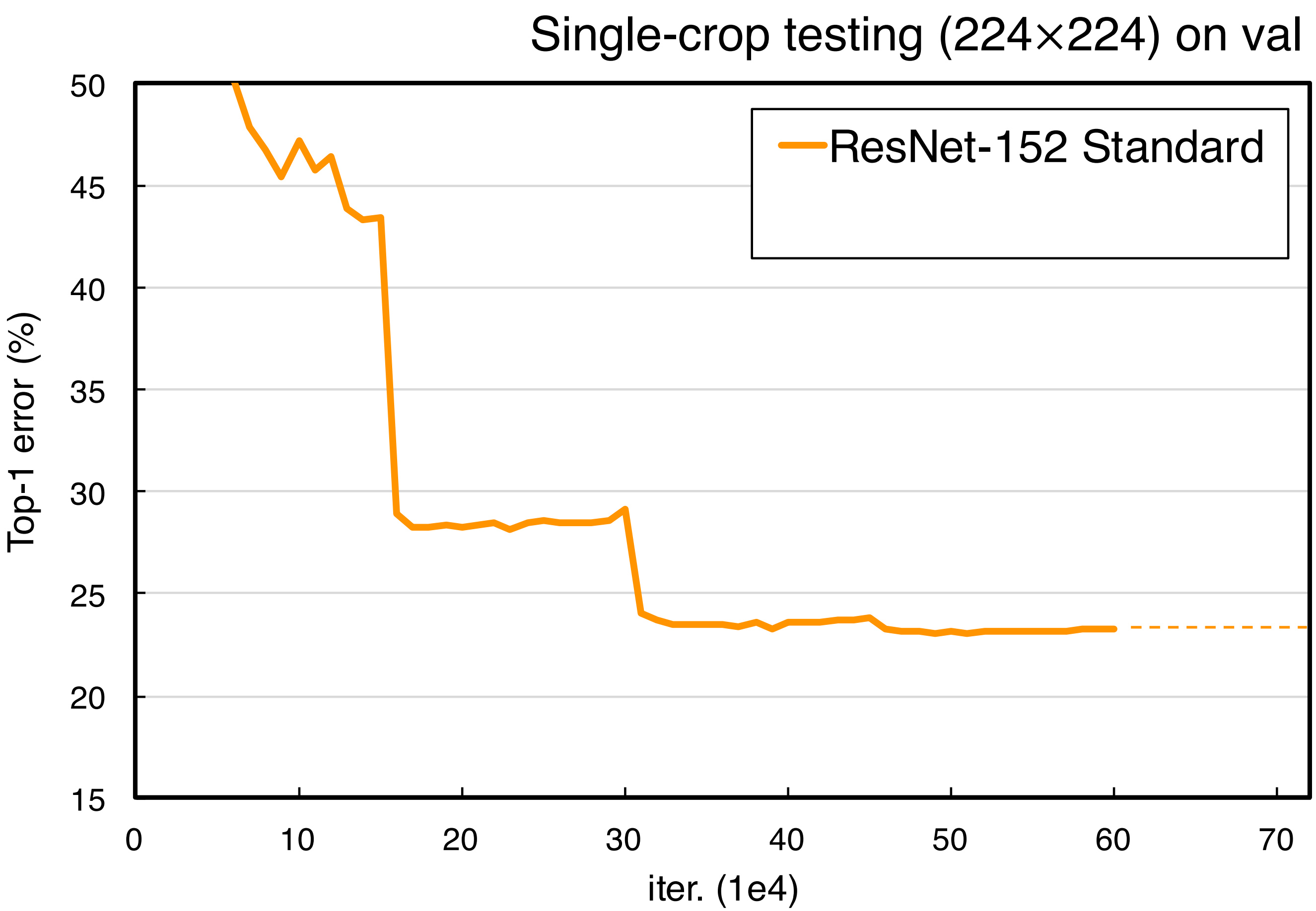




Model	Standard	TZ (ours)
11-layer ConvNet	7.20	**6.06** **(1.14% gain)**
ResNet-110	5.69	**5.24** **(0.45% gain)**
ResNeXt-29 (2×64d)	4.31	**3.58** **(0.73% gain)**

Error rate % (avg. of 5 trials)

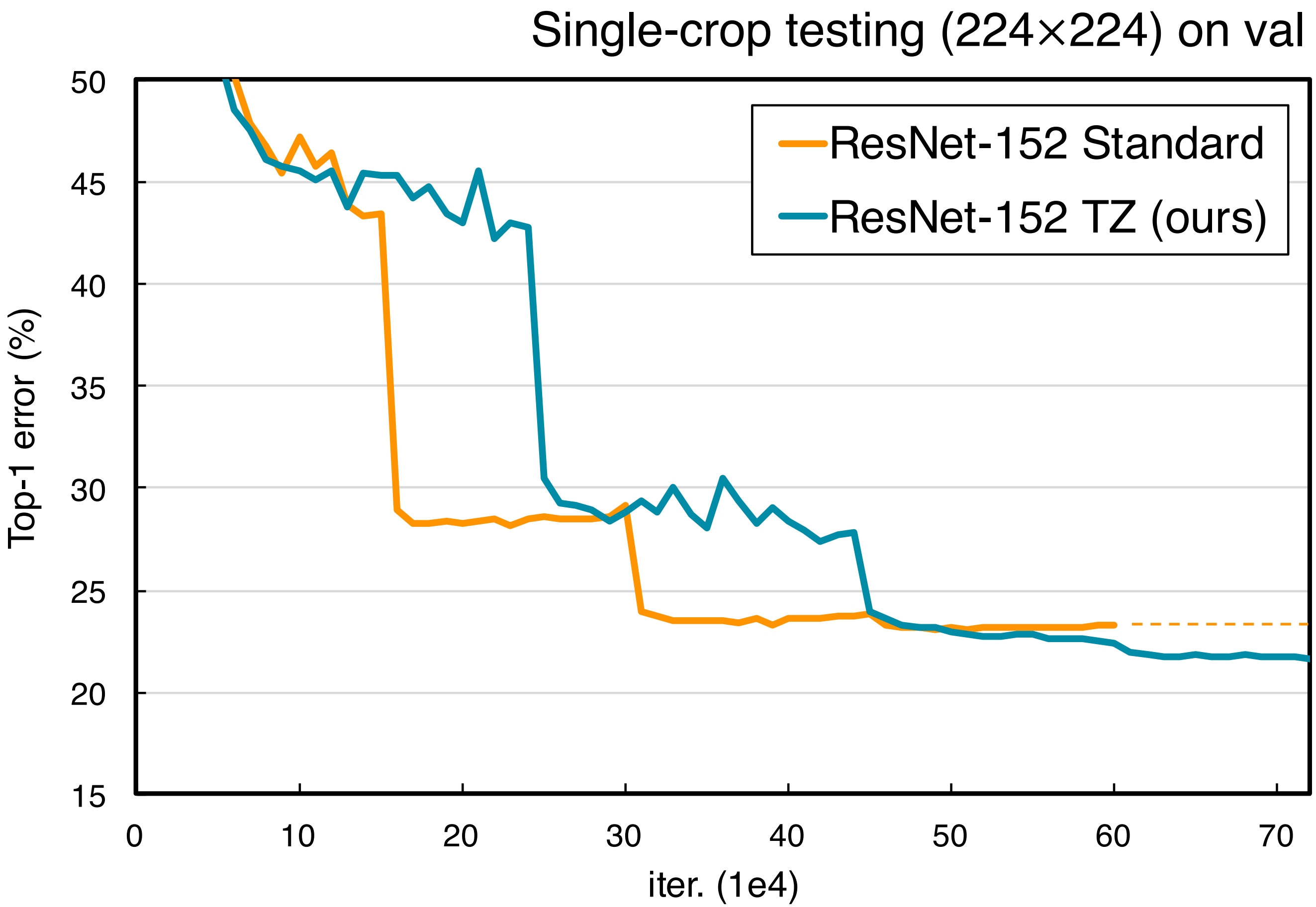
ImageNet



Single-model performance (%) (top-1/top-5 err.)

23.28 (19.40/**4.75**)

ImageNet



Single-model performance (%) (top-1/top-5 err.)

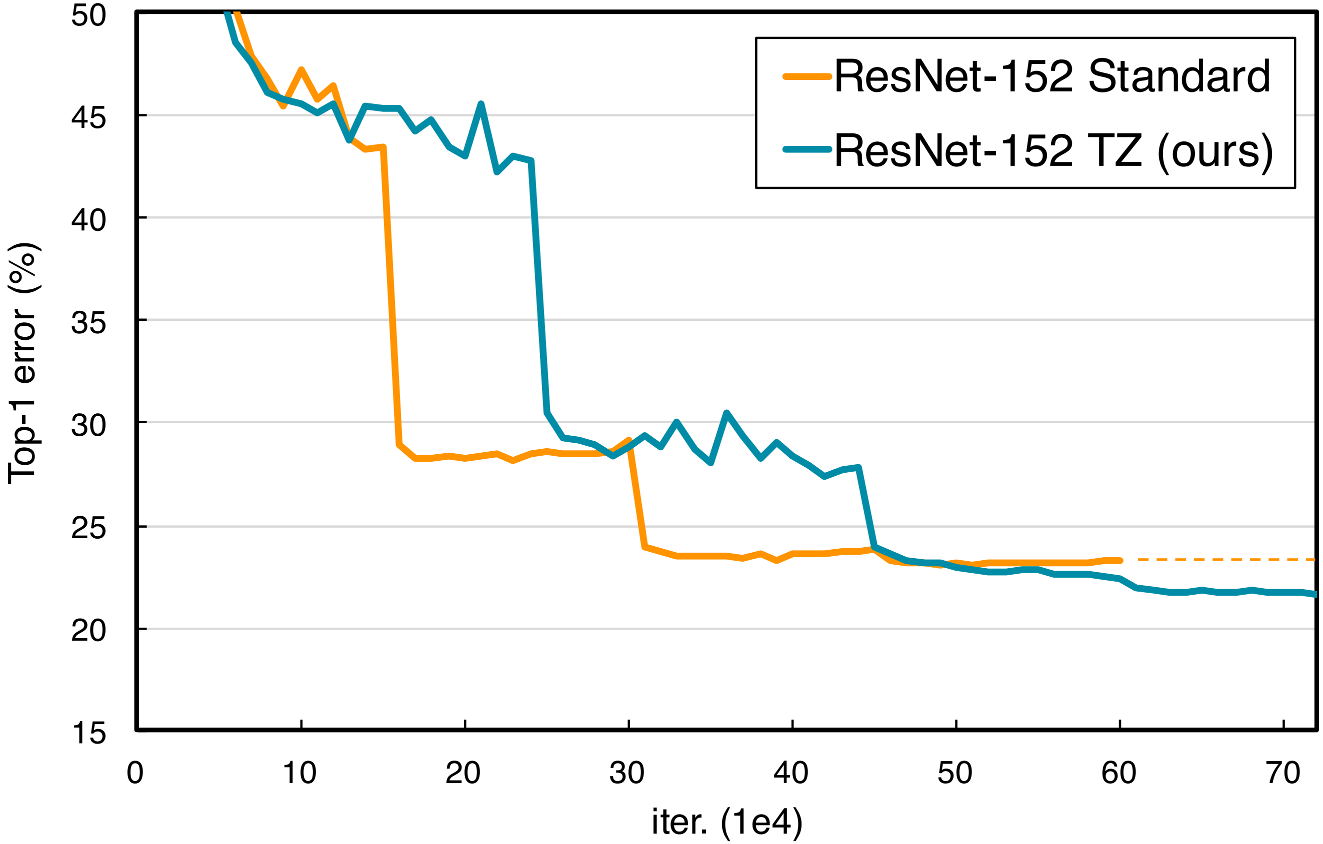
23.28 (19.40/**4.75**)

21.62 (18.69/**4.14**)

0.61% gain

ImageNet

Single-crop testing (224×224) on val



Single-model performance (%) (top-1/top-5 err.)

23.28 (19.40/**4.75**)

21.62 (18.69/**4.14**)

0.61% gain

- Final top-5 error on test: 3.205% (5th place)
- We are currently conducting further experiments.

IMAGENET Large Scale Visual Recognition Challenge 2017 (ILSVRC2017)

Deep Pyramidal Residual Networks

(for classification + localization task)

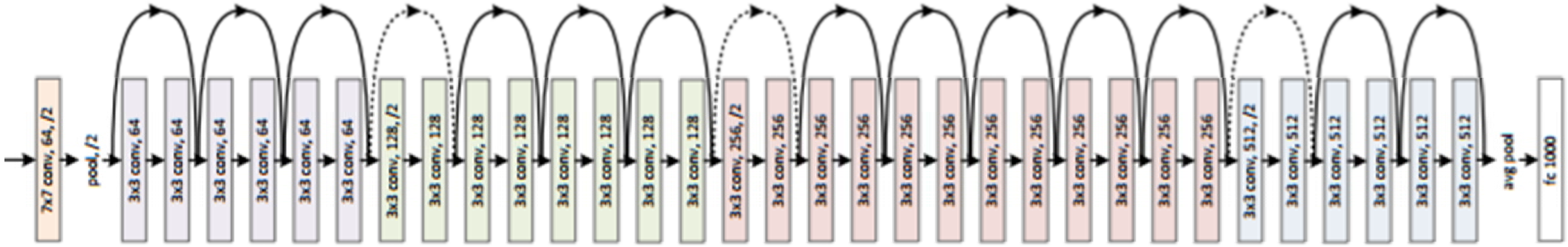
Dongyoon Han*, Jiwhan Kim*, Gwang-Gook Lee, and Junmo Kim
(equally contributed by the authors*)
{dyhan, jhkim89}@kaist.ac.kr, gwanggook.lee@sk.com, junmo.kim@kaist.ac.kr

Presenter: Dongyoon Han

TEAM: SIIT_KAIST+ SKT

Deep Pyramidal Residual Networks

- Deep residual networks (ResNet) [1] have shown a remarkable performance in image recognition.

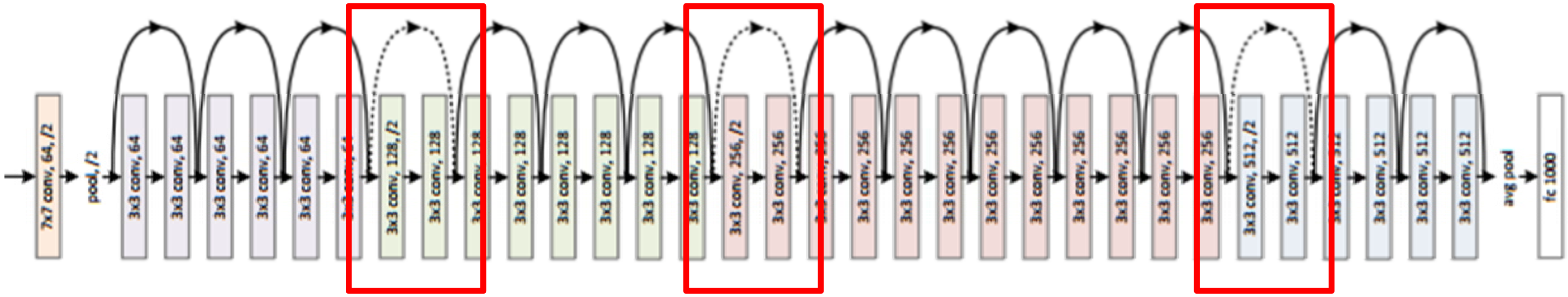


- According to [2], ResNet can be viewed like ensembles of relatively shallow networks.

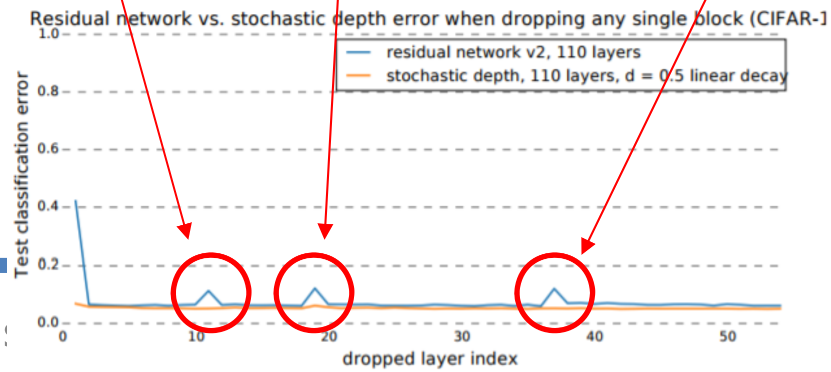
[1] K. He et al., "Deep Residual Learning for Image Recognition", CVPR 2016.
[2] A. Veit et al., "Residual Networks Behave Like Ensembles of Relatively Shallow Networks", NIPS 2016.

Deep Pyramidal Residual Networks

- Deep residual networks (ResNet) [1] have shown a remarkable performance in image recognition.

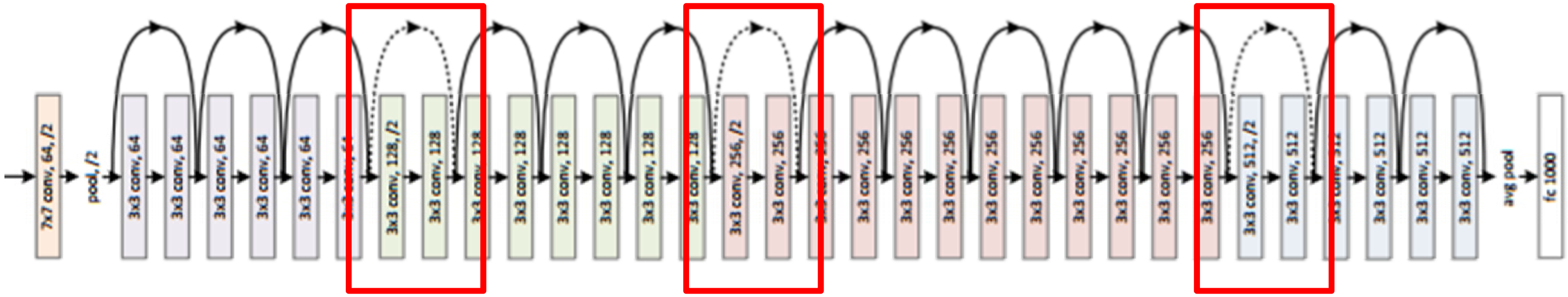


- According to [2], ResNet can be viewed like ensembles of relatively shallow networks.
  - Exp: deleting individual layers from networks at test time.
  - Deleting **a layer with increasing feature dimensions** leads to degrade performance, which is shown with a error fluctuation:



Deep Pyramidal Residual Networks

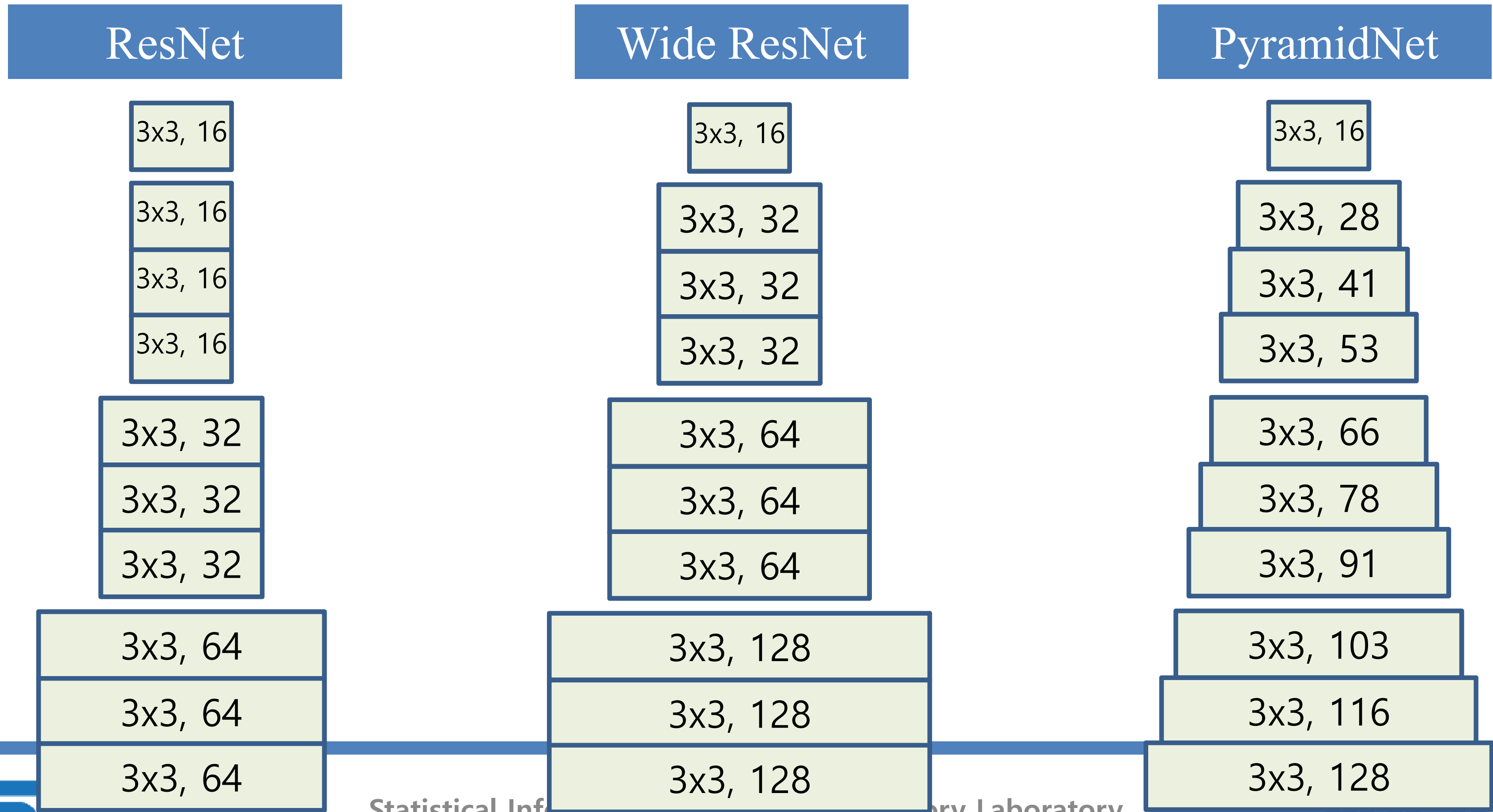
- Deep residual networks (ResNet) [1] have shown a remarkable performance in image recognition.



- According to [2], ResNet can be viewed like ensembles of relatively shallow networks.
  - Exp: deleting individual layers from networks at test time.
  - Deleting **a layer with increasing feature dimensions** leads to degrade performance shown with a error fluctuation.
- We conjectured that **increasing the feature dimension gradually**, instead of sharply increasing only at some blocks can
  - diminish the error fluctuation phenomenon and
  - increase ResNet's ensembling effect.

Deep Pyramidal Residual Networks

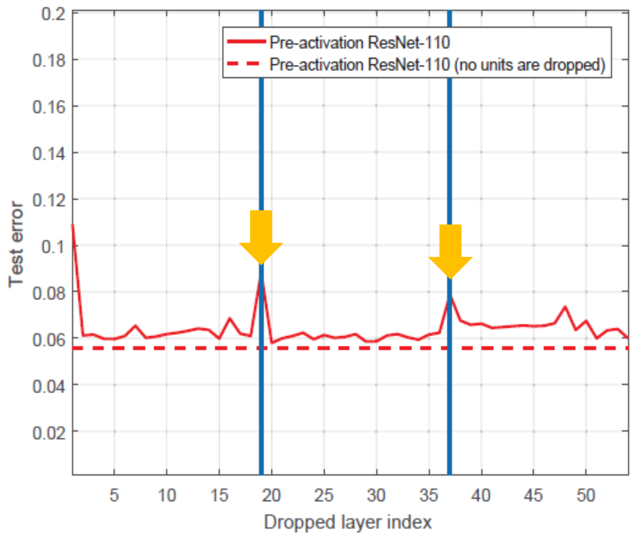
- Schematic illustrations of ResNet, Wide ResNet and PyramidNet.
- Each block denotes conv stacks (units) with feature map dimension.



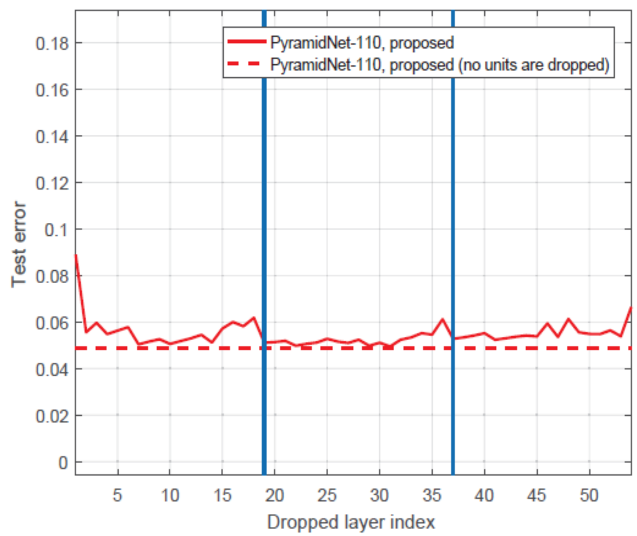
Deep Pyramidal Residual Networks

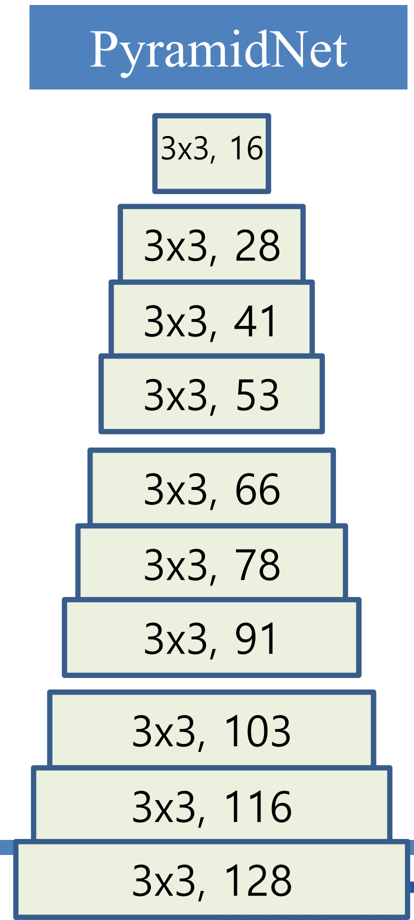
- Experimental results of dropping a single layer at test time:

(a) Pre-activation ResNet



(b) PyramidNet





Please come to our poster for more details!

Thank you!

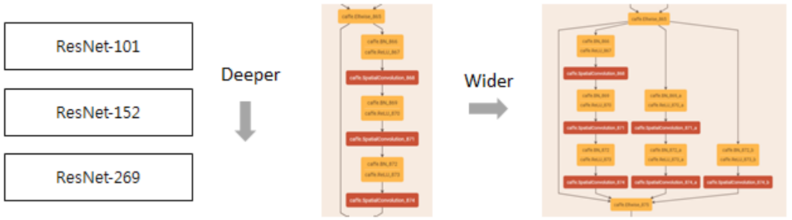
Aggregating multi-level/shape features and confidence penalty for object detection

Keun Dong Lee, Seungjae Lee, Jong Gook Ko

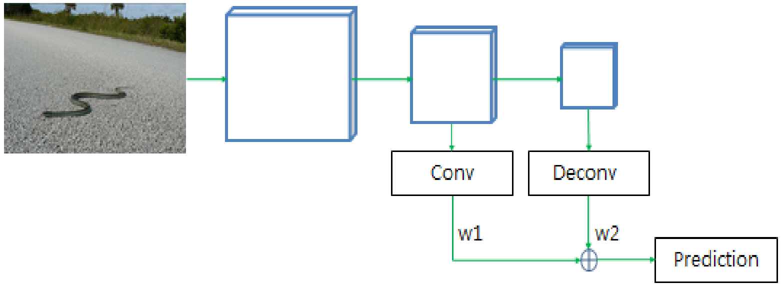
Jaehyung Kim, Jun Hyun Nam, Jinwoo Shin

Improving Detection Networks beyond GBD Network [Xingyu et al. 2017]

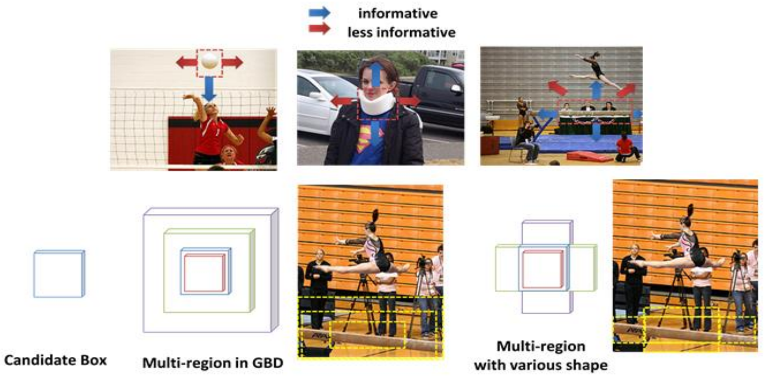
- ## Width and Depth
  - Train various depths (101/152/269) and widths for multi-region networks.
  - Some classes has better results in the shallower network (e.g. orange, burrito) and in the wider network (e.g. baby bed, violin and ladybug).



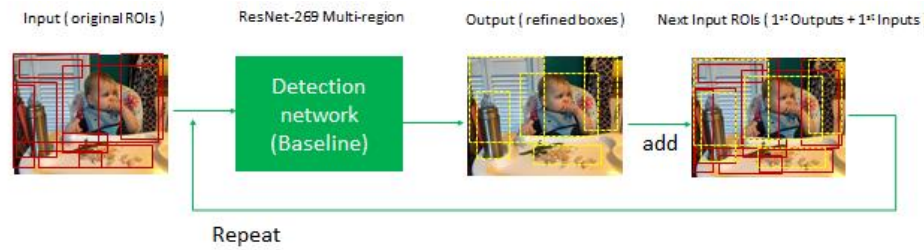
- ## Multi-level Features
  - Train model with weighted addition fusion of different layer feature maps
  - Upper level feature map has more weight value
  - It is effective for recognizing small size objects such as wine bottle, puck, band aid and remote control, etc



- ## Multi-shape Features
  - Train model with various shape of surrounding regions for context pooling
  - Informativeness of surrounding regions is varying according to the directions (noise or context)
  - AP gain in 90 classes such as balance beam, neck brace, volleyball



- **Iterative Region Proposals**
  - Cascaded RPN: Train a baseline model to generate ROIs and take an ensemble of two models trained independently.
  - Iterative box refinement: Use predicted boxes generated by a trained detection network as new ROIs together with previous input ROIs.



- **Confidence Penalty**
  - Detection network often fails because of high scored background or unlabeled objects.
  - To resolve this issue, we added negative entropy to the original loss function to regularize highly confident background output.

$$\mathcal{L}(\theta) = -\sum \log p_\theta(\mathbf{y}|\mathbf{x}) - \beta H(p_\theta(\mathbf{y}|\mathbf{x}))$$

- **Experimental Results**
  - Apply aggregating multi-level/shape features and confidence penalty
  - Commonly used techniques such as global context, box averaging and different ensemble rules

No	Model	mAP (val2)	mAP (test)
1	baseline/ baseline with aggregating RPs	0.622/0.626	-
2	1 + confidence penalty	0.635	-
3	2 + width and depth	0.642	0.60827
4	3 + multi-shape features	0.645	0.60829
5	4 + multi-level features (different ensembles)	0.650	0.61022