IMAGENET COCO Common Objects in Context

Towards Good Practices for Recognition & Detection

Qiaoyong Zhong, Chao Li, Yingying Zhang, Haiming Sun

Shicai Yang, Di Xie, Shiliang Pu

{zhongqiaoyong, sunhaiming, yangshicai, xiedi}@hikvision.com

Hikvision Research Institute

October 9, 2016

Team Members

HIKVISION

Scene Classification:

- *Shicai Yang*

Scene Parsing:

- *Haiming Sun*
- *Di Xie*

DET + LOC:

- *Qiaoyong Zhong*
- *Chao Li*
- *Yingying Zhang*
- *Di Xie*

Summary of Our Submissions

HIKVISION

- **Scene Classification**
  - 1st place, 0.0901 top5 error
- **Scene Parsing**
  - 7th place, 0.53335 average mIoU & pixel accuracy
- **Object Detection**
  - 2nd place, 0.653 mAP
- **Object Localization**
  - 2nd place, 0.0874 localization error

Scene Classification

HIKVISION

- **Data Augmentation**
  - Color Augmentation (directly adopted from [1])
  - PCA Jittering (from [2])
  - Random Image Interpolation
  - Crop Sampling
    - scale jittering (from [3][4])
    - scale and aspect ratio augmentation (from [5])
      - random area ratio (a = [0.08, 1])
      - random aspect ratio (s = [3/4, 4/3])
      - crop size: W' = sqrt(W*H*a*s); H' = sqrt(W*H*a/s)
      - random offset to pick crop center, then crop and resize
    - **Supervised Data Augmentation**

[1] https://github.com/facebook/fb.resnet.torch/
[2] A. Krizhevsky, et al. ImageNet Classification with Deep Convolutional Neural Networks. NIPS, 2012.
[3] K. Simonyan, et al. Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR, 2015.
[4] K. He, et al. Deep Residual Learning for Image Recognition. CVPR, 2016.
[5] C. Szegedy, et al. Going Deeper with Convolutions. CVPR, 2015.

Scene Classification

HIKVISION

- **Supervised Data Augmentation (SDA)**
  - train a model from scratch (coarse model)
  - use coarse model to generate ground truth class activation
  - randomly select a location based on prob. of target class
  - map this location to original image
  - randomly select a crop center near that location in original image
  - other steps are similar with the method in GoogLeNet paper

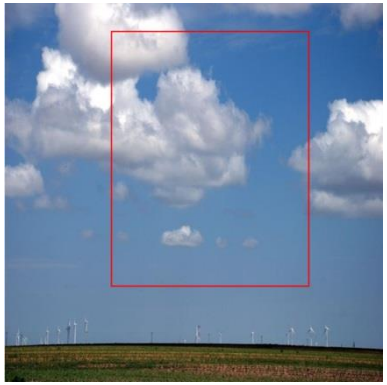
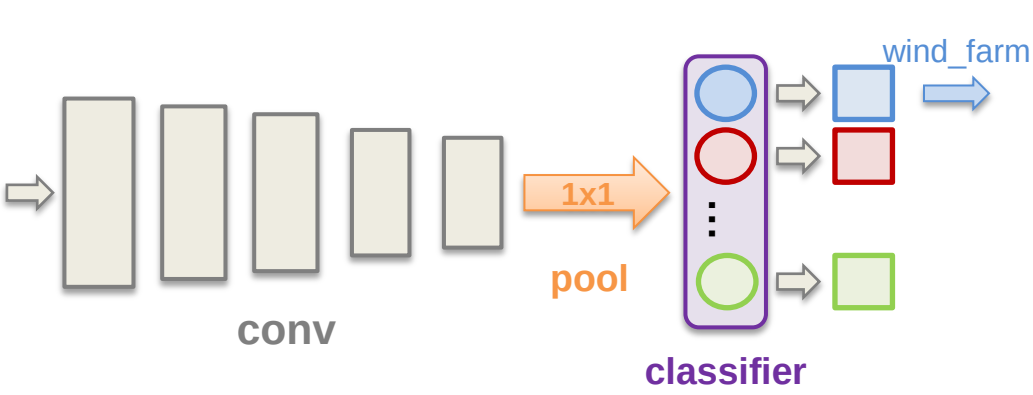
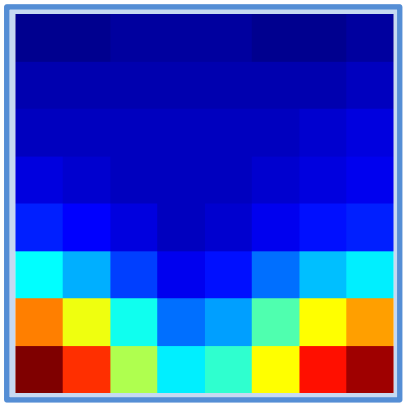


image
(ground truth: wind_farm)





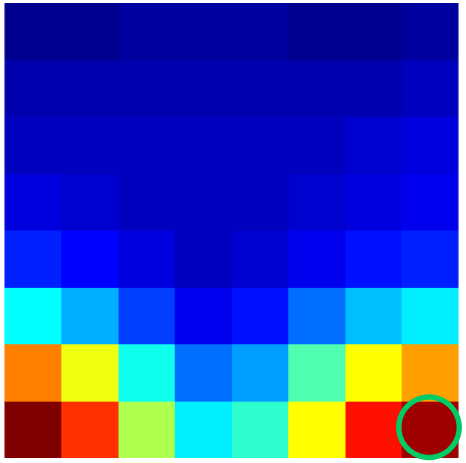
ground truth class activation

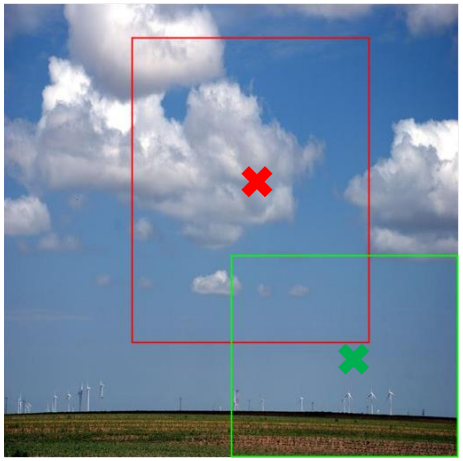
Inspired by: [6] B. Zhou, et al. Learning Deep Features for Discriminative Localization. CVPR, 2016.

Scene Classification

HIKVISION

- **Supervised Data Augmentation (SDA)**
  - train a model from scratch (coarse model)
  - use coarse model to generate ground truth class activation
  - randomly select a location based on prob. of target class
  - map this location to original image
  - randomly select a crop center near that location in original image
  - other steps are similar with the method in GoogLeNet paper





Randomly Selected Crop:

Original (in red)

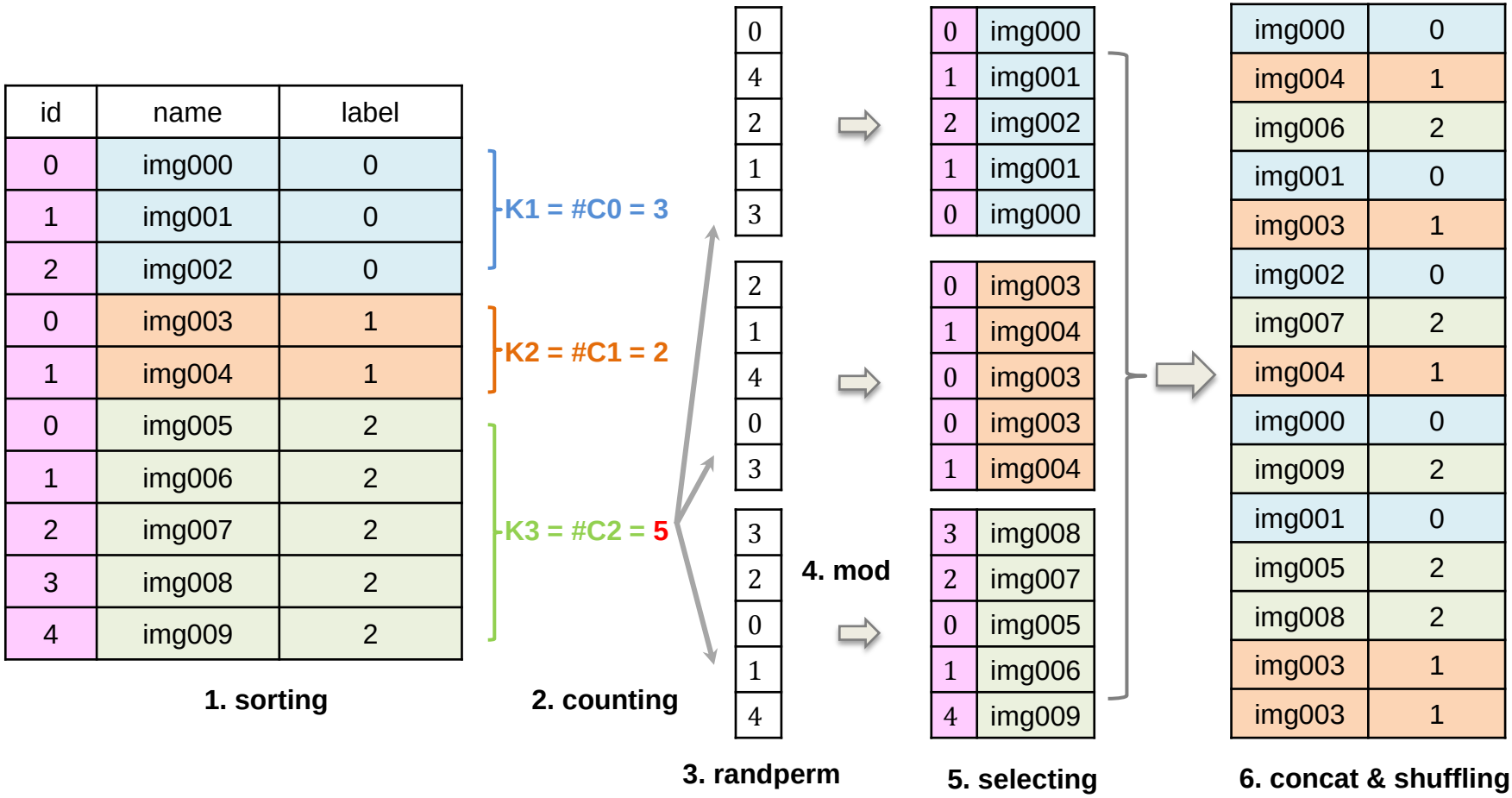
SDA (in green)

Inspired by: [6] B. Zhou, et al. Learning Deep Features for Discriminative Localization. CVPR, 2016.

Scene Classification

HIKVISION

- **Imbalanced Class Problem**
  - Balanced Sampling via *Label Shuffling*

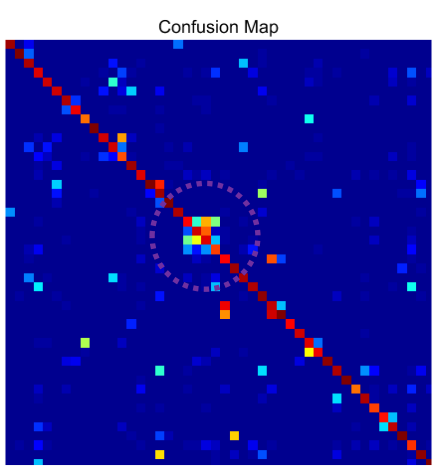


[7] L. Shen, et al. Relay Backpropagation for Effective Learning of Deep Convolutional Neural Networks. ECCV, 2016.

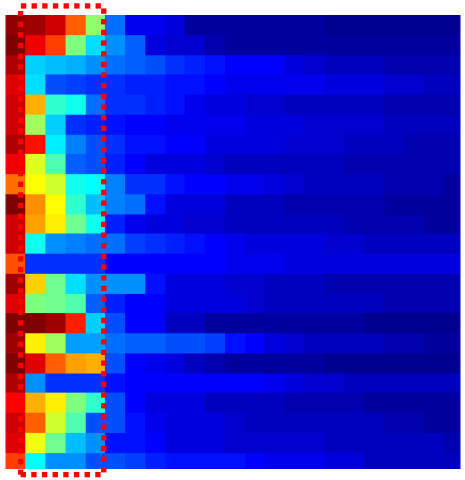
Scene Classification

HIKVISION

- **label smoothing** (LS) via prior label distribution



Co-label Similarity



Top 4 Similar Labels

0	0	0	0	0	0	1	0	0	0	0	...	0
ε/K	ε/K	ε/K	ε/K	ε/K	ε/K	$1-\varepsilon+\varepsilon/K$	ε/K	ε/K	ε/K	ε/K	...	ε/K
ε/K	ε/K	$\tau/4$	ε/K	$\tau/4$	$\tau/4$	$1-\tau-\varepsilon$	ε/K	$\tau/4$	ε/K	ε/K	...	ε/K

Hard Label

Original LS

Proposed LS

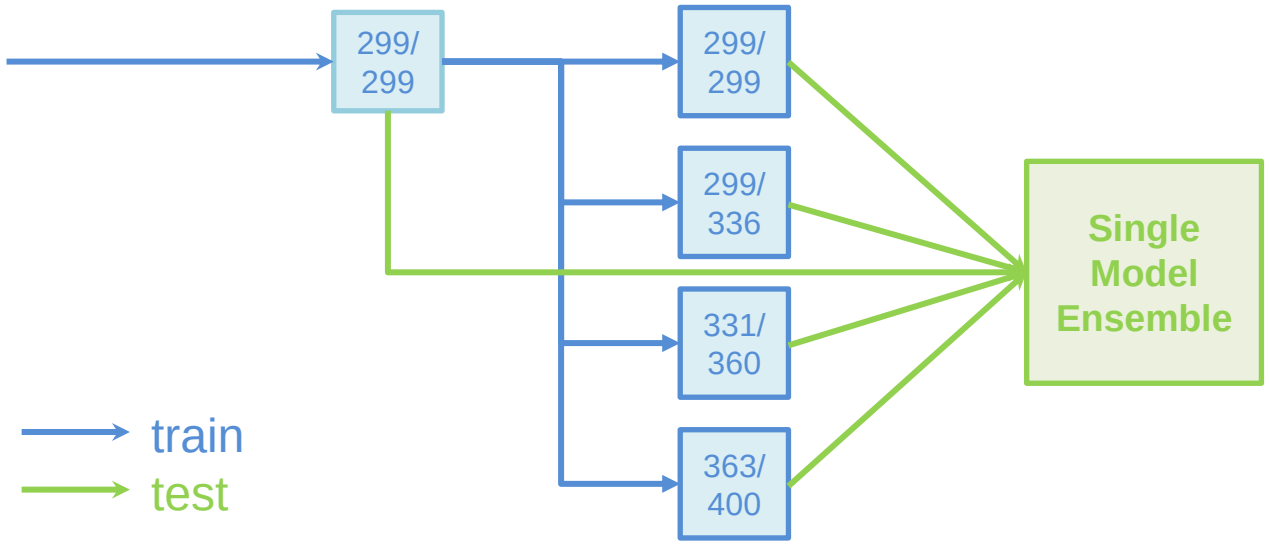
labels of similar classes used for label smoothing

[8] C. Szegedy, et al. Rethinking the Inception Architecture for Computer Vision. CVPR, 2016.

Scene Classification

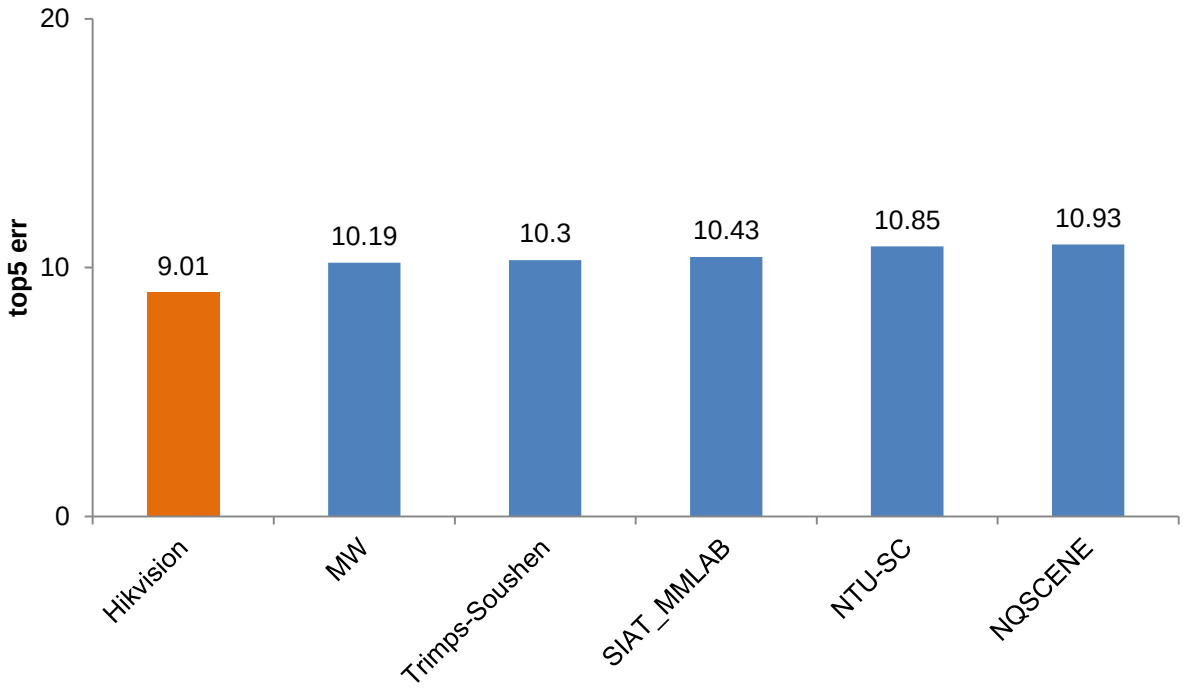
HIKVISION

- Train and Test in **Harmony**
  - train in the same way as you test
    - remove augmentation for the last several epochs (+0.3%)
  - test in the same way as you train
    - test 32 random crops from scale and ratio augmentation (+0.3%)
  - multi-scale testing over multi-scale training
  - use checkpoints from the middle of training to avoid overfitting (+0.3%)

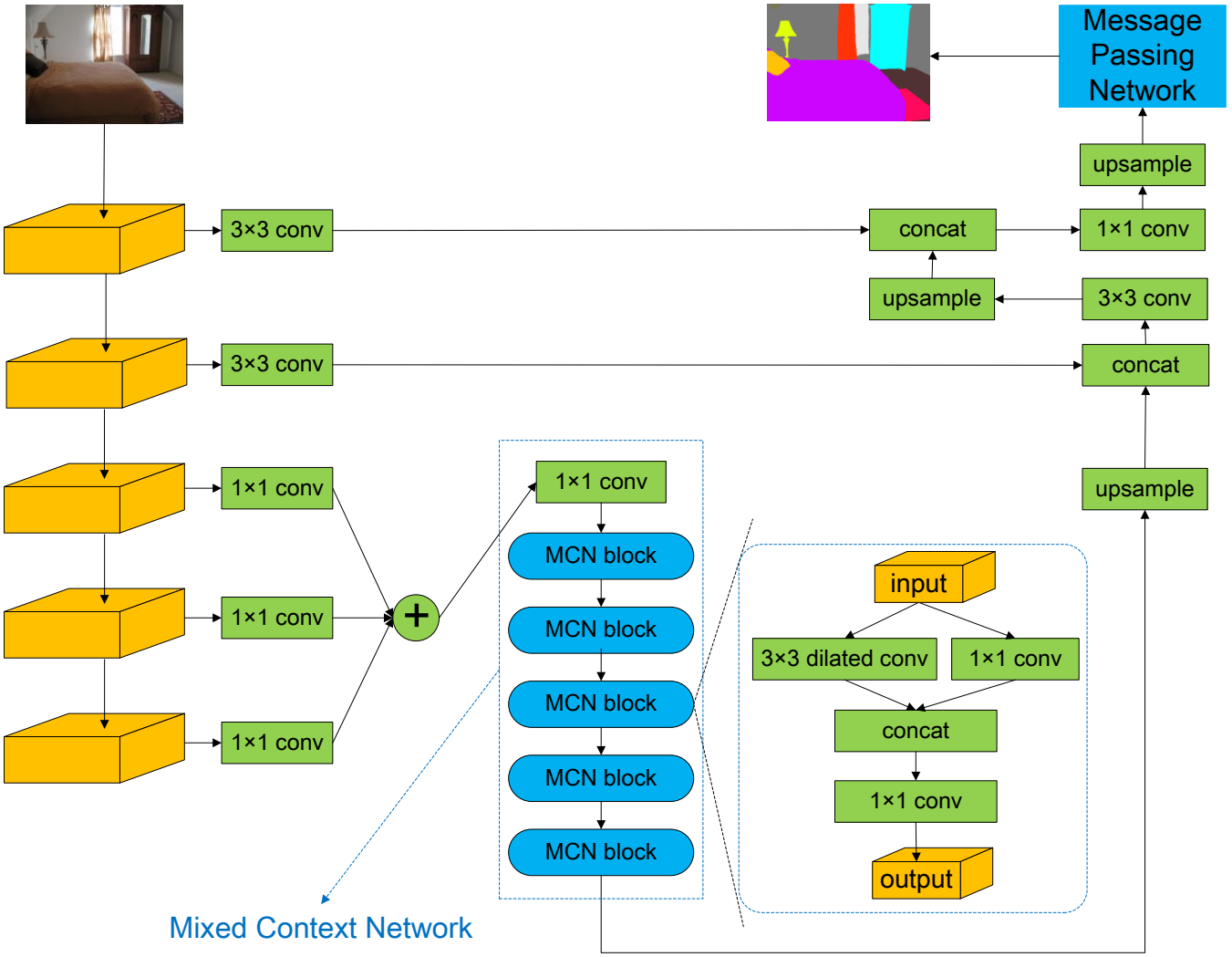


Scene Classification

- **Models**
  - Inception v3/Inception ResNet v2, and their variants
  - Wider ResNet with 50/64 layers
- **Results**



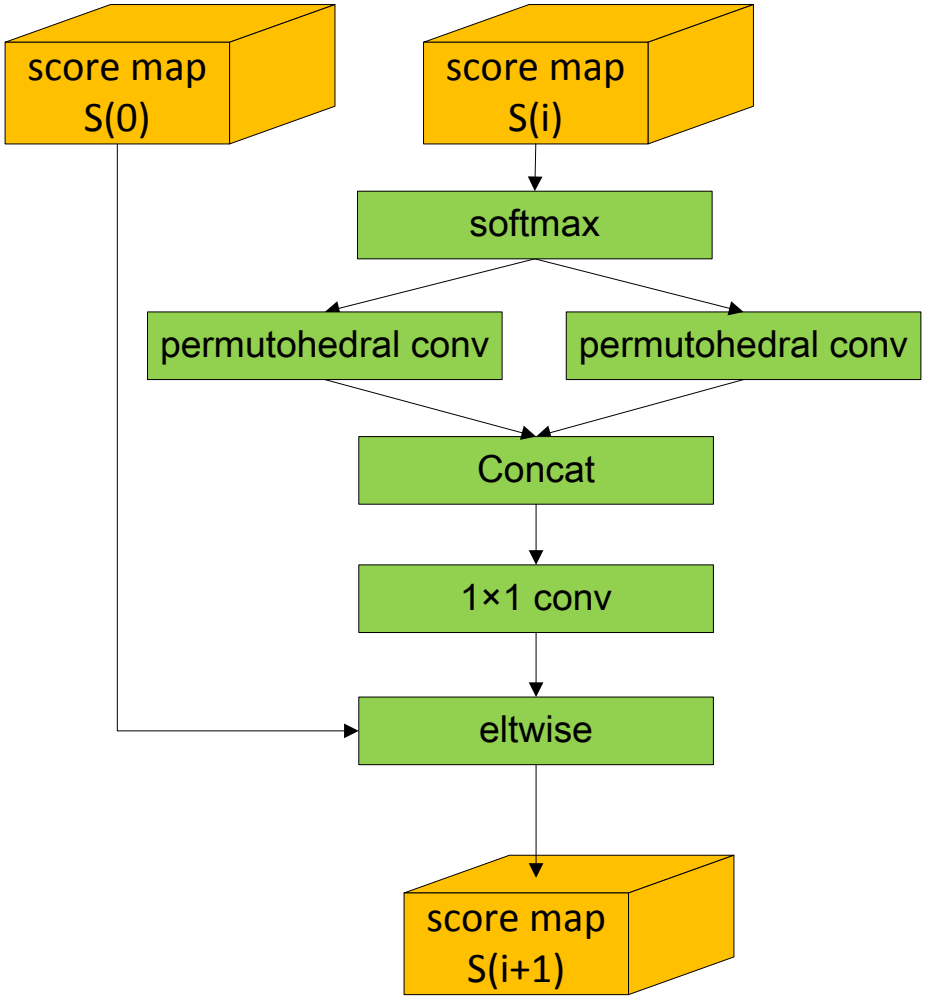
Scene Parsing



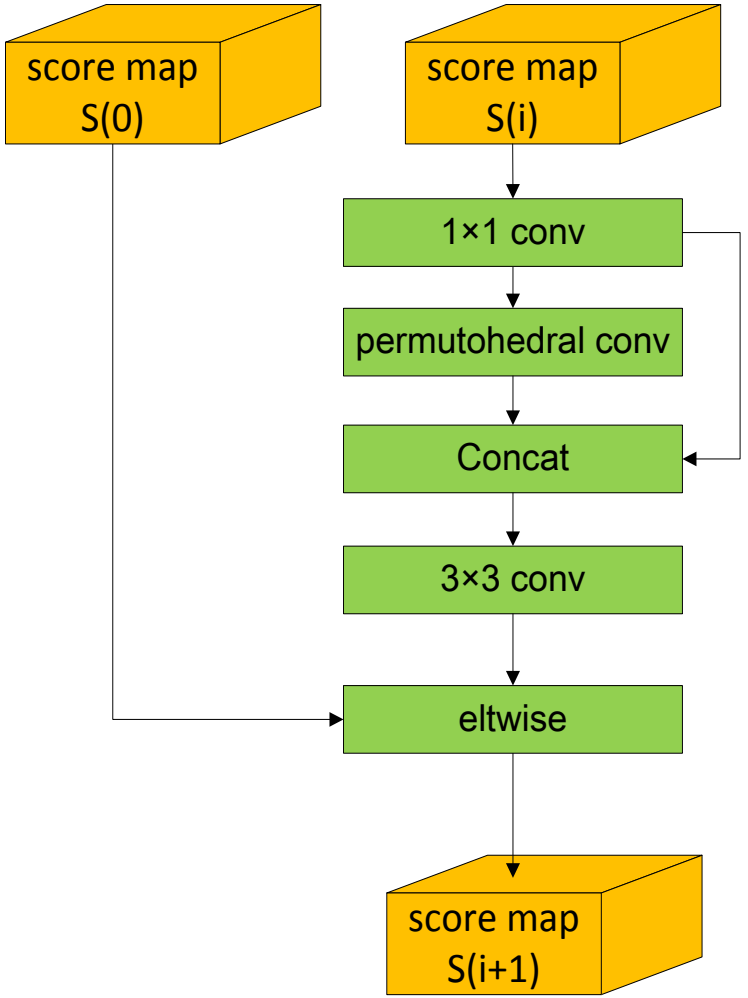
overall architecture for scene parsing

Scene Parsing

HIKVISION



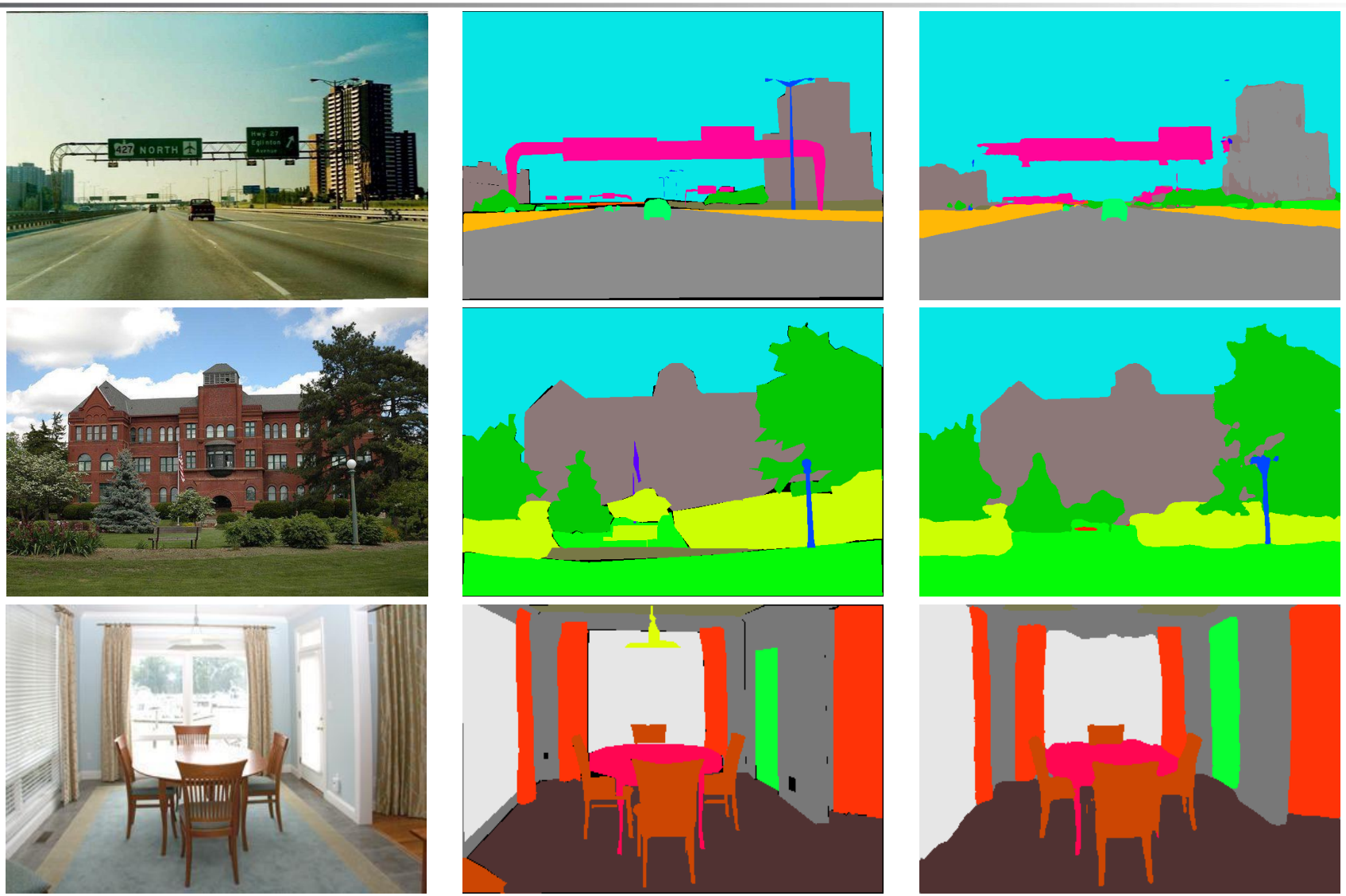
CRF as RNN



Message Passing Network

Parsing Results

HIKVISION



Object Detection Elements

HIKVISION

ResNet-101 Variants

facebook

Identity Mapping

Training Tricks

Balanced Sampling

Multi-Scale Training

OHEM

Testing Tricks

Multi-Scale Testing

HFlip

Box Voting

RPN Proposal

Cascade RPN

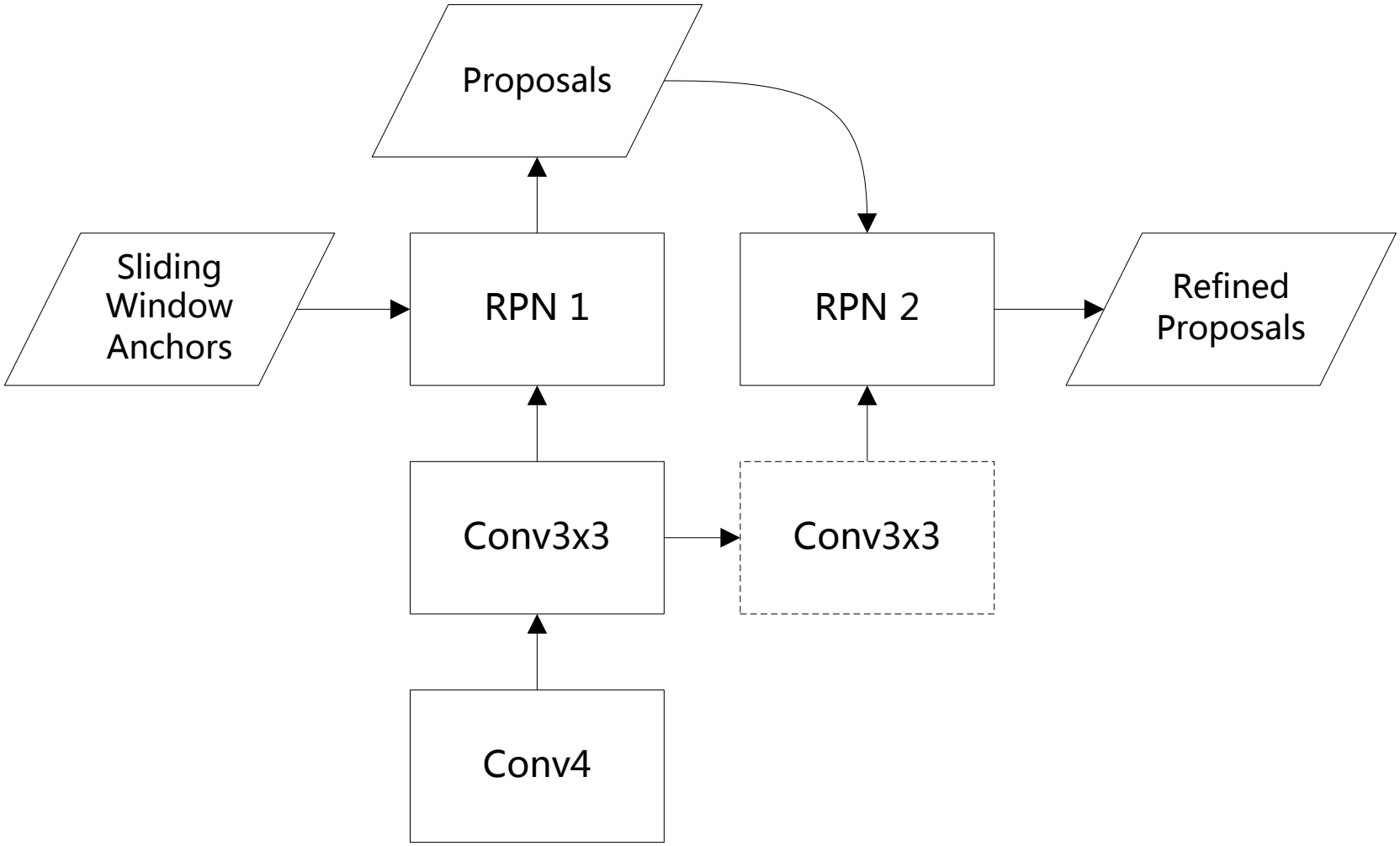
Constrained Neg/Pos Anchor Ratio

Pretraining

Pretrained Global Context

Pretrain on LOC

Cascade RPN


Proposals
Sliding Window Anchors
RPN 1
RPN 2
Refined Proposals
Conv3x3
Conv3x3
Conv4

Constrained NEG/POS Anchor Ratio

HIKVISION

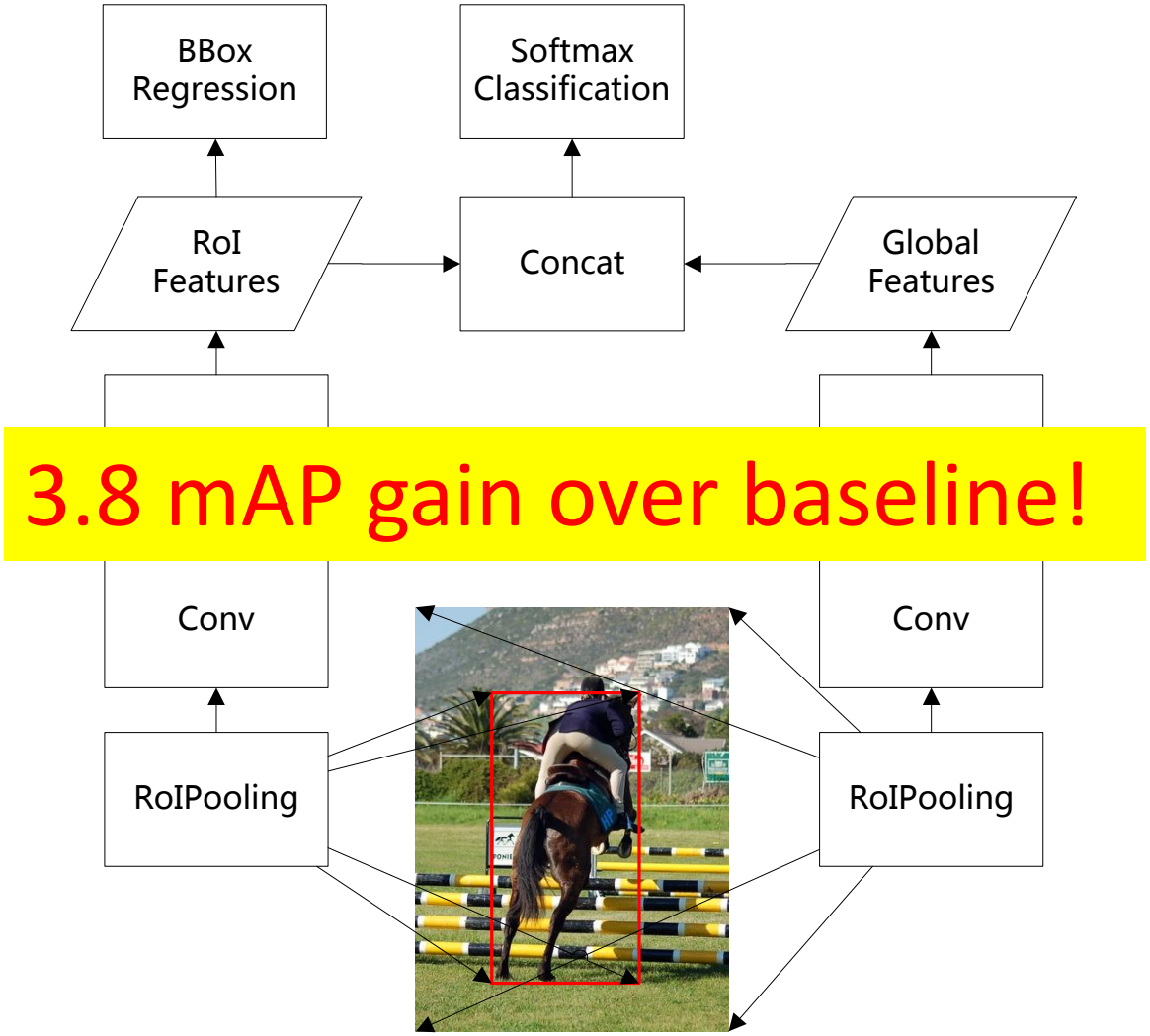
- Naïve RPN
  - Batch size: 256
  - Expected N/P ratio: 1
  - Real N/P ratio: usually **> 10**
- Our RPN
  - Min batch size: 32
  - Max N/P ratio: 1.5

Ablation Study				
Cascade RPN		✓		✓
Constrained N/P			✓	✓
Recall@0.5	91.0	91.2	92.0	91.9
Recall@0.7	70.2	77.9	74.0	79.7
Average Recall	52.5	57.2	54.6	57.9

5.4 AR gain

9.5 Recall@0.7 gain

Pre-trained Global Context



Pre-training on LOC

HIKVISION

0.5 mAP gain over baseline!

Balanced Sampling

- Adapted from Shen et al. [7] for detection task

[7] L. Shen, et al. Relay Backpropagation for Effective Learning of Deep Convolutional Neural Networks. ECCV, 2016.

Performance

HIKVISION

- ImageNet DET

	Team	mAP	Rank
Single Model	Hikvision	63.40	1
	CUImage	63.36	2
Ensemble	CUImage	66.3	1
	Hikvision	65.3	2

- PASCAL VOC2012

Team	mAP	Rank
Hikvision	87.9	1
ResNet Baseline	83.8	2

- ImageNet CLS-LOC

	CLS	LOC	Rank
LOC (Ensemble)	3.7	8.7	2

Take Home Message

HIKVISION

- **Scene Classification**
  - better utilize your data and model (SDA)
  - label smoothing via soft label
  - balanced sampling via label shuffling
  - train and test in harmony
- **Scene Parsing**
  - Mixed Context Net & Message Passing Net
- **Object Detection**
  - use identity mapping
  - cascade RPN
  - constrained NEG/POS anchor ratio
  - pre-trained global context
  - balanced sampling
- **Object Localization**
  - LOC = CLS + DET

Acknowledgments

- We would like to thank our HPC team:
  - Peng Wang
  - Jianfeng Peng
  - Xing Zheng
  - Zhiqiang Zhou
  - etc…

Thank you!