Understanding Scene in the Wild

SenseCUSceneParsing

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi
Xiaogang Wang, Tong Xiao, Jiaya Jia

SenseTime 商湯集團

香港中文大學
The Chinese University of Hong Kong

Features of ADE20K Dataset

- Number of image
  - ADE20K Dataset: 20k
- Number of scene
  - Image label: 1038
- Number of semantic label
  - Wall / Building, Field / Earth, Mountain / Hill, Stair / Stairway……

Our baseline

- Pretrained Resnet 101 + FCN pixel prediction
- Result regarding to mIOU / pixel accuracy
  - Train Data: 70.20 / 90.34
  - Val Data: **35.08/76.87**
- Of course, we have many pre-baselines that are not so good

Our result improves

— Evils in the details

- Various data augmentation
- Dropout to the last convolution layers
- Using dilated convolution
- Learning rate policy
- Total iteration number
- Correct way to use batch normalization
- Larger cropsize and larger receptive field
- ……

Code and model will be released later

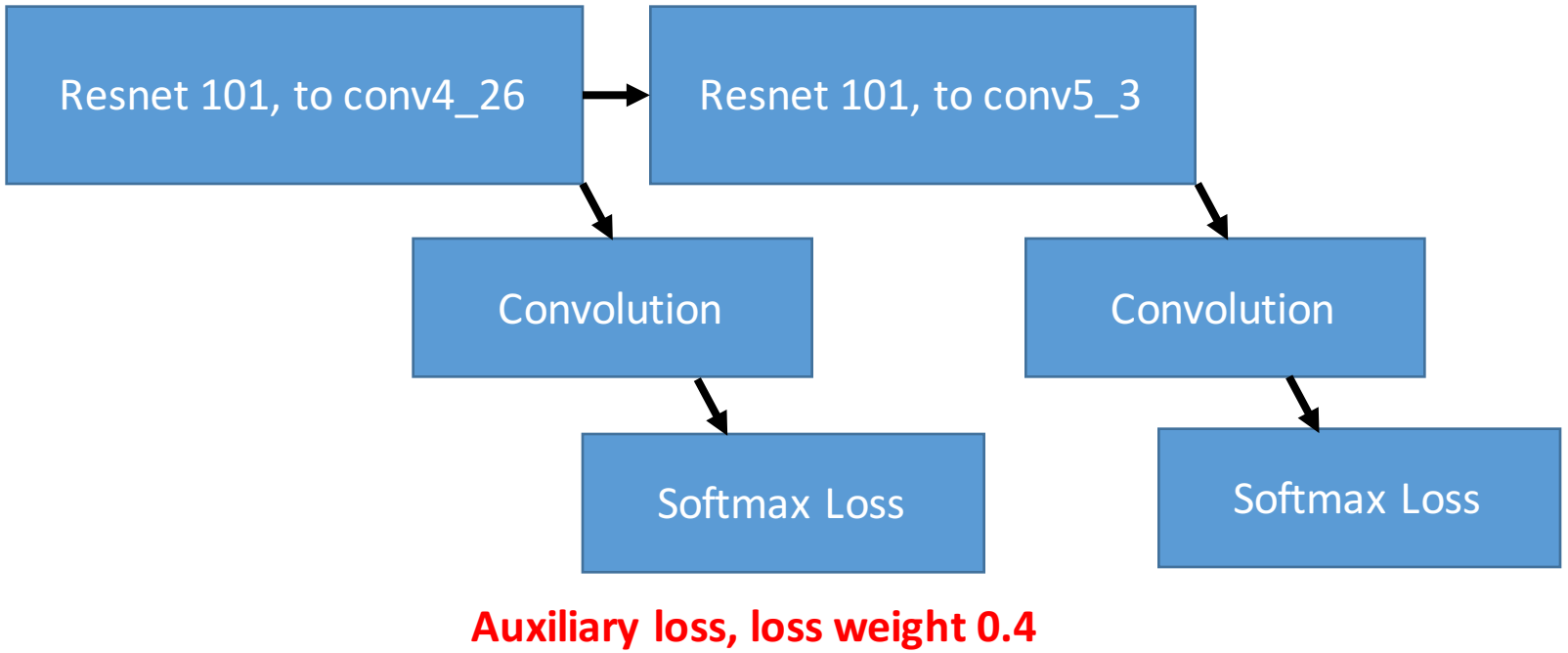
Our result improves

— Evils in the details

- Previous baseline result by mIOU / pixel accuracy
  - Train Data: 70.20 / 90.34
  - Val Data: **35.08/76.87**
- Current Resnet101 result
  - Train Data: 75.16/91.99
  - Val Data: **36.85/77.65**

Our result improves

- Deeply supervise for better optimization



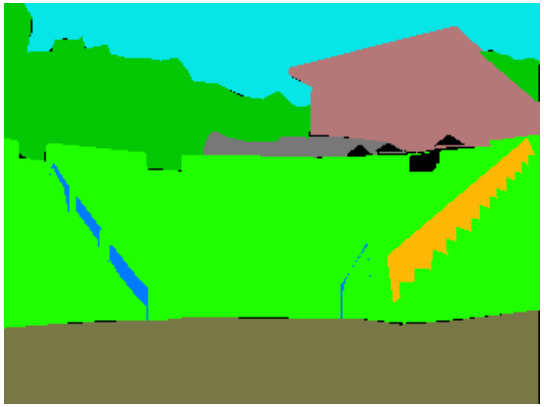
Our result improves

— Deep resnet improves by additional loss

- Previous result by mIOU / pixel accuracy
  - Train Data: 75.16/91.99
  - Val Data: **36.85/77.65**
- Current result by deeply supervised training
  - Train Data: 77.70/93.15
  - Val Data: **38.28/78.63**
- Better optimization policy improves confusion label and inconspicuous object

Image level information may help scene parsing

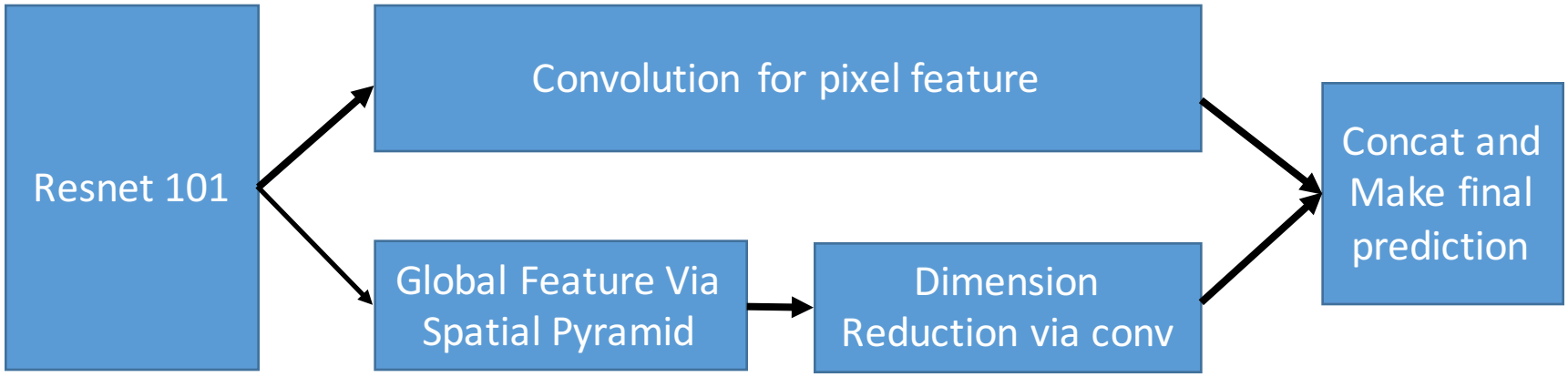
- A failure example



Recognize scene in image level

- State-of-the-art Image classification
  - FCN + Average Pooling
- Classical scene understanding
  - Spatial Pyramid Matching
- Better scene recognition
  - FCN + Spatial Pyramid Matching Pooling

Pixel Prediction with Image Level Information



- Utilizing image level information for scene parsing
- End to end learning
- Marginal computation cost

Our result improves

— Spatial pyramid global feature

- Previous result by mIOU / pixel accuracy
  - Train Data: 77.70/93.15
  - Val Data: **38.28/78.63**
- Current result by image level information
  - Train Data: 79.51/93.65
  - Val Data: **41.29/80.04**
- Global feature improves error failure to sense image label

Deeper Pretrained Model

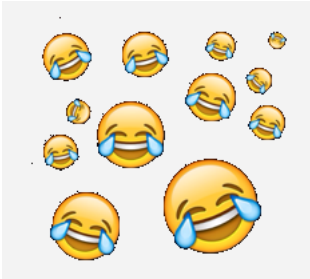
Pretrained Model	Result
Resnet 50	40.11/79.55
Resnet 101	**41.29/80.04**
Resnet 152	42.23/80.46
Resnet 269	**43.39/80.90**

- Better and Deeper pretrained model improves the result consistently

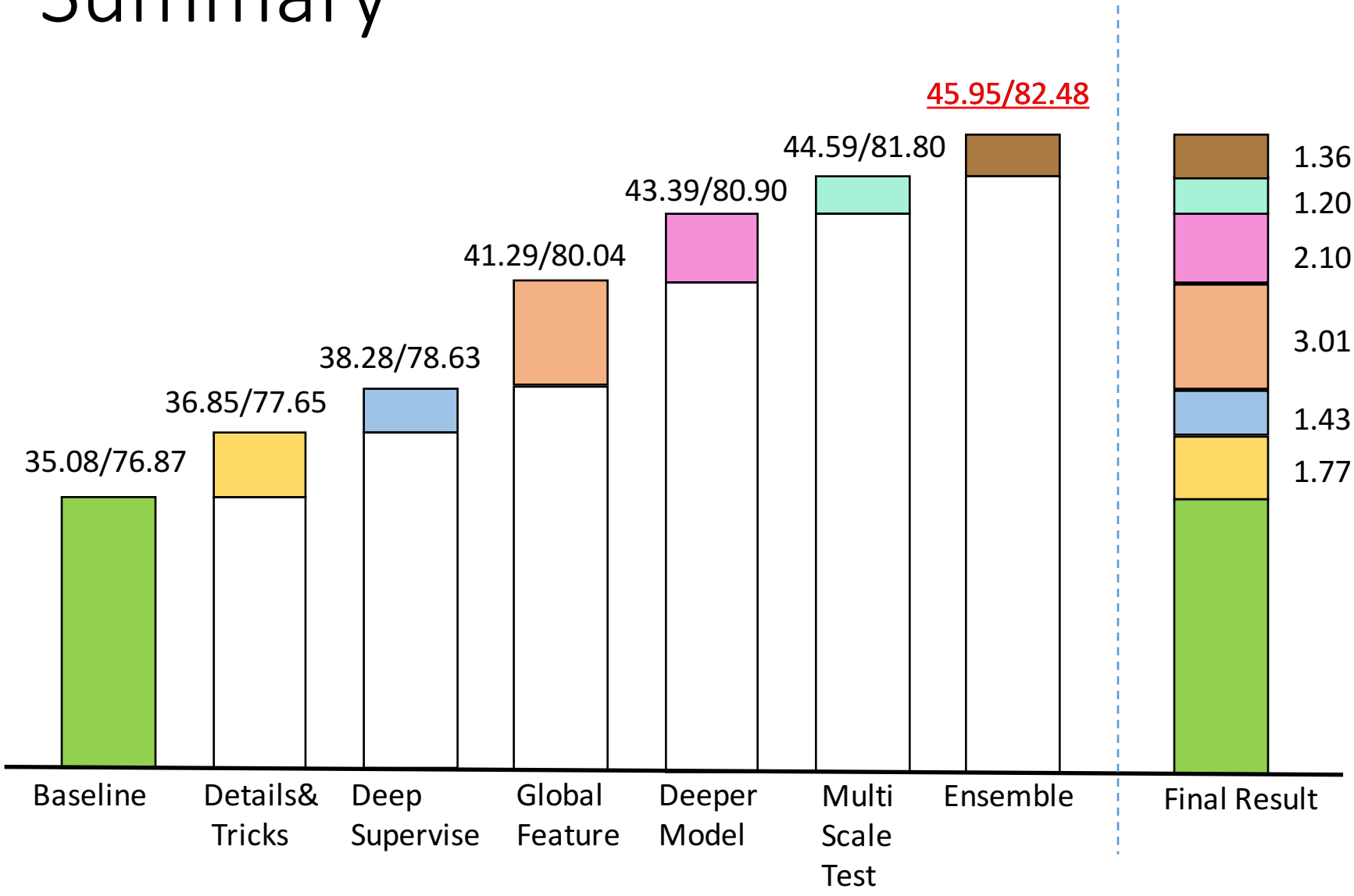
Testing and Ensemble

Method	Result
Resnet 269 Single Scale Test	**43.39/80.90**
Resnet 269 Multi Scale Test	44.59/81.80
Ensemble of 5 Models	**45.95/82.48**

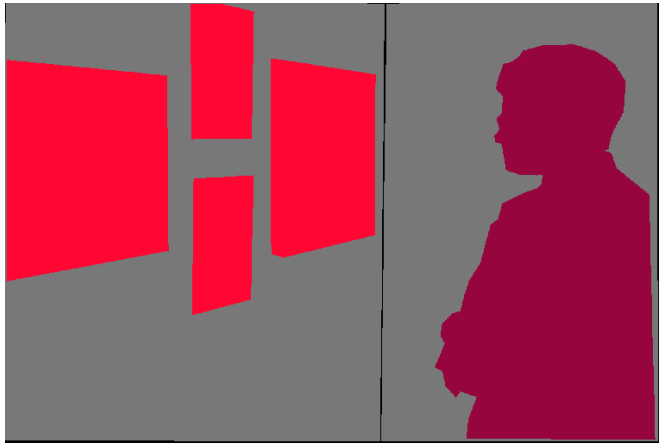
- Proper testing scheme improves the result
- But it is time consuming and only useful for competitions



Summary


45.95/82.48
44.59/81.80
43.39/80.90
41.29/80.04
38.28/78.63
36.85/77.65
35.08/76.87
1.36
1.20
2.10
3.01
1.43
1.77
Baseline
Details& Tricks
Deep Supervise
Global Feature
Deeper Model
Multi Scale Test
Ensemble
Final Result

Visual Results



wall
person
painting

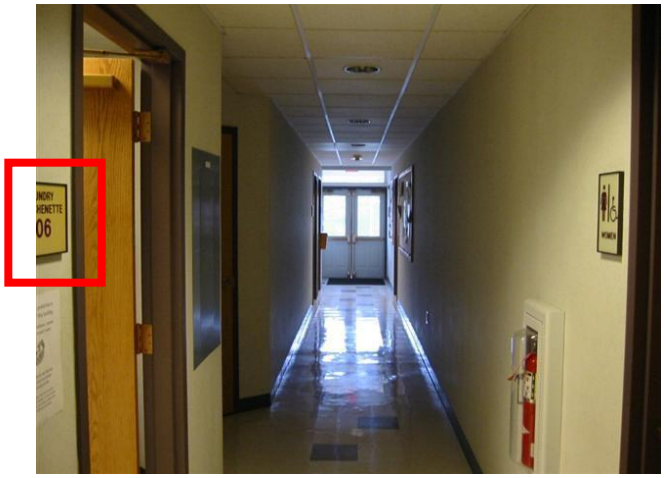
Visual Results

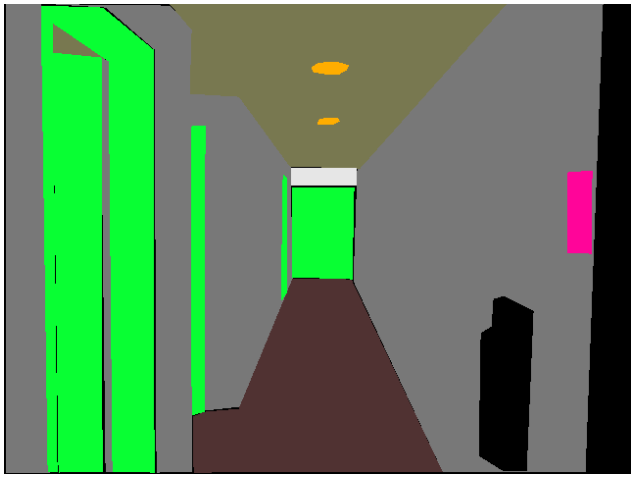
wall	floor
ceiling	bed
windowpane	door
curtain	rug
wardrobe	lamp
cushion	

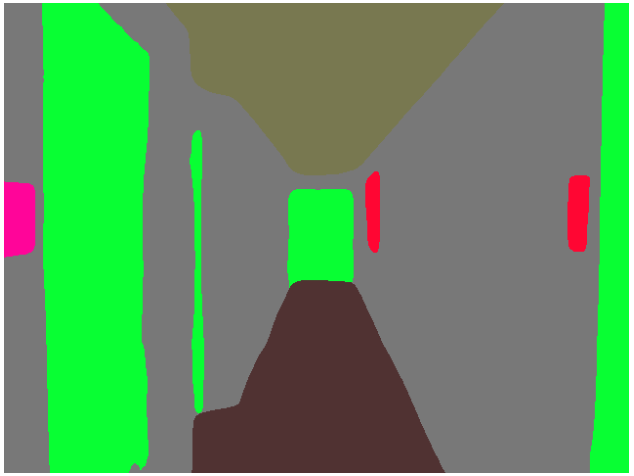
Visual Results

wall	building
sky	tree
road	earth
plant	fence
streetlight	

Visual Results

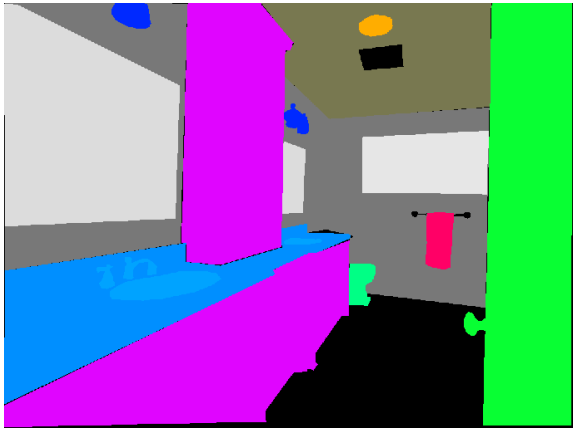


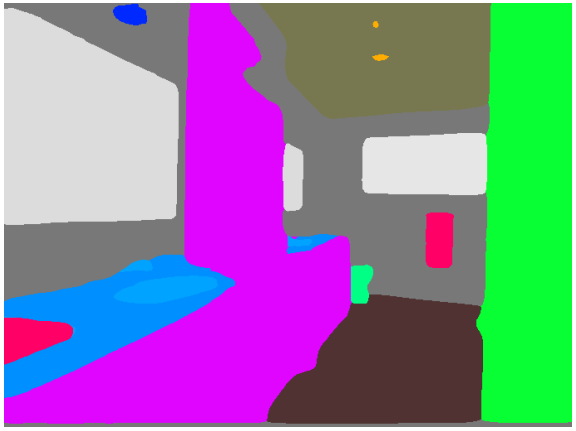




wall	floor
ceiling	windowpane
door	painting
signboard	light

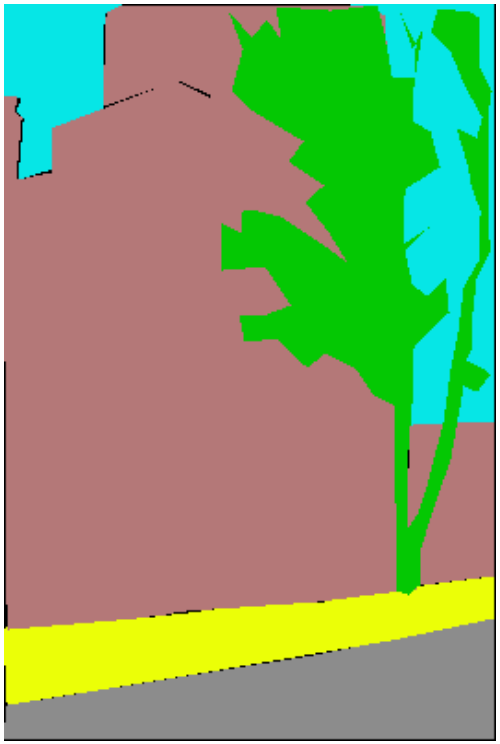
Visual Results

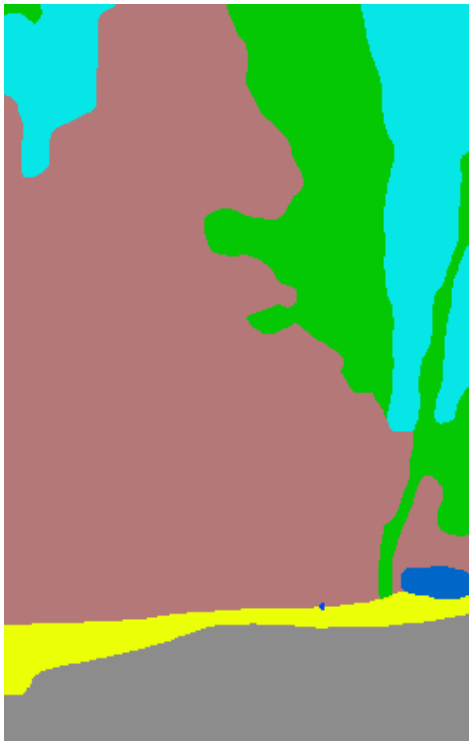




wall	floor
ceiling	windowpane
cabinet	door
mirror	sink
toilet	countertop
towel	light
sconce	

Visual Results





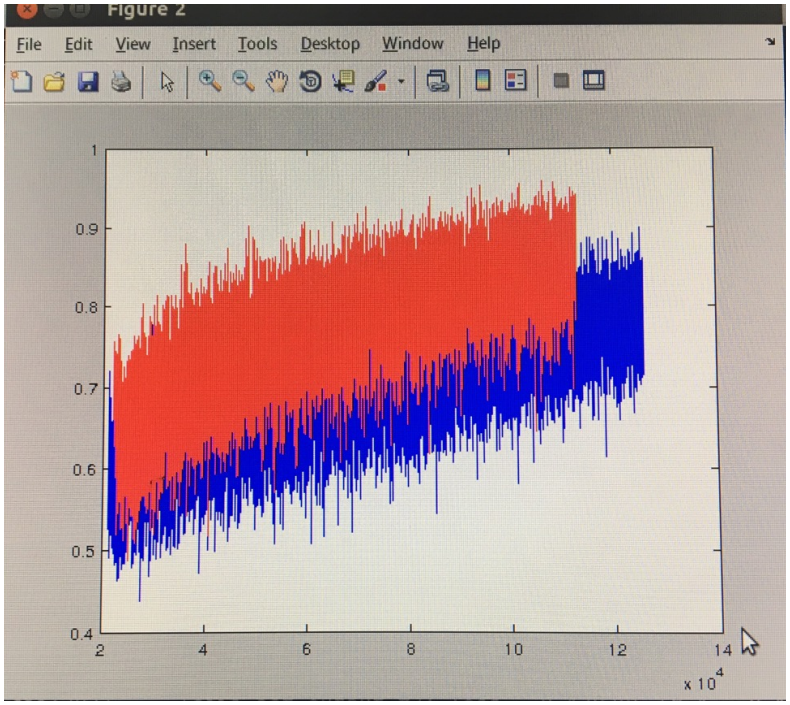
Future Direction

- More labeled data, and more clear definition
- Use of human semantic similarity matrix
- Small and rare object
- Scene parsing in video
- Speedup

It is not yet finished…

Learn by failure – Balance Sample

- Sample training image to uniform distribution
- Better training accuracy
- But overfitting and worse validation accuracy



Blue is baseline training accuracy,
Red is training accuracy after balance sample

Learn by no significant improvement

- Hard sample mining
- CRF
- Stochastic depth
- Using predefined class correlation
- ……

Thanks & Questions