SenseCUSceneParsing: Understanding Scene In The Wild

Hengshuang Zhao[1,2] Jianping Shi[1] Xiaojuan Qi[2] Xiaogang Wang[2] Tong Xiao[2] Jiaya Jia[2]

SenseTime Group Limited[1] The Chinese University of Hong Kong[2]

Problem Definition

Predict the label of each pixel in wild scene

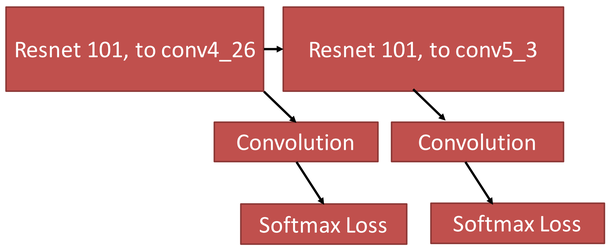
Our Baseline

- Pretrained Resnet101
- Fully Convolutional Network
- Dilated convolution with stride 8

Evils in the Details

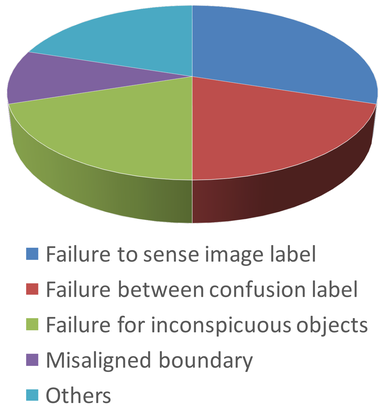
- Various data augmentation
- Add dropout to the last convolution layers
- Using dilated convolution
- Learning rate policy
- Total iteration number
- Correct way to use batch normalization
- Larger cropsize and larger receptive field
- etc.

Deeply Supervised Loss



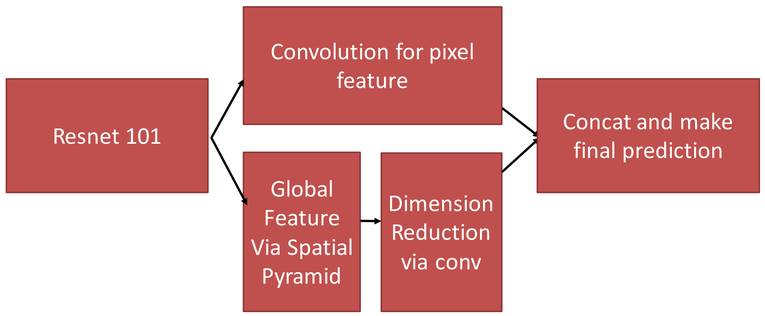
Auxiliary loss, loss weight 0.4

Error analysis for our baseline model



Scene Parsing by Scene Understanding

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



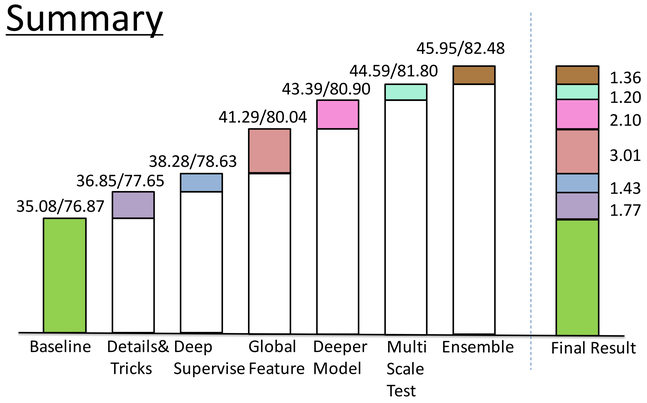
Deeper Pretrained Model

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

Test & 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

Summary



Learn by Failure

- Sample training image to uniform distribution
- Hard sample mining
- CRF
- Stochastic depth
- Stuff / object training
- ASPP
- Using predefined class correlation

Visual Result