Places Challenge 2017
Scene Parsing

WinterIsComing

Riwei Chen, Qi Chen, Xinglong Wu

Yifan Lu, Yudong Jiang, Linfu Wen

TOUTIAO AI LAB

今日头条人工智能实验室

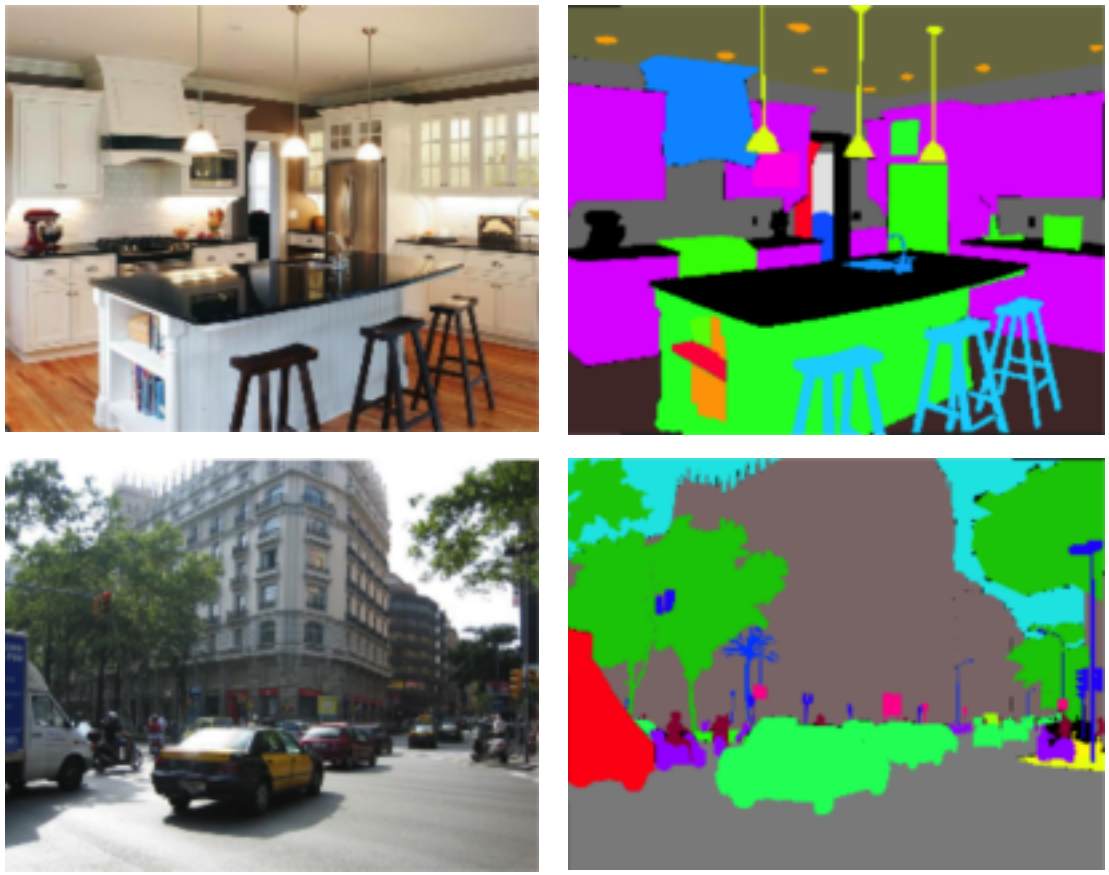
ByteDance 字节跳动

Outline

- Single Model Results
- Method Overview
- Method Details
  - Model Pretraining
  - Pyramid Pooling
  - Batch Size & BN
  - Other details
  - Submissions
- Visual Results
- Future Direction

Features of ADE20K Dataset—Scene Parsing

- Number of image
  - Training: 20K
  - Validation: 2K
  - Testing: 3K
- Number of category
  - Semantic category: 150



Single Model Results on Validation Set

- Single model
  - Compared with the best single model result of 2016

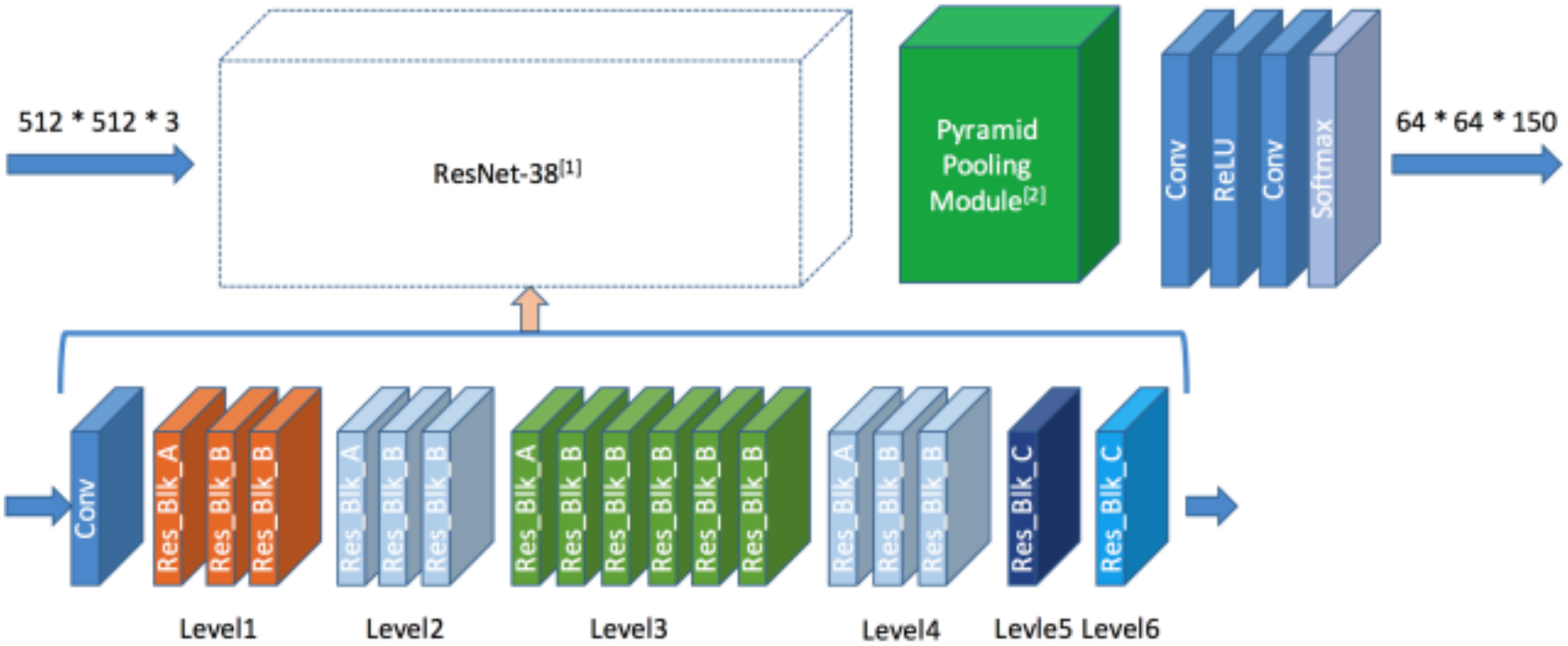
Team	mIoU	pixel accuracy
SenseCuSceneParsing[1]	43.39%	80.90%
Adelaide[2]*	43.06%	80.53%
WinterIsComming(ours)	**43.98%**	**81.13%**

[1] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network, CVPR 2017
[2] Wu Z, Shen C, Hengel A V D. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. 2016
* The result of "Model C, 2 conv"

Method Overview

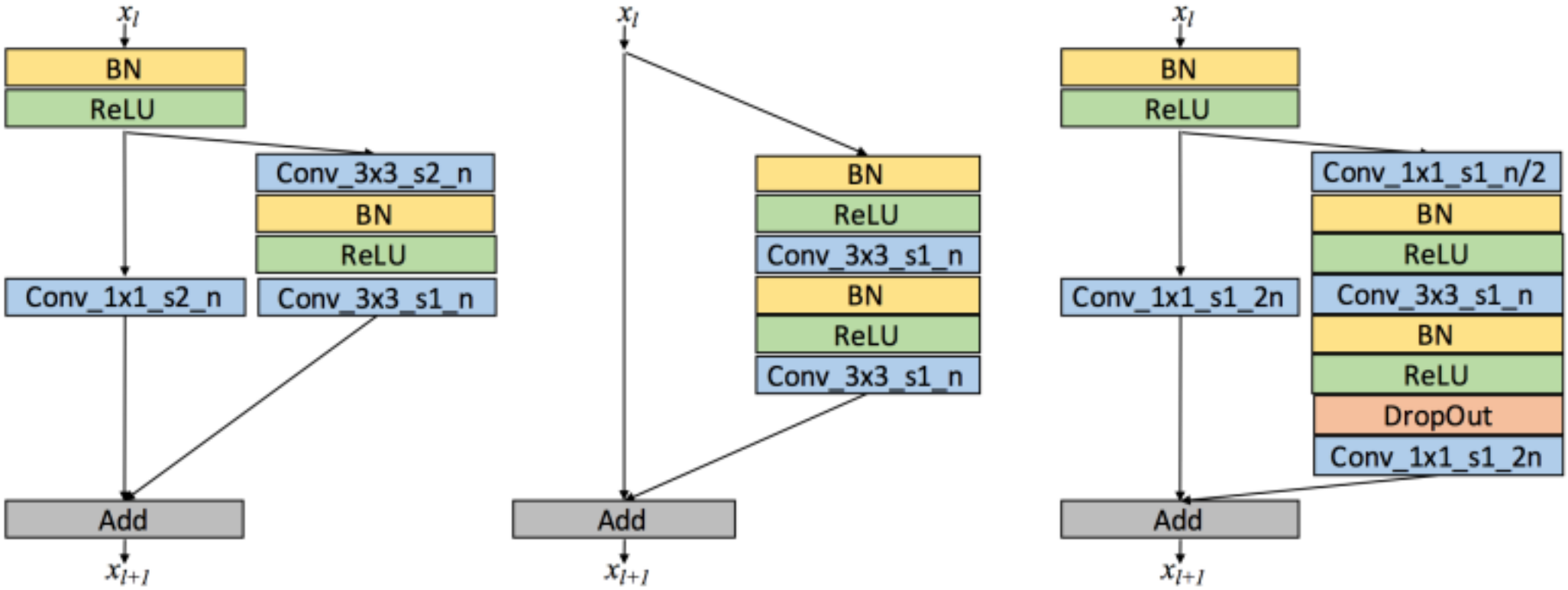
- Base Network: ResNet38
- Pyramid Pooling
- ImageNet and Places2 pretraining
- Batch Size is critical
- Ensemble models trained with different epochs

Network Structure



[1] Wu Z, Shen C, Hengel A V D. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. arXiv 2016
[2] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network, CVPR 2017
* Our implement is based on: https://github.com/itijyou/ademxapp

Building Blocks

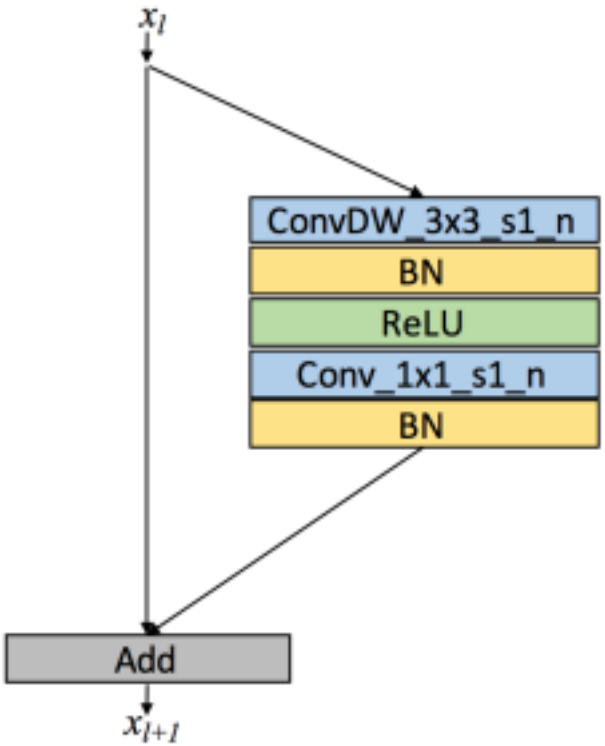


(a) Res_Blk_A

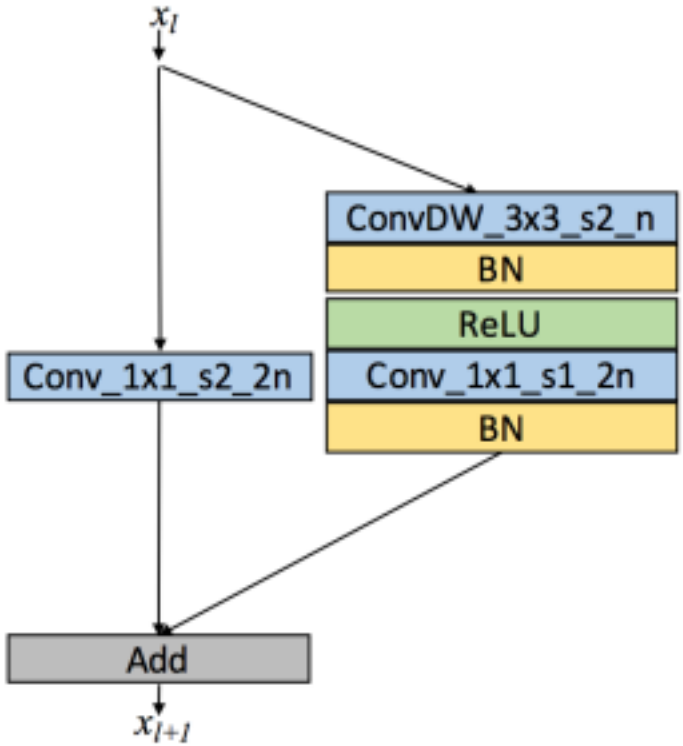
(b) Res_Blk_B

(c) Res_Blk_C

Res-MobileNet



(a) Res_Mobile_Blk_A

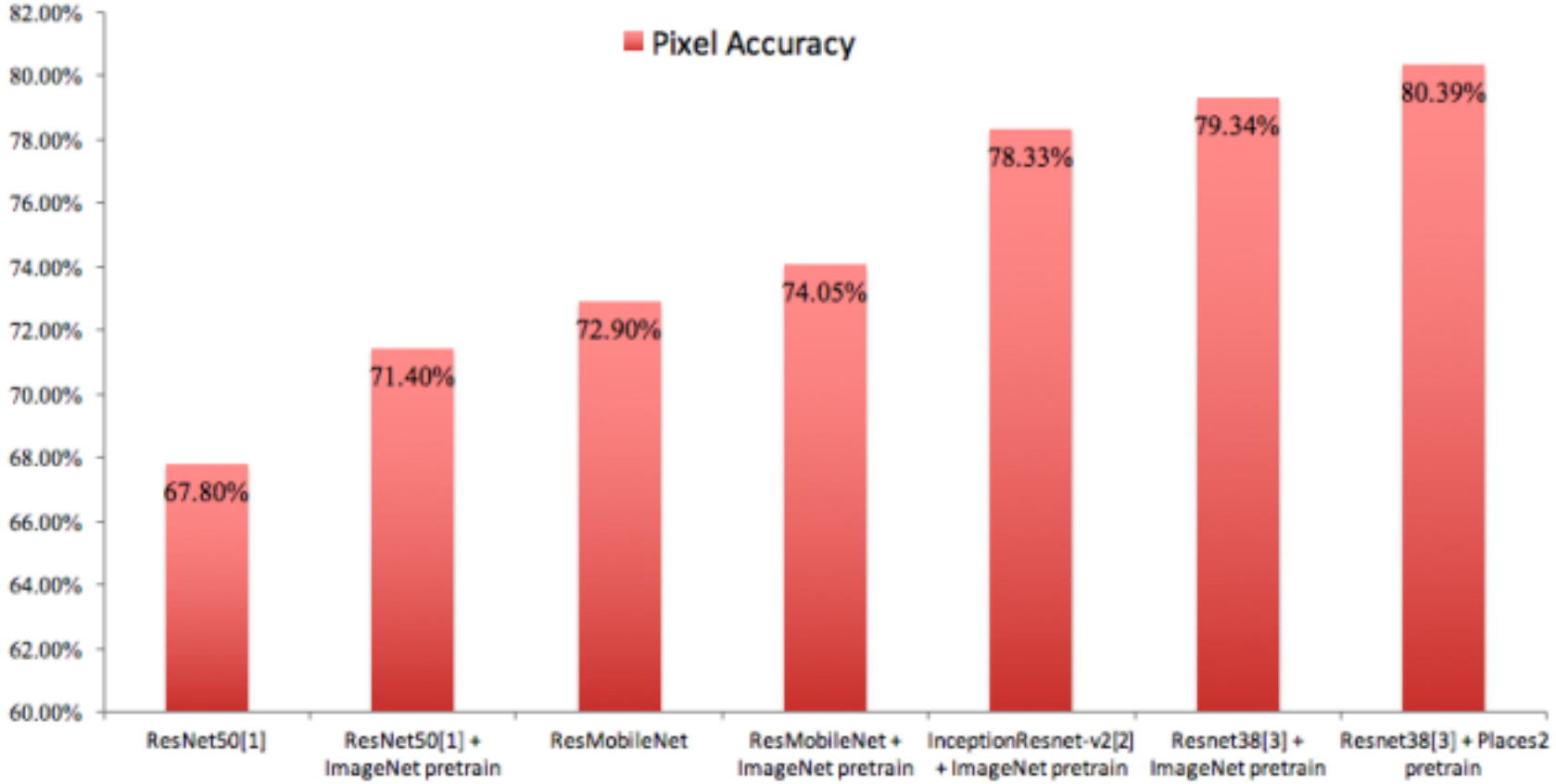


(b) Res_Mobile_Blk_B

Model	computation (macc)
ResNet50	109.4G
Res-MobileNet	**32.5G**
ResNet38	415.5G
VGG16	618.0G

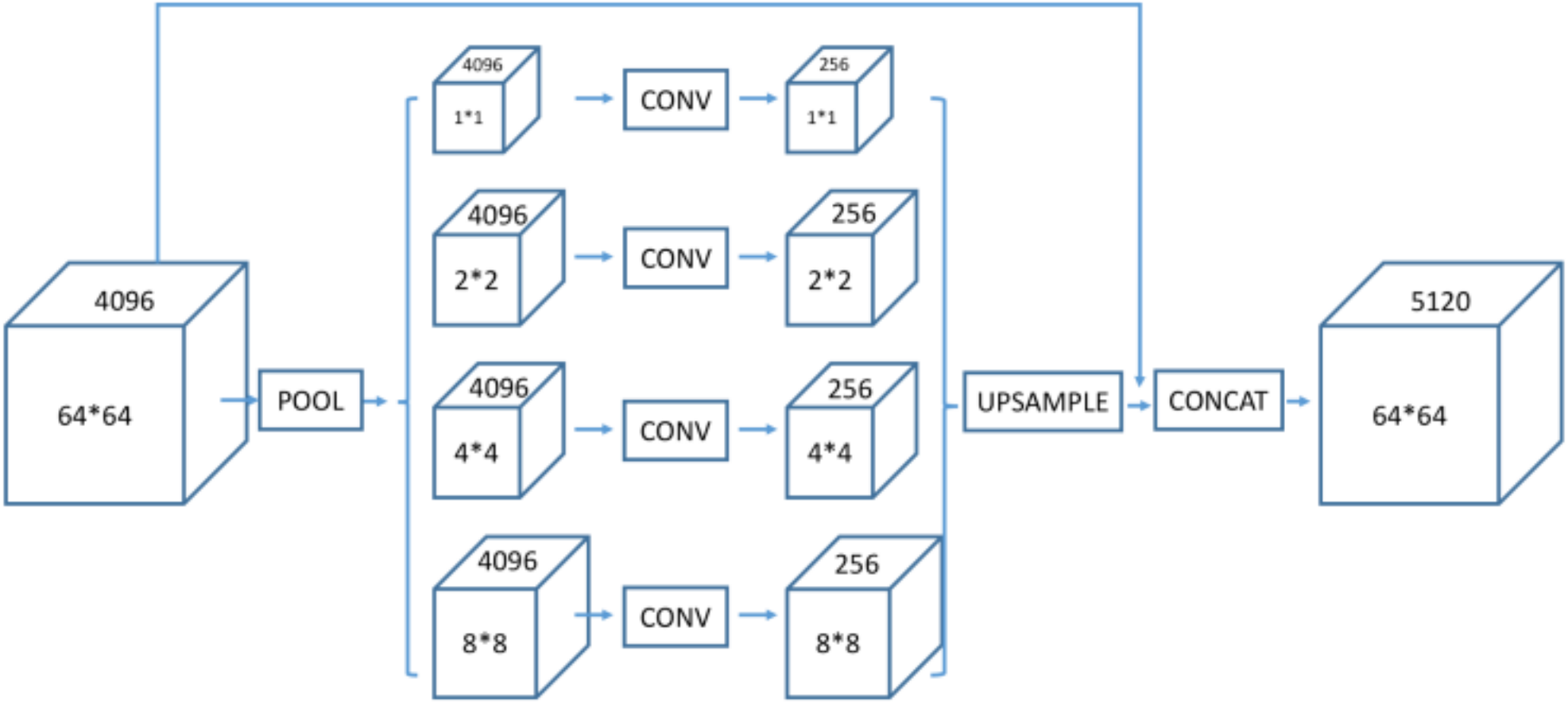
* The computation cost of models when input size is 512x512

Model Performance



[1] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network, CVPR 2017
[2] Szegedy C, Ioffe S, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv 2016
[3] Wu Z, Shen C, Hengel A V D. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. arXiv 2016

Pyramid Pooling



Pyramid Pooling

- Pyramid Pooling improves the integrity of segmentation

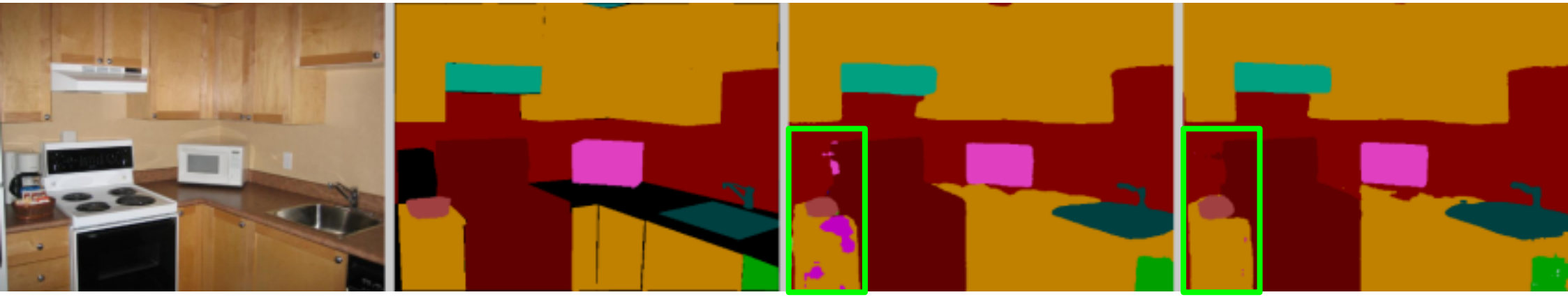
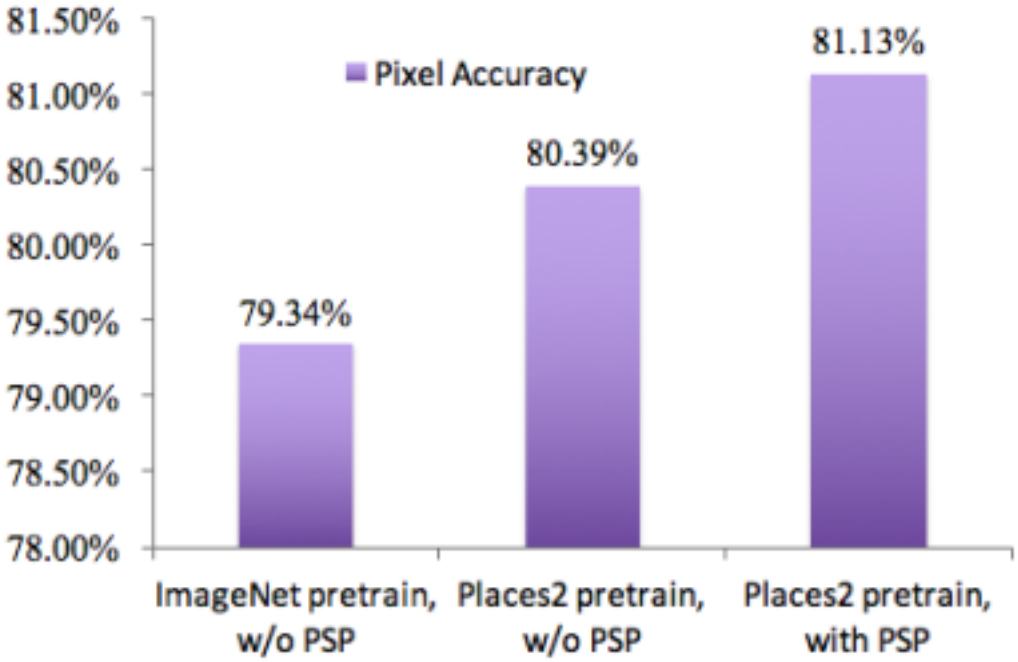
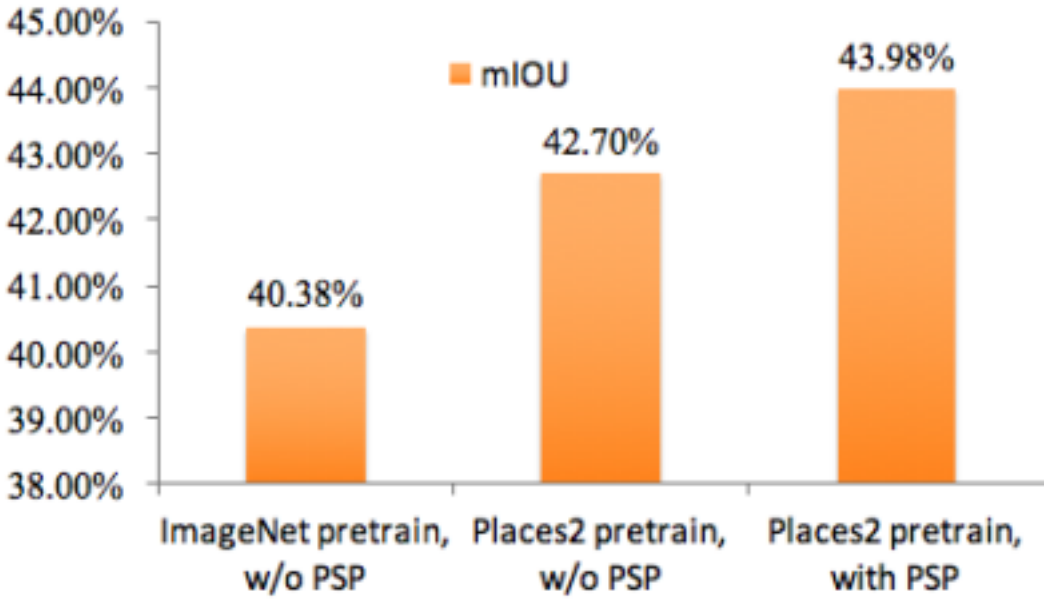


Image | Ground Truth | without Pyramid Pooling | with Pyramid Pooling

Pretraining

- ResNet50 without ImageNet pretraining has the lowest accuracy
- Places2 pretraining helps improve accuracy





Batch size & Batch Norm

- Training batch size is critical
- Experiment with Res-MobileNet
- ResNet38 w/o PP, batch size = 6
- After adding PP, batch size = 2
- Usually use 4 GTX 1080Ti GPUs

Training Batch Size per GPU	Testing Pixel Accuracy
1	68.4%
2	69.7%
4	70.7%
finetune with fixed BN	72.9%
finetune ImageNet pretrained model with fixed BN	74.1%

Other Details

- Training augmentation
  - Multi-scale: [0.7, 1.3]
  - Flip
  - Random crop to 512x512
- Testing augmentation
  - Flip
  - No multi-scale
- SGD solver with lr = 1e-4 for 64 epochs

Submissions

- Submit 1: train with only ADE20K training set
  - we get 81.13%/43.98% pixel accuracy/mIOU on validation set
- Submit 2-4: finetune the model with both training and validation set for 5, 22, 29 epochs respectively
- Submit 5: ensemble submit 1-4 models by voting

Summary

- Pretraining is critical and datasets of similar tasks work better
- Batch size should be large enough
- Fix BN params can further improve result (when batchsize is small)
- Pyramid Pooling can improve region integrity of segmentation

Visual Results

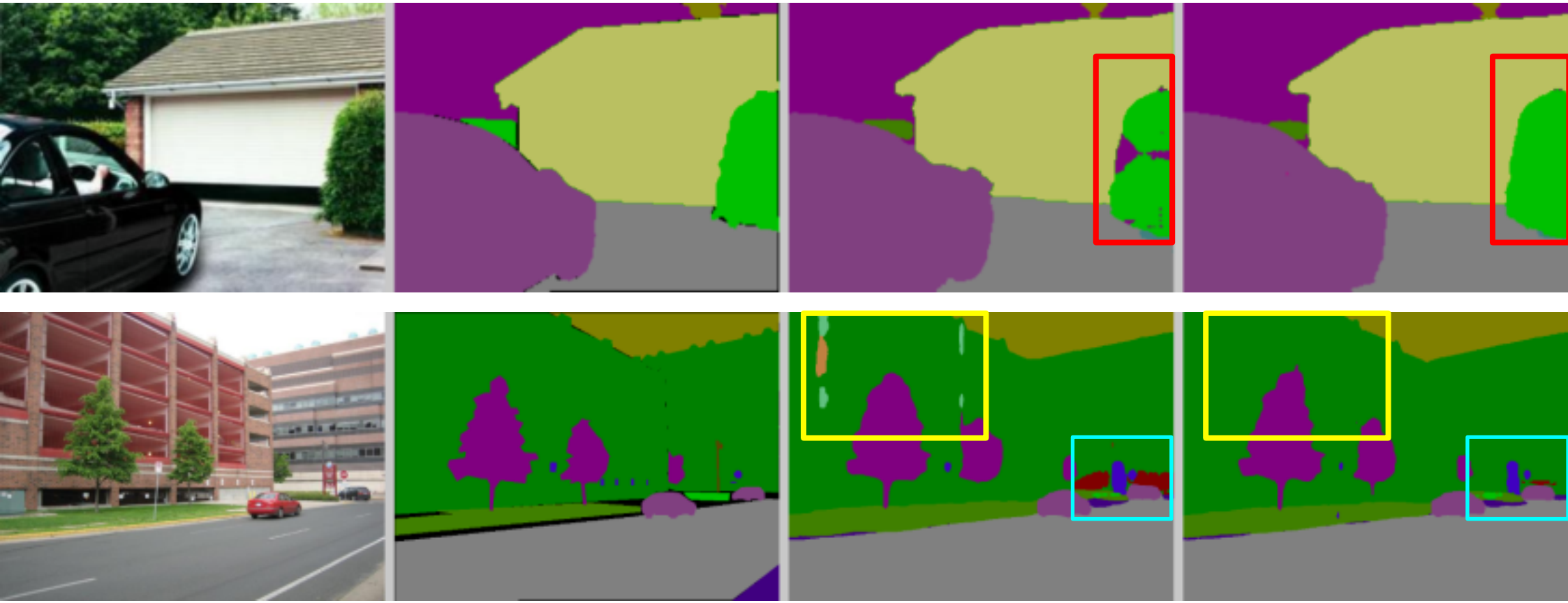


Image | Ground Truth | without Pyramid Pooling | with Pyramid Pooling

Future work

- Memory-efficient deep learning framework
- Well-Pretrained Res-MobileNet
- Focal loss
- Expert model

Thanks & Questions