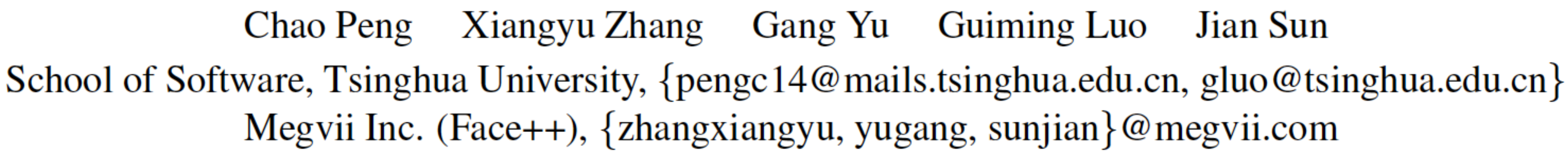
Large Kernel Matters—— Improve Semantic Segmentation by Global Convolutional Network



Chao Peng Xiangyu Zhang Gang Yu Guiming Luo Jian Sun
School of Software, Tsinghua University, {pengc14@mails.tsinghua.edu.cn, gluo@tsinghua.edu.cn}
Megvii Inc. (Face++), {zhangxiangyu, yugang, sunjian}@megvii.com

82.2% on PASCAL VOC 2012 and 76.9% on the Cityscapes.

Two Challenges in Seg

- 1) classification : an object associated to a specific semantic concept should be marked correctly.
- 2) localization : the classification label for a pixel must be aligned to the appropriate coordinates in output score map.

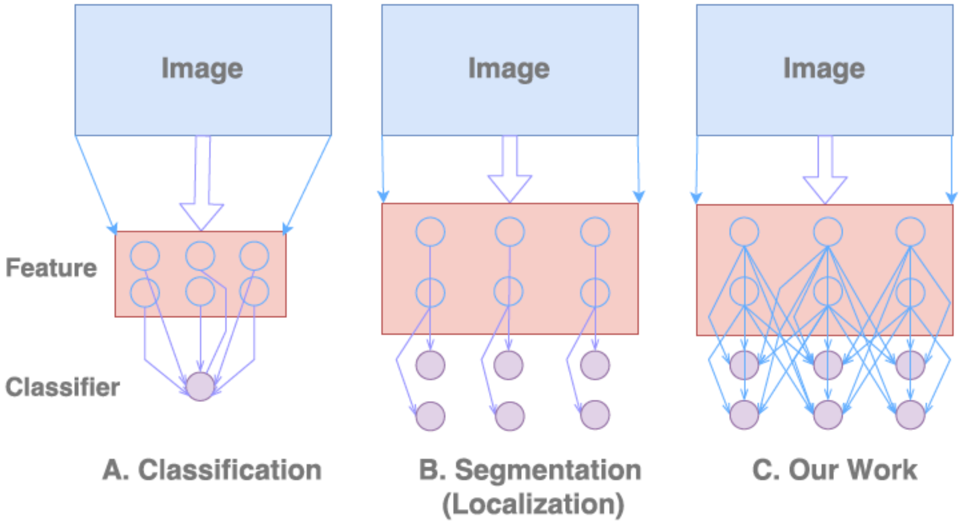


Figure 1. A: Classification network; B: Conventional segmentation network, mainly designed for localization; C: Our Global Convolutional Network.

Contributions

- 1) we propose Global Convolutional Network for semantic segmentation which explicitly address the “classification” and “localization” problems simultaneously.

- 2) a Boundary Refinement block is introduced which can further improve the localization performance near the object boundaries.

- 3) we achieve state-of-art results on two standard benchmarks, with 82.2% on PASCAL VOC 2012 and 76.9% on the Cityscapes.

Conventional method

- Multi-Context Embedding
  - PaserNet, Dilated-Net, ASPP in Deeplabv2
- Resolution Enlarging
  - Unpooling in Segnet
  - Dilation Convolutuin
  - Deconvolution
  - Bilinear Upsampling(LRR)
- Boundary Alignment
  - denseCRF
  - CRFasRNN

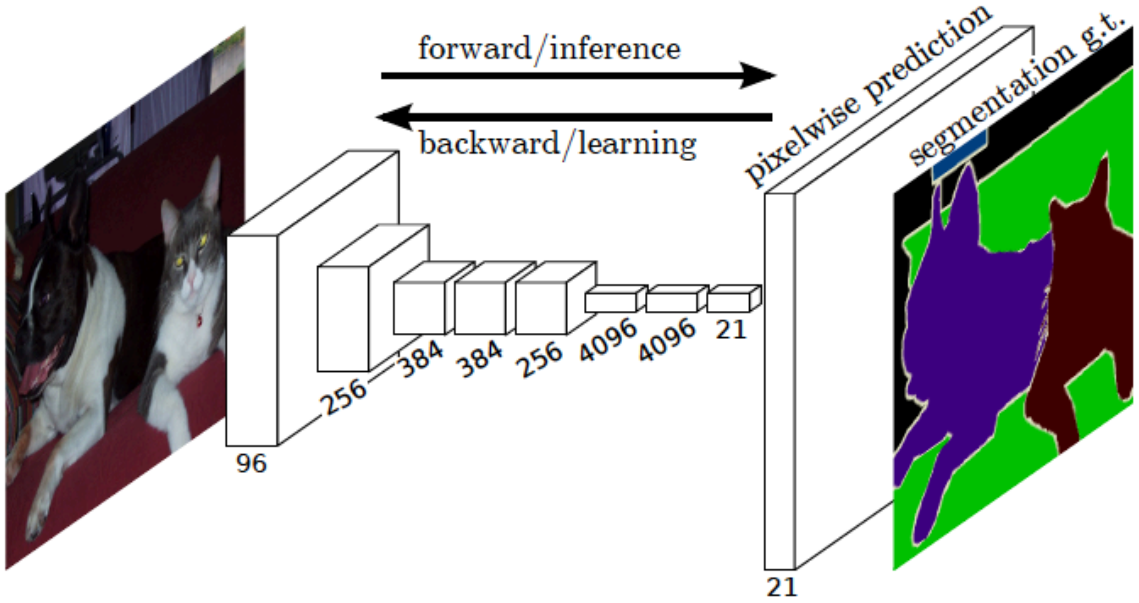


Figure 1. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation.

Framework

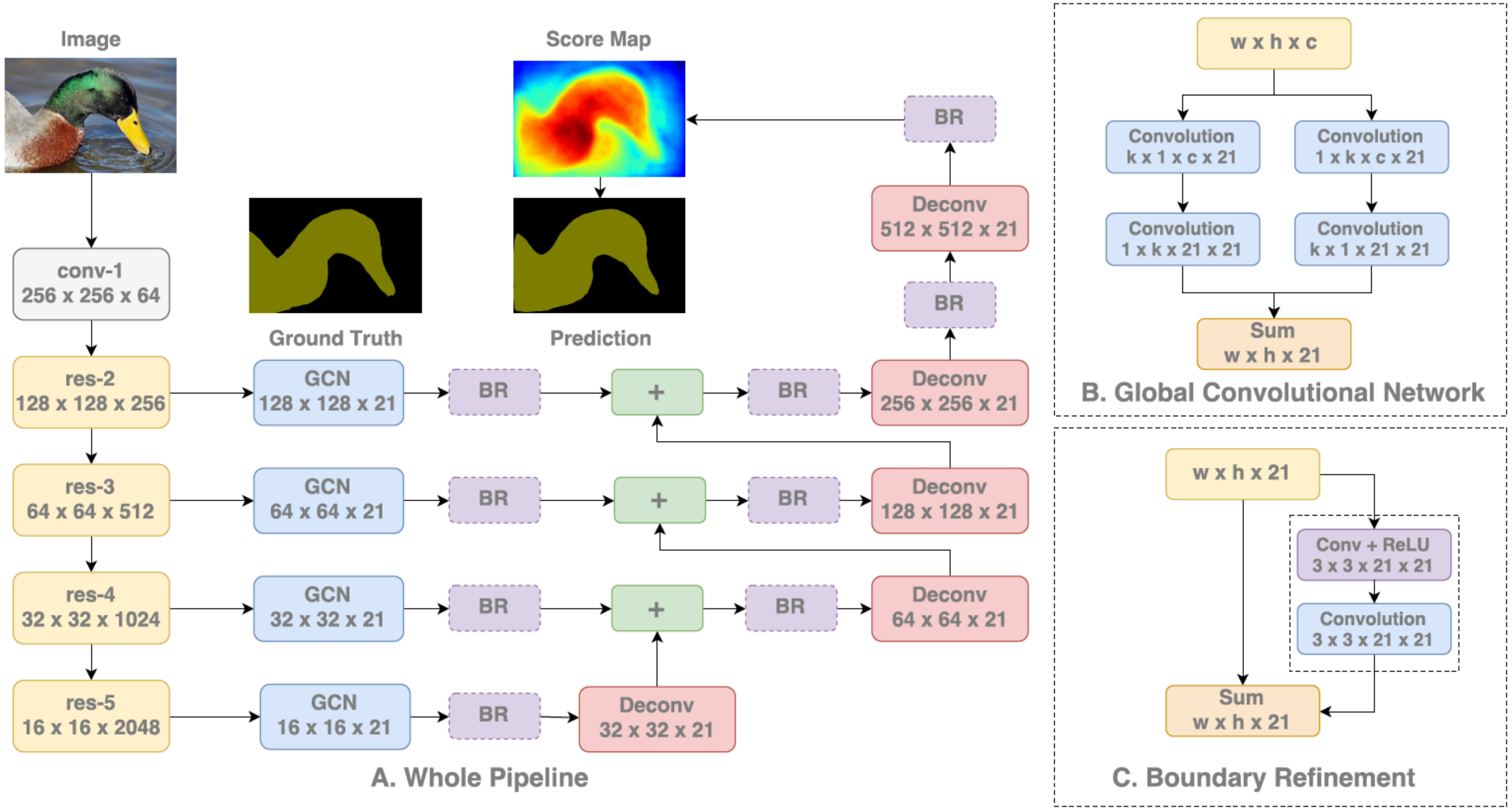


Figure 2. An overview of the whole pipeline in (A). The details of Global Convolutional Network (GCN) and Boundary Refinement (BR) block are illustrated in (B) and (C), respectively.

VRF

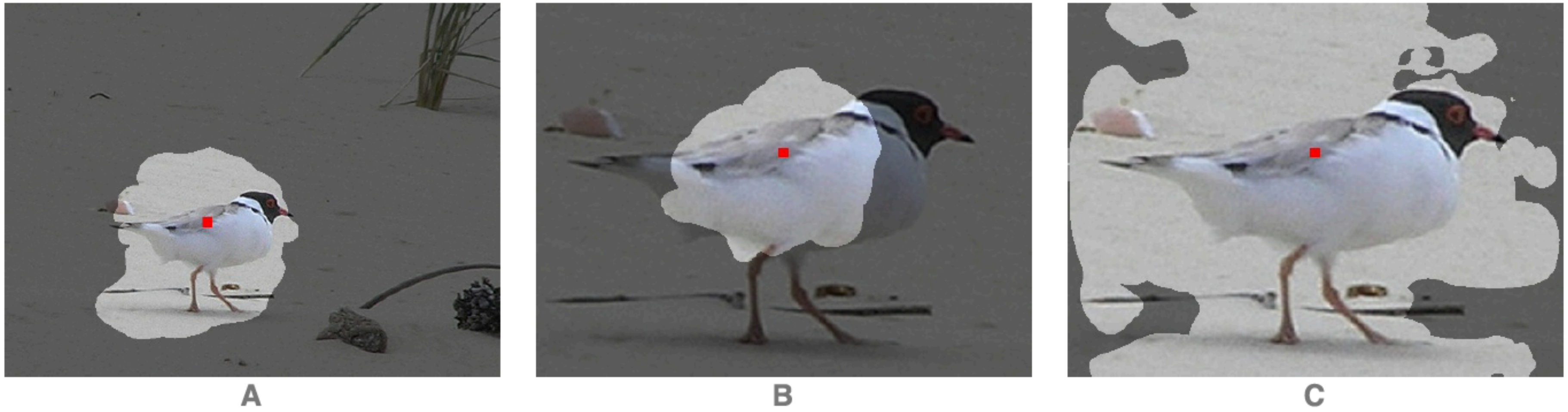
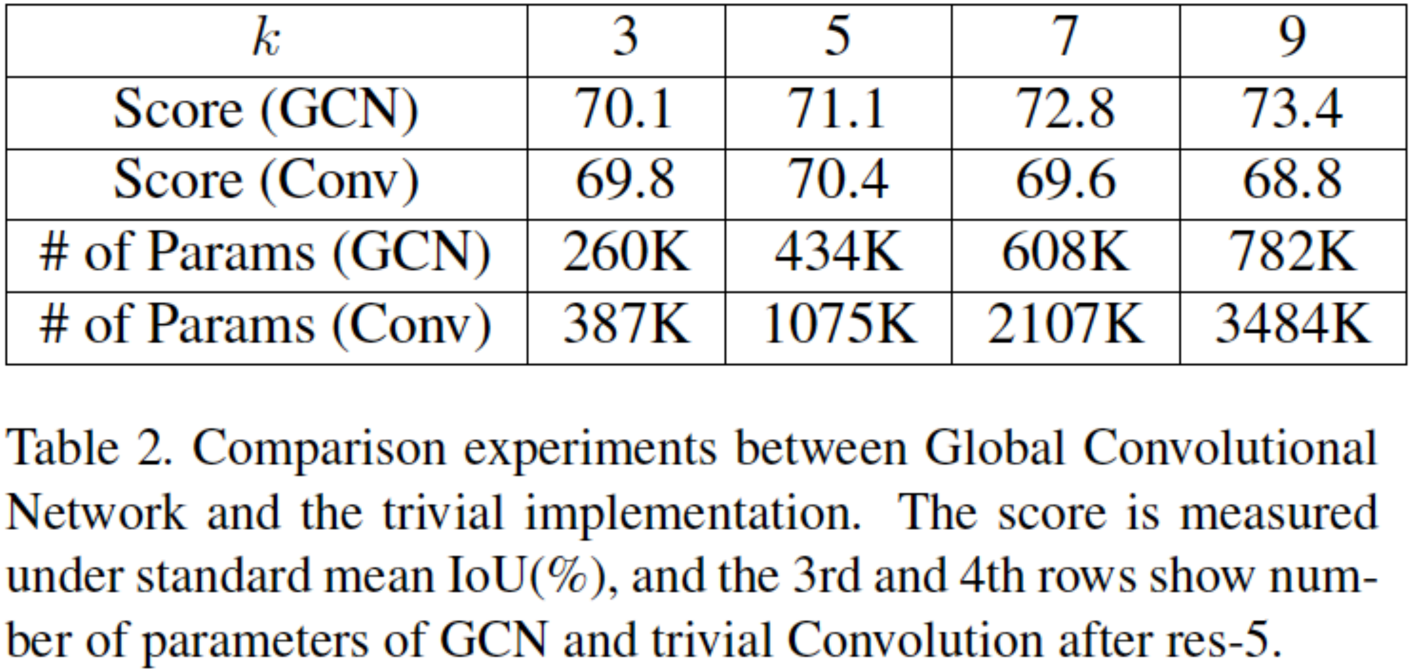


Figure 3. Visualization of *valid receptive field* (VRF) introduced by [38]. Regions on images show the VRF for the score map located at the center of the bird. For traditional segmentation model, even though the receptive field is as large as the input image, however, the VRF just covers the bird (A) and fails to hold the entire object if the input resized to a larger scale (B). As a comparison, our Global Convolution Network significantly enlarges the VRF (C).

Ablation Study

- (1) Are more parameters helpful? (性价比)

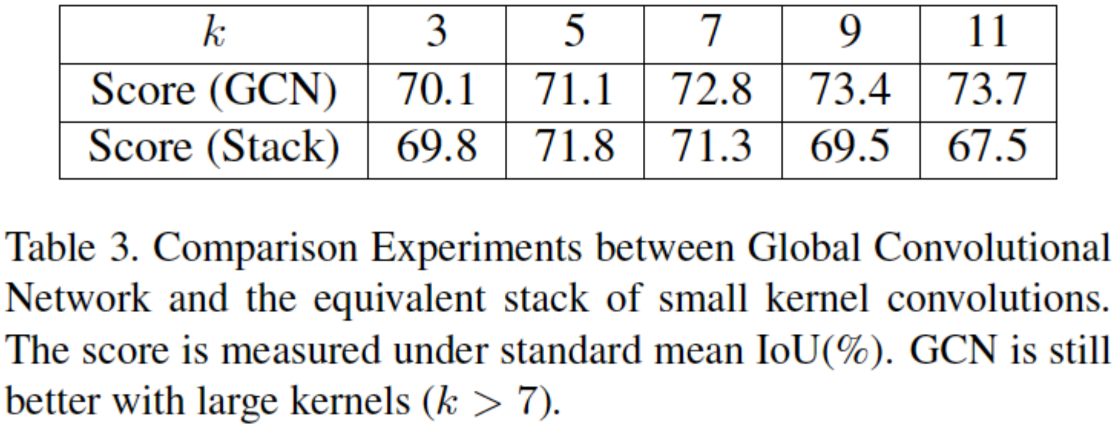


k	3	5	7	9
Score (GCN)	70.1	71.1	72.8	73.4
Score (Conv)	69.8	70.4	69.6	68.8
# of Params (GCN)	260K	434K	608K	782K
# of Params (Conv)	387K	1075K	2107K	3484K

Table 2. Comparison experiments between Global Convolutional Network and the trivial implementation. The score is measured under standard mean IoU(%), and the 3rd and 4th rows show number of parameters of GCN and trivial Convolution after res-5.

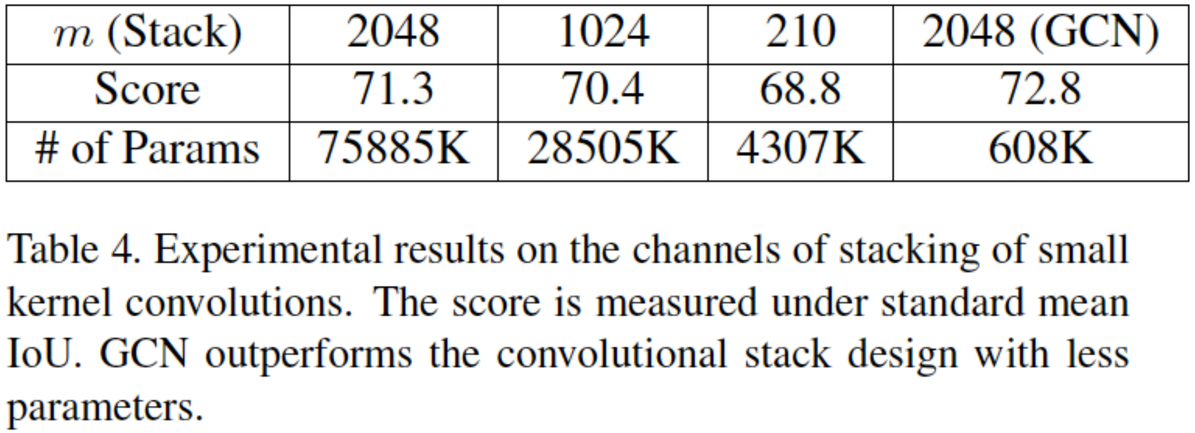
Ablation Study

- (2) GCN vs. Stack of small convolutions.



k	3	5	7	9	11
Score (GCN)	70.1	71.1	72.8	73.4	73.7
Score (Stack)	69.8	71.8	71.3	69.5	67.5

Table 3. Comparison Experiments between Global Convolutional Network and the equivalent stack of small kernel convolutions. The score is measured under standard mean IoU(%). GCN is still better with large kernels ($k > 7$).

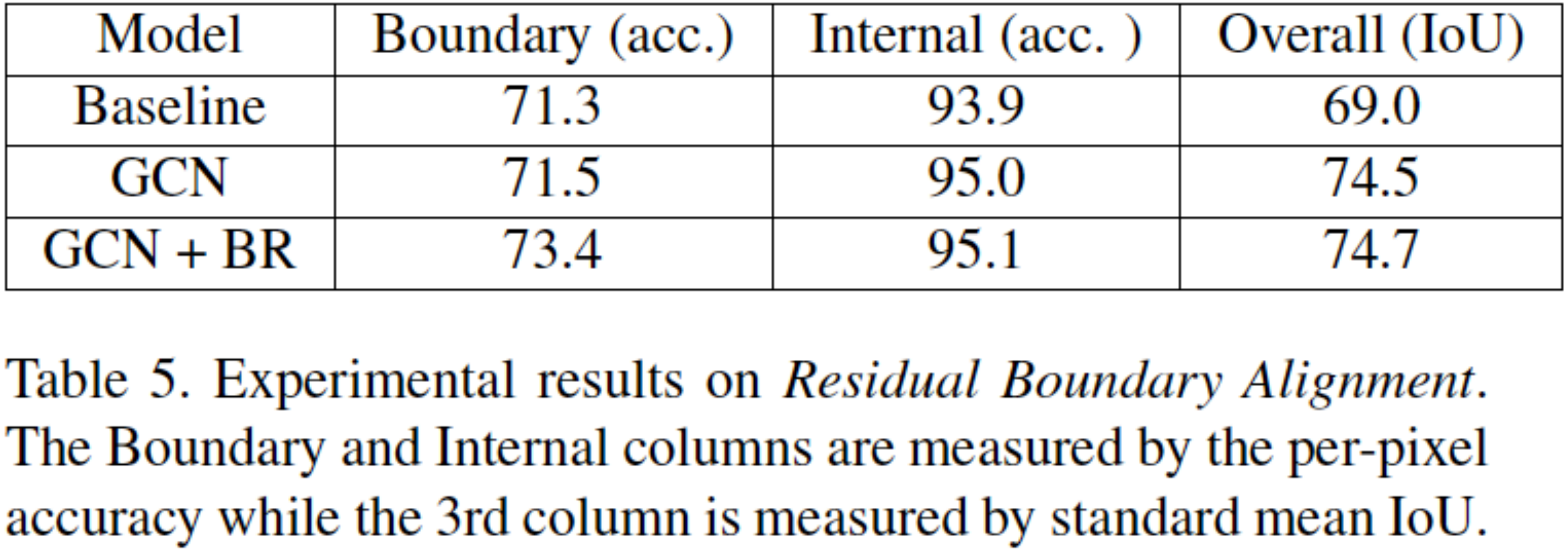


m (Stack)	2048	1024	210	2048 (GCN)
Score	71.3	70.4	68.8	72.8
# of Params	75885K	28505K	4307K	608K

Table 4. Experimental results on the channels of stacking of small kernel convolutions. The score is measured under standard mean IoU. GCN outperforms the convolutional stack design with less parameters.

Ablation Study

- (3) How GCN contributes to the segmentation results?



Model	Boundary (acc.)	Internal (acc.)	Overall (IoU)
Baseline	71.3	93.9	69.0
GCN	71.5	95.0	74.5
GCN + BR	73.4	95.1	74.7

Table 5. Experimental results on *Residual Boundary Alignment*. The Boundary and Internal columns are measured by the per-pixel accuracy while the 3rd column is measured by standard mean IoU.

Ablation Study

- (4) Global Convolutional Network for Pretrained Model

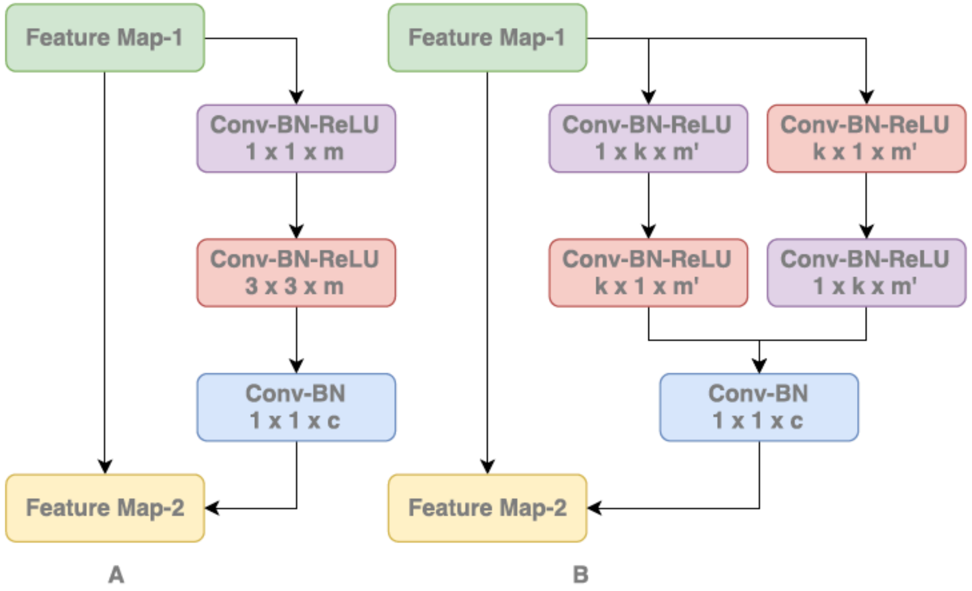
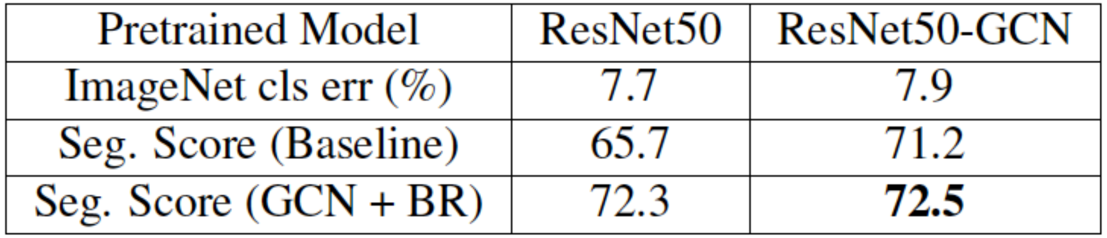


Figure 5. A: the bottleneck module in original ResNet. B: our *Global Convolutional Network* in ResNet-GCN.

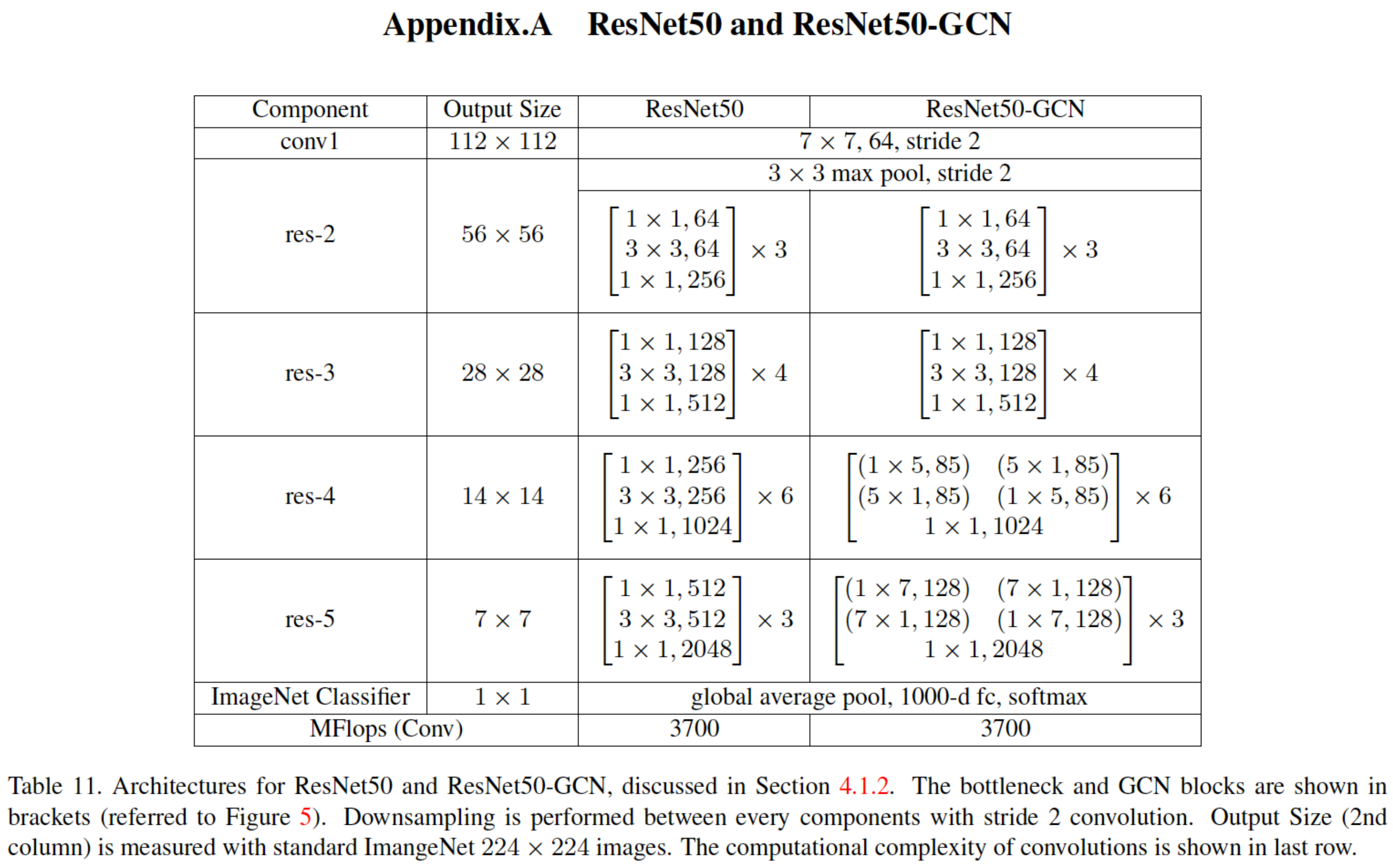


Pretrained Model	ResNet50	ResNet50-GCN
ImageNet cls err (%)	7.7	7.9
Seg. Score (Baseline)	65.7	71.2
Seg. Score (GCN + BR)	72.3	**72.5**

Table 6. Experimental results on ResNet50 and ResNet50-GCN. Top-5 error of 224×224 center-crop on 256×256 image is used in ImageNet classification error. The segmentation score is measured under standard mean IoU.

Ablation Study

- (4) Global Convolutional Network for Pretrained Model



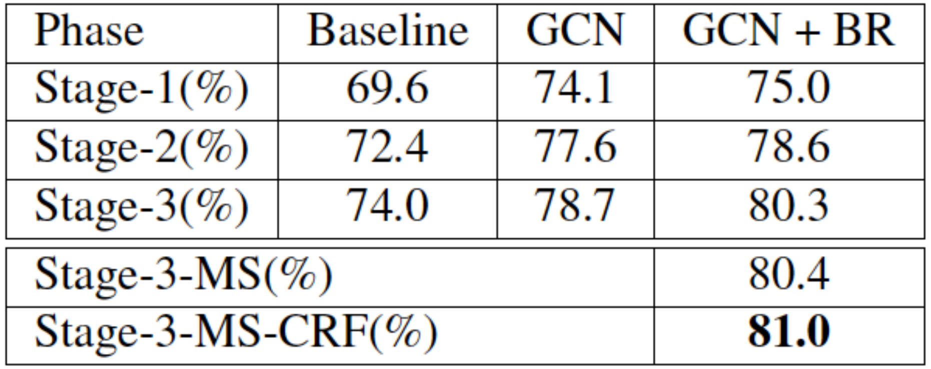
Appendix.A ResNet50 and ResNet50-GCN

Component	Output Size	ResNet50	ResNet50-GCN
conv1	112 × 112	7 × 7, 64, stride 2	
res-2	56 × 56	3 × 3 max pool, stride 2	
		$\left[\begin{array}{c}1\times1,64\\3\times3,64\\1\times1,256\end{array}\right]\times3$	$\left[\begin{array}{c}1\times1,64\\3\times3,64\\1\times1,256\end{array}\right]\times3$
res-3	28 × 28	$\left[\begin{array}{c}1\times1,128\\3\times3,128\\1\times1,512\end{array}\right]\times4$	$\left[\begin{array}{c}1\times1,128\\3\times3,128\\1\times1,512\end{array}\right]\times4$
res-4	14 × 14	$\left[\begin{array}{c}1\times1,256\\3\times3,256\\1\times1,1024\end{array}\right]\times6$	$\left[\begin{array}{cc}(1\times5,85)&(5\times1,85)\\(5\times1,85)&(1\times5,85)\\\multicolumn{2}{c}{1\times1,1024}\end{array}\right]\times6$
res-5	7 × 7	$\left[\begin{array}{c}1\times1,512\\3\times3,512\\1\times1,2048\end{array}\right]\times3$	$\left[\begin{array}{cc}(1\times7,128)&(7\times1,128)\\(7\times1,128)&(1\times7,128)\\\multicolumn{2}{c}{1\times1,2048}\end{array}\right]\times3$
ImageNet Classifier	1 × 1	global average pool, 1000-d fc, softmax	
MFlops (Conv)		3700	3700

Table 11. Architectures for ResNet50 and ResNet50-GCN, discussed in Section 4.1.2. The bottleneck and GCN blocks are shown in brackets (referred to Figure 5). Downsampling is performed between every components with stride 2 convolution. Output Size (2nd column) is measured with standard ImangeNet 224 × 224 images. The computational complexity of convolutions is shown in last row.

Experiment on Pascal VOC

- (1) In Stage-1 , we mix up all the images from COCO, SBD and standard PASCAL VOC 2012, resulting in 109,892 images for training.
- (2) During the Stage-2 , we use the SBD and standard PASCAL VOC 2012 images, the same as Section 4.1 .
- (3) For Stage-3 , we only use the standard PASCAL VOC 2012 dataset. The input image is padded to 640*640 in Stage-1 and 512*512 for Stage-2 and Stage-3.
- The evaluation on validation set is shown in Table 7 .



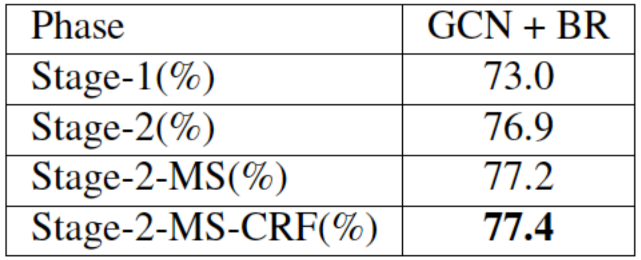
Phase	Baseline	GCN	GCN + BR
Stage-1(%)	69.6	74.1	75.0
Stage-2(%)	72.4	77.6	78.6
Stage-3(%)	74.0	78.7	80.3
Stage-3-MS(%)			80.4
Stage-3-MS-CRF(%)			**81.0**

Table 7. Experimental results on PASCAL VOC 2012 validation set. The results are evaluated by standard mean IoU.

82.2% on test set

Experiment on Cityscapes

- (1) In Stage-1 , we mix up the coarse annotated images and the training set, resulting in 22,973 images.
- (2) For Stage-2 , we only finetune the network on training set. During the evaluation phase, we split the images into four 1024*1024 crops and fuse their score maps.

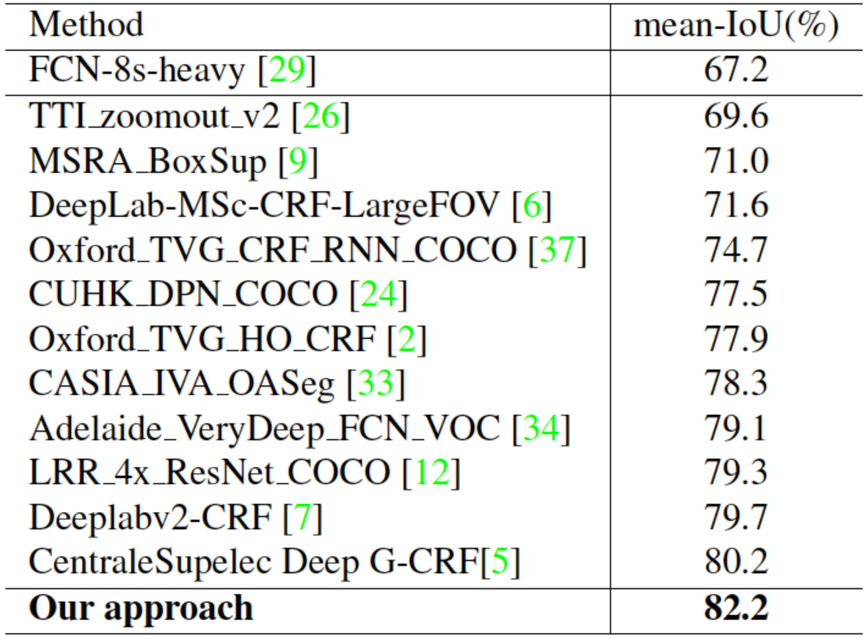


Phase	GCN + BR
Stage-1(%)	73.0
Stage-2(%)	76.9
Stage-2-MS(%)	77.2
Stage-2-MS-CRF(%)	**77.4**

Table 9. Experimental results on Cityscapes validation set. The standard mean IoU is used here.

76.9% on test set

Result



Method	mean-IoU(%)
FCN-8s-heavy [29]	67.2
TTI_zoomout_v2 [26]	69.6
MSRA_BoxSup [9]	71.0
DeepLab-MSc-CRF-LargeFOV [6]	71.6
Oxford_TVG_CRF_RNN_COCO [37]	74.7
CUHK_DPN_COCO [24]	77.5
Oxford_TVG_HO_CRF [2]	77.9
CASIA_IVA_OASeg [33]	78.3
Adelaide_VeryDeep_FCN_VOC [34]	79.1
LRR_4x_ResNet_COCO [12]	79.3
Deeplabv2-CRF [7]	79.7
CentraleSupelec Deep G-CRF[5]	80.2
Our approach	**82.2**

Table 8. Experimental results on PASCAL VOC 2012 test set.

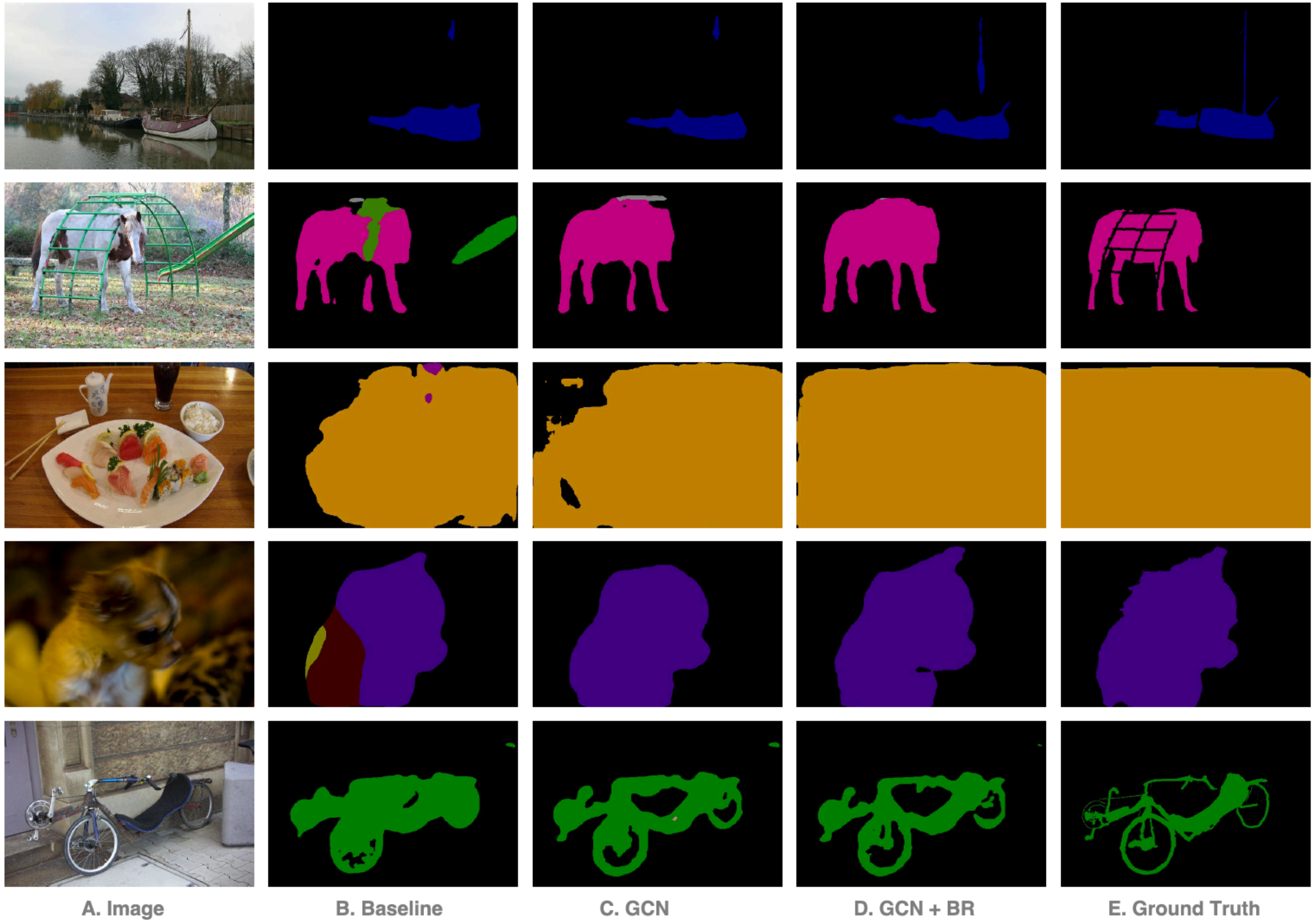


Figure 6. Examples of semantic segmentation results on PASCAL VOC 2012. For every row we list input image (A), 1 × 1 convolution baseline (B), Global Convolutional Network (GCN) (C), Global Convolutional Network plus Boundary Refinement (GCN + BR) (D), and Ground truth (E).

Thank you