

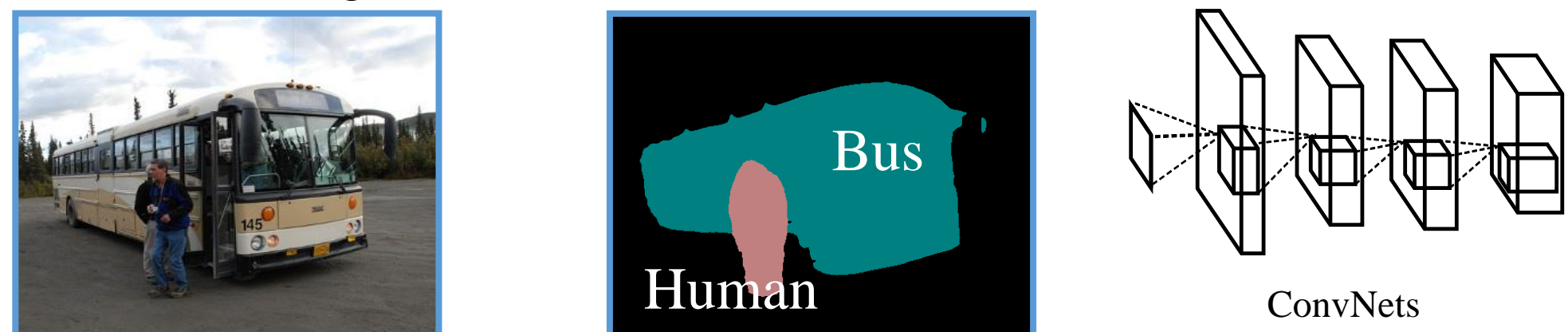
Not All Pixels Are Equal: Difficulty-Aware Semantic Segmentation via Deep Layer Cascade

Xiaoxiao Li, Ziwei Liu, Ping Luo, Chen Change Loy, Xiaoou Tang
Department of Information Engineering, The Chinese University of Hong Kong
{lx015,lz013,pluo,ccloy,xtang}@ie.cuhk.edu.hk



1. Introduction

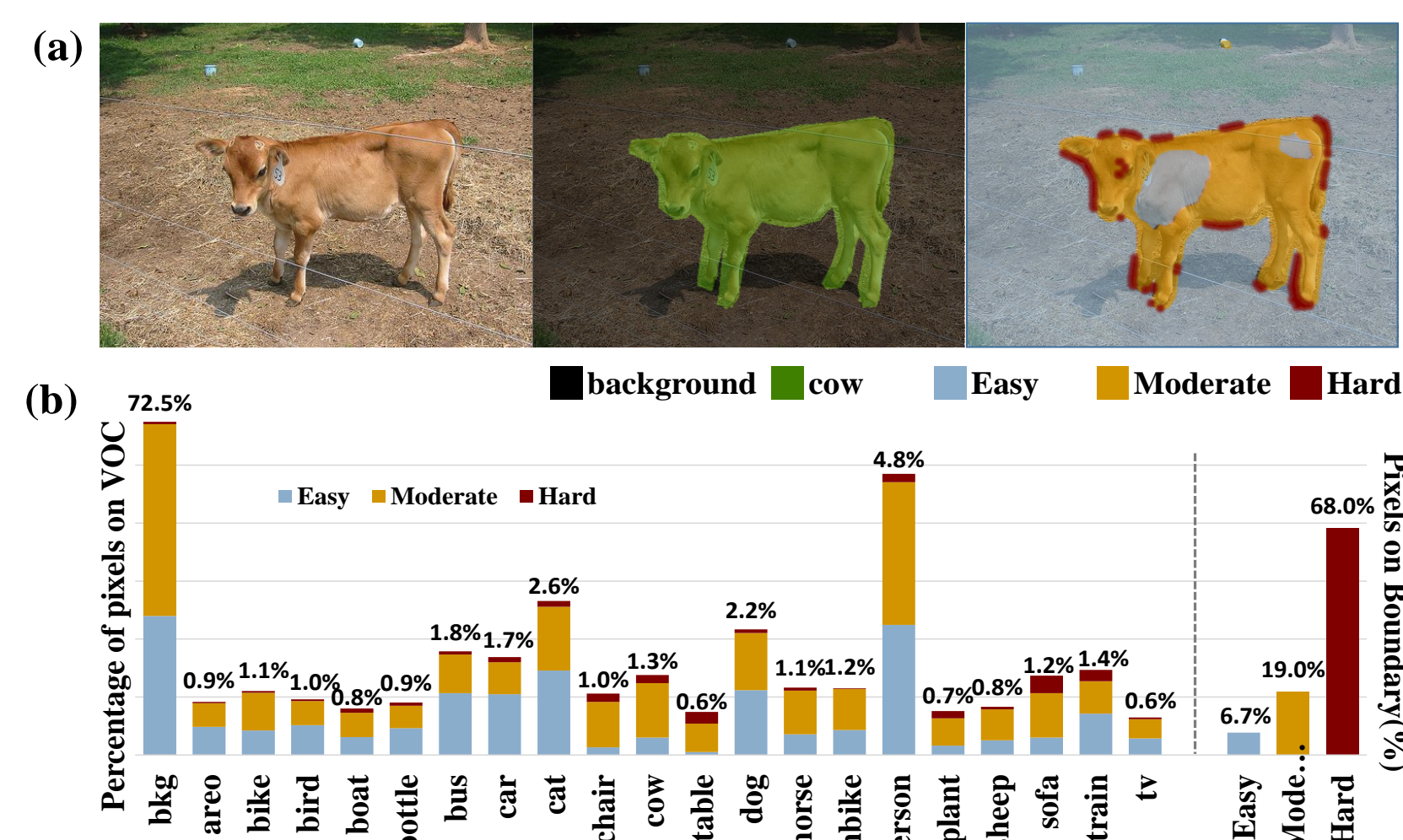
- Task & General Approaches:**
Semantic Segmentation



- Existing Works:**
Increase the model capacity to achieve promising results
High runtime complexity

| Backbone Network | Speed(300*500 image) |
|----------------------|----------------------|
| VGG16 | 5.7 fps |
| ResNet-101 | 7.1 fps |
| Inception-ResNet | 9.0 fps |
| Layer Cascade | 14.7 fps |
| Layer Cascade (fast) | 23.6 fps |

- Motivation:**
Partition all pixels into three sets by classification confidence

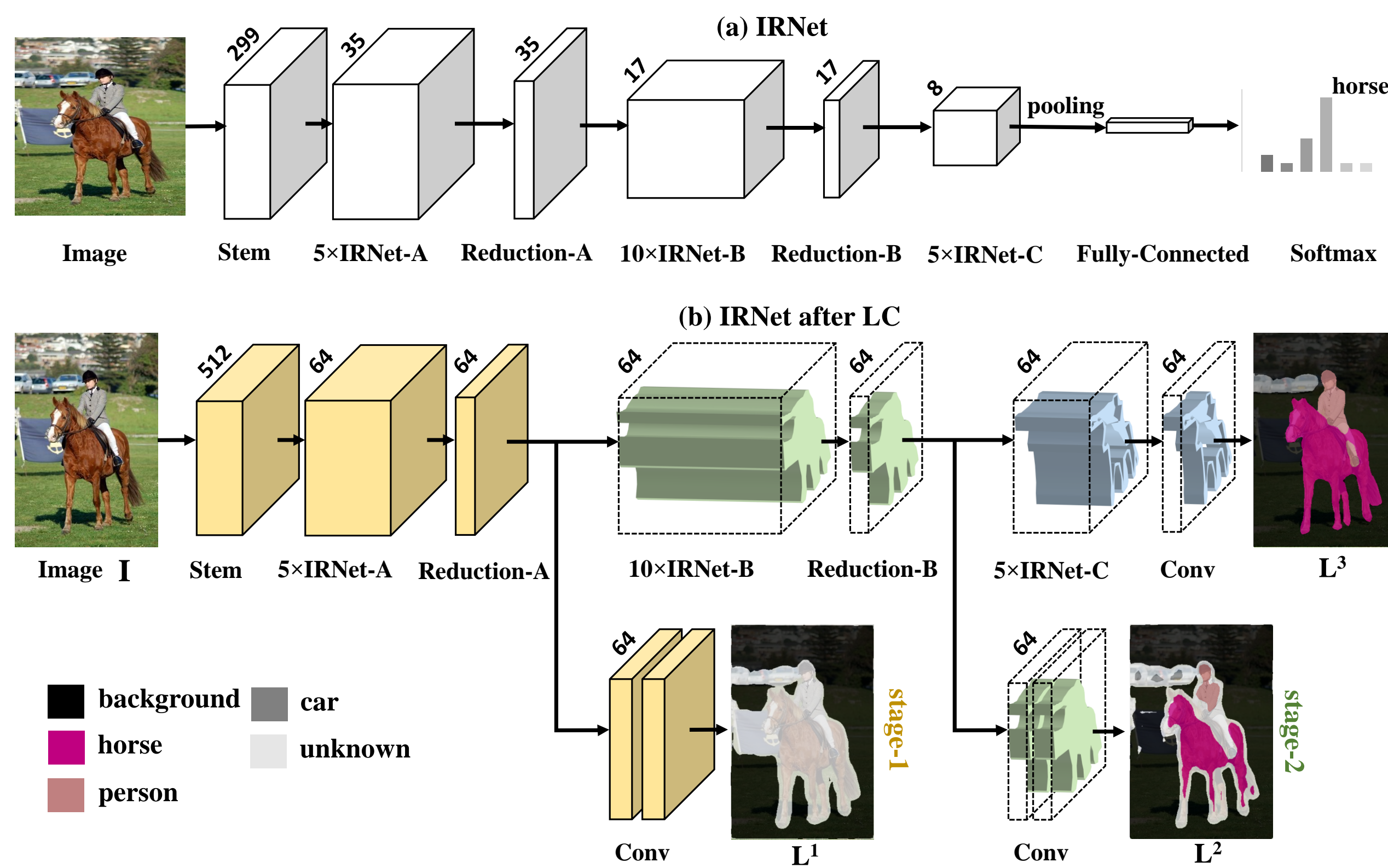


Nearly 40% easy region

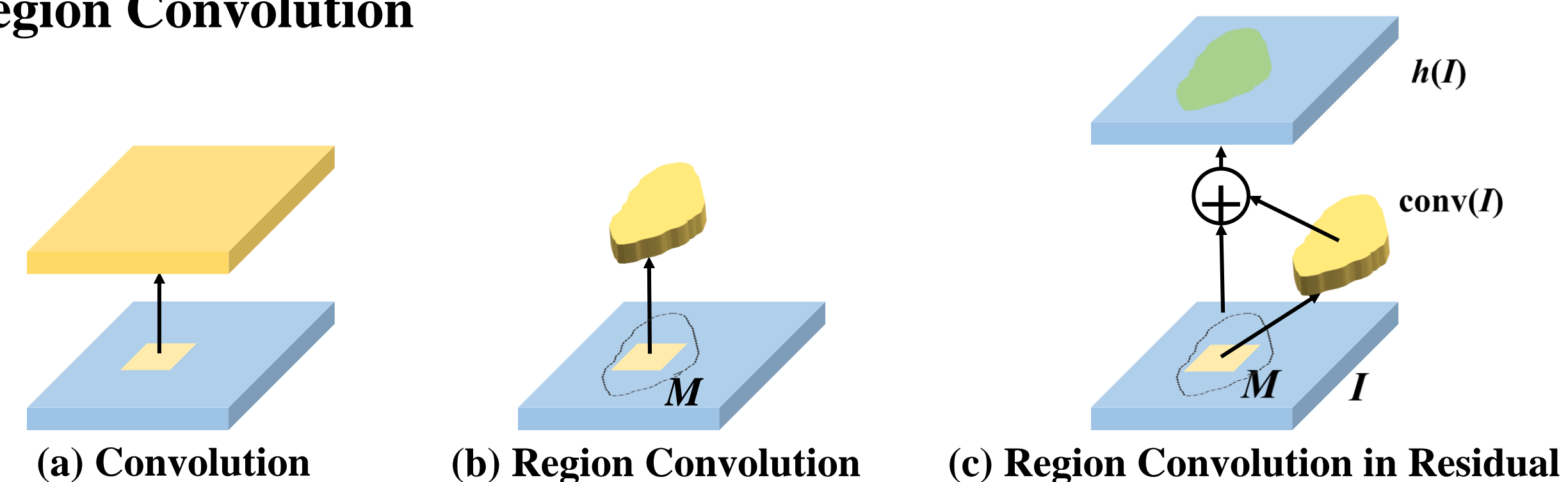
- Our Idea:**
Treats a **single** deep model as a **cascade of several sub-models**
Earlier sub-models are trained to handle **easy and confident regions**
Feed-forward **harder regions** to the **next sub-model** for processing

2. Approach

- Turning Inception-ResNet into Deep Layer Cascade (LC)**



- Region Convolution**



5. Overall Performance

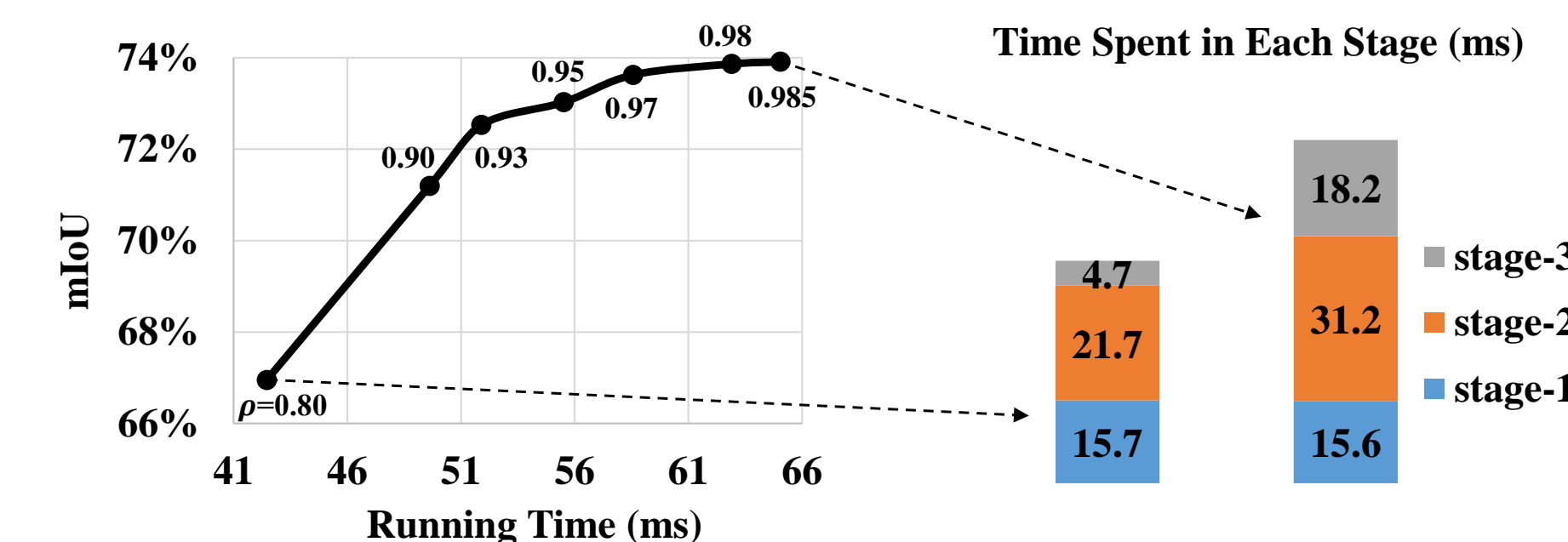
| | areo | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | person | plant | sheep | sofa | train | tv | mIoU |
|-------------------------|------|------|------|------|--------|------|------|------|-------|------|-------|------|-------|-------|--------|-------|-------|------|-------|------|-------------|
| FCN | 76.8 | 34.2 | 68.9 | 49.4 | 60.3 | 75.3 | 74.7 | 77.6 | 21.4 | 62.5 | 46.8 | 71.8 | 63.9 | 76.5 | 73.9 | 45.2 | 72.4 | 37.4 | 70.9 | 55.1 | 62.2 |
| DeepLab | 84.4 | 54.5 | 81.5 | 63.6 | 65.9 | 85.1 | 79.1 | 83.4 | 30.7 | 74.1 | 59.8 | 79.0 | 76.1 | 83.2 | 80.8 | 59.7 | 82.2 | 50.4 | 73.1 | 63.7 | 71.6 |
| RNN | 87.5 | 39.0 | 79.7 | 64.2 | 68.3 | 87.6 | 80.8 | 84.4 | 30.4 | 78.2 | 60.4 | 80.5 | 77.8 | 83.1 | 80.6 | 59.5 | 82.8 | 47.8 | 78.3 | 67.1 | 72.0 |
| Adelaide | 91.9 | 48.1 | 93.4 | 69.3 | 75.5 | 94.2 | 87.5 | 92.8 | 36.7 | 86.9 | 65.2 | 89.1 | 90.2 | 86.5 | 87.2 | 64.6 | 90.1 | 59.7 | 85.5 | 72.7 | 79.1 |
| RNN [†] | 90.4 | 55.3 | 88.7 | 68.4 | 69.8 | 88.3 | 82.4 | 85.1 | 32.6 | 78.5 | 64.4 | 79.6 | 81.9 | 86.4 | 81.8 | 58.6 | 82.4 | 53.5 | 77.4 | 70.1 | 74.7 |
| BoxSup [†] | 89.8 | 38.0 | 89.2 | 68.9 | 68.0 | 89.6 | 83.0 | 87.7 | 34.4 | 83.6 | 67.1 | 81.5 | 83.7 | 85.2 | 83.5 | 58.6 | 84.9 | 55.8 | 81.2 | 70.7 | 75.2 |
| DPN [†] | 89.0 | 61.6 | 87.7 | 66.8 | 74.7 | 91.2 | 84.3 | 87.6 | 36.5 | 86.3 | 66.1 | 84.4 | 87.8 | 85.6 | 85.4 | 63.6 | 87.3 | 61.3 | 79.4 | 66.4 | 77.5 |
| DeepLab-v2 [†] | 92.6 | 60.4 | 91.6 | 63.4 | 76.3 | 95.0 | 88.4 | 92.6 | 32.7 | 88.5 | 67.6 | 89.6 | 92.1 | 87.0 | 87.4 | 63.3 | 88.3 | 60.0 | 86.8 | 74.5 | 79.7 |
| LC | 94.1 | 63.0 | 91.2 | 67.9 | 79.5 | 93.4 | 90.0 | 93.8 | 37.4 | 83.7 | 65.9 | 90.7 | 86.1 | 88.8 | 87.5 | 68.5 | 86.9 | 64.3 | 85.6 | 72.2 | 80.3 |
| LC [†] | 85.5 | 66.7 | 94.5 | 67.2 | 84.0 | 96.1 | 89.8 | 93.5 | 47.2 | 90.4 | 71.5 | 88.9 | 91.7 | 89.2 | 89.1 | 70.4 | 89.4 | 70.7 | 84.2 | 79.6 | 82.7 |

Per-class results on VOC12 test set. Approaches pre-trained on COCO are marked with [†].

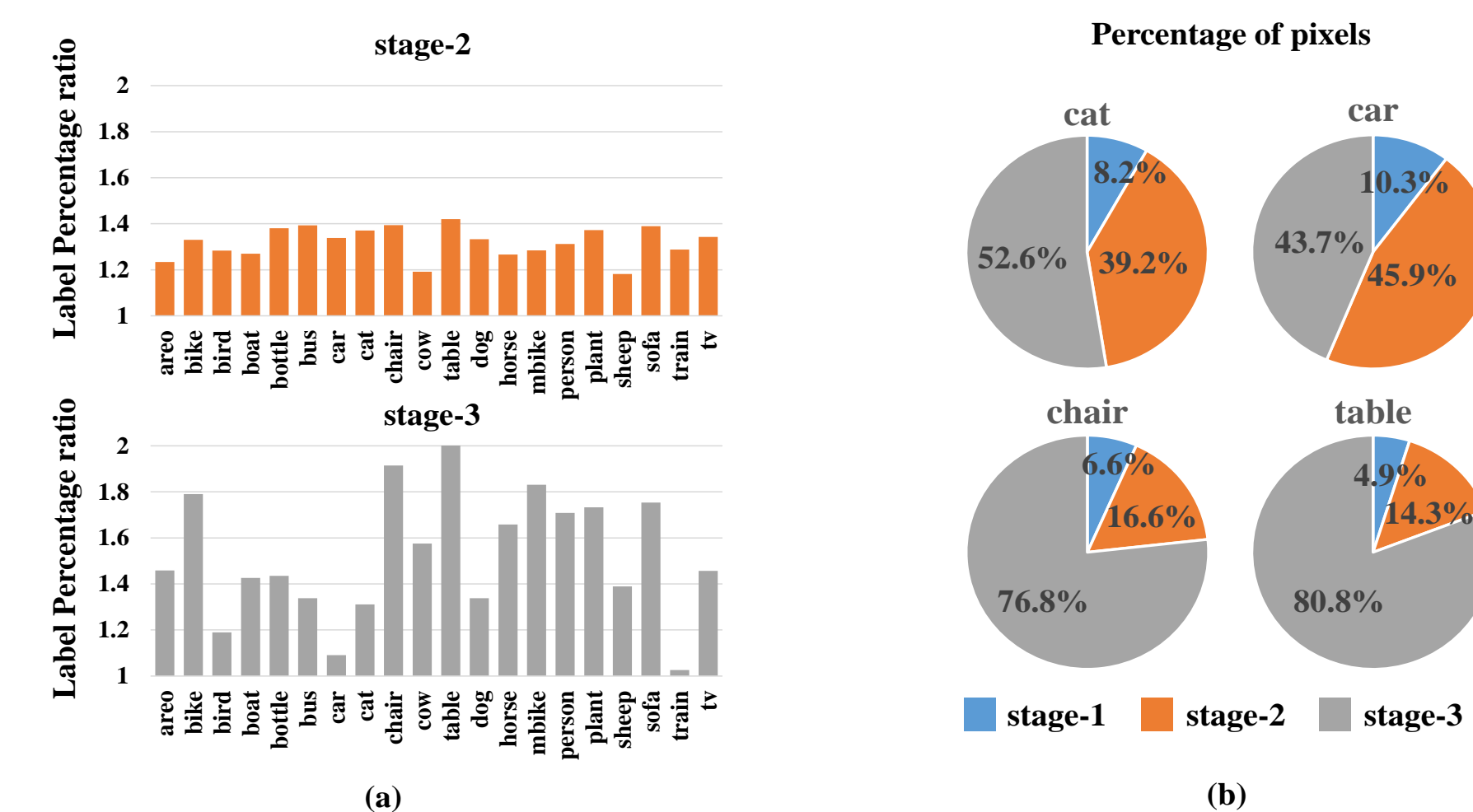
3. Experiments

- Ablation Study on Probability Thresholds ρ**
 ρ controls percentage of easy and hard regions
- Performance and Speed Trade-off**

| ρ | 1 | 0.995 | 0.985 | 0.970 | 0.950 | 0.930 | 0.900 | 0.800 |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|
| stage-1 (%) | 0 | 15 | 23 | 30 | 35 | 35 | 44 | 56 |
| stage-2 (%) | 0 | 14 | 29 | 31 | 30 | 41 | 31 | 29 |
| mIoU (%) | 72.70 | 73.56 | 73.91 | 73.63 | 73.03 | 72.53 | 71.20 | 66.95 |



- Stage-wise Label Distribution**

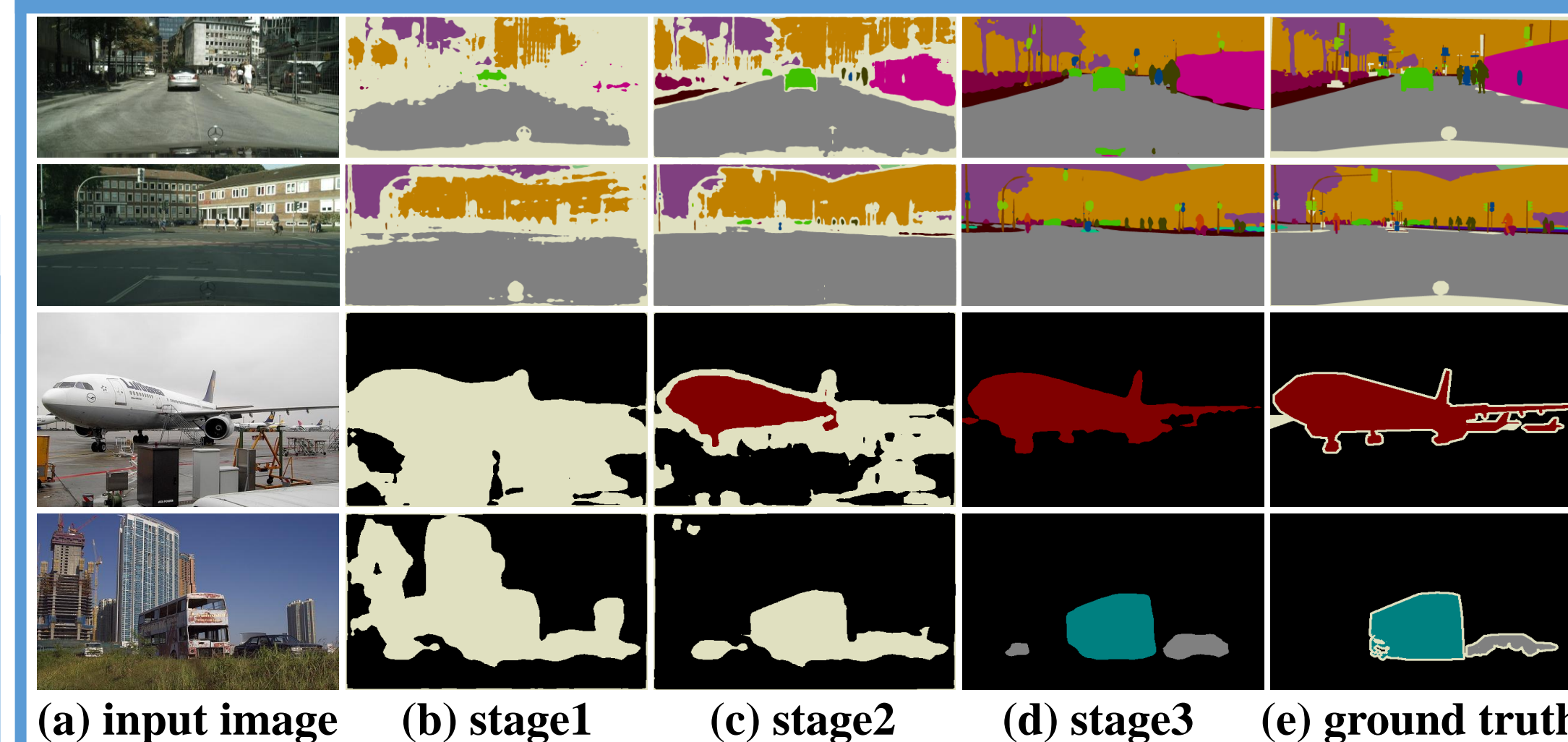


Stage-1 handles the easy regions (background)

Stage-2 focus more on the foreground than stage-1 does

Stage-3 further focus on harder classes

4. Stages' Outputs



6. Conclusion

- LC adopts a “**difficulty-aware**” learning paradigm.
- LC accelerates both training and testing by the usage of **region convolution**. It is capable of running in **real-time**.
- LC is an **end-to-end trainable** framework that jointly optimizes the feature learning for different regions.

