# Appendix

This appendix presents additional results. We study the impact of ImageNet pretraining on the performance and demonstrate its importance in Section A. To gain more insight about our approach Segmenter, we analyze its attention maps and the learned class embeddings in Section B. Finally, we give an additional qualitative comparison of Segmenter to DeepLabv3+ on ADE20K, Cityscapes and Pascal Context in Section C.

## A. ImageNet pre-training

To study the impact of ImageNet pre-training on Segmenter, we compare our model pre-trained on ImageNet with equivalent models trained from scratch. To train from scratch, the weights of the model are initialized randomly with a truncated normal distribution. We use a base learning rate of $10^{-3}$ and two training procedures. First, we follow the fine-tuning procedure and use SGD optimizer with "poly" scheduler. Second, we follow a more standard procedure when training a transformer from scratch where we use AdamW with a cosine scheduler and a linear warmup for $16K$ iterations corresponding to 10% of the total number of iterations. Table 9 reports results for Seg-S/16. We observe that when pre-trained on ImageNet-21k using SGD, Seg-S/16 reaches 45.37% yielding a 32.9% improvement over the best randomly initialized model.

| Method | Pre-training | Optimizer | mIoU (SS) |
|---|---|---|---|
| Seg-S/16 | None | AdamW | 4.42 |
| Seg-S/16 | None | SGD | 12.51 |
| Seg-S/16 | ImageNet-21k | AdamW | 34.77 |
| Seg-S/16 | ImageNet-21k | SGD | **45.37** |

Table 9: Impact of pretraining on the performance on ADE20K validation set.

## B. Attention maps and class embeddings

To better understand how our approach Segmenter processes images, we display attention maps of Seg-B/8 for 3 images in Figure 6. We resize attention maps to the original image size. For each image, we analyze attention maps of a patch on a small instance, for example lamp, cow or car. We also analyze attention maps of a patch on a large instance, for example bed, grass and road. We observe that the attention map field-of-view adapts to the input image and the instance size, gathering global information on large instances and focusing on local information on smaller instances. This adaptability is typically not possible with CNN which have a constant field-of-view, independently of the data. We also note there is progressive gathering of information from bottom to top layers, as for example on the cow instance, where the model first identifies the cow the patch belongs to, then identifies other cow instances. We observe that attention maps of lower layers depends strongly on the selected patch while they tend to be more similar for higher layers.

Additionally, to illustrate the larger receptive field size of Segmenter compared to CNNs, we reported the size of the attended area in Figure 7, where each dot shows the mean attention distance for one of the 12 attention heads at each layer. Already for the first layer, some heads attend to distant patches which clearly lie outside the receptive field of ResNet/ResNeSt initial layers.

To gain some understanding of the class embeddings learned with the mask transformer, we project embeddings into 2D with a singular value decomposition. Figure 8 shows that these projections group instances such as means of transportation (bottom left), objects in a house (top) and outdoor categories (middle right). It displays an implicit clustering of semantically related categories.

## C. Qualitative results

We present additional qualitative results including comparison with DeepLabv3+ ResNeSt-101 and failure cases in Figures 9, 10 and 11. We can see in Figure 9 that Segmenter produces more coherent segmentation maps than DeepLabv3+. This is the case for the wedding dress in (a) or the airplane signalmen's helmet in (b). In Figure 10, we show how for some examples, different segments which look very similar are confused both in DeepLabv3+ and Segmenter. For example, the armchairs and couchs in (a), the cushions and pillows in (b) or the trees, flowers and plants in (c) and (d). In Figure 11, we can see how DeepLabv3+ handles better the boundaries between different people entities. Finally, both Segmenter and DeepLabv3+ have problems segmenting small instances such as lamp, people or flowers in Figure 12 (a) or the cars and signals in Figure 12 (b).

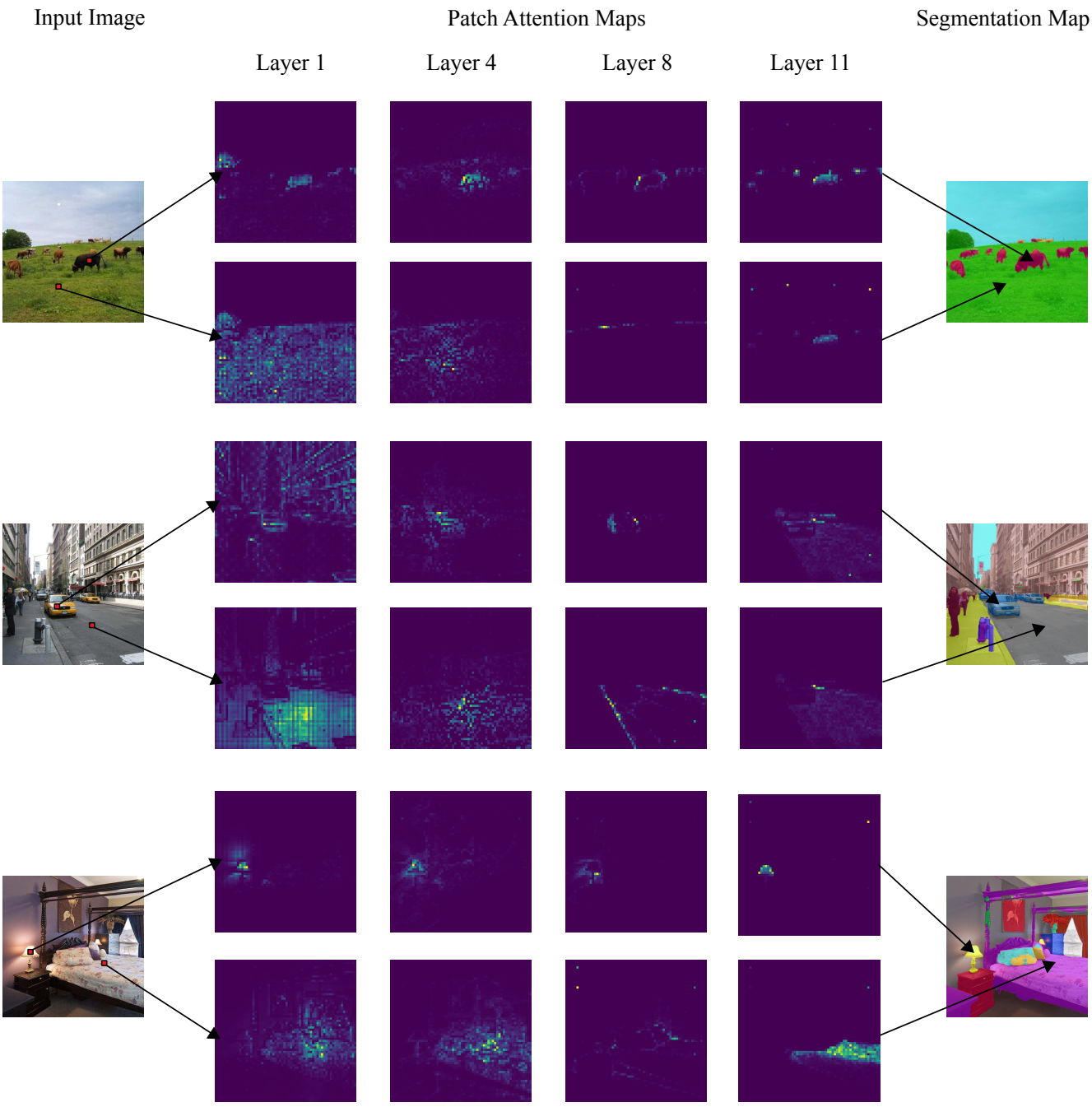


Figure 6: Seg-B/8 patch attention maps for the layers 1, 4, 8 and 11.

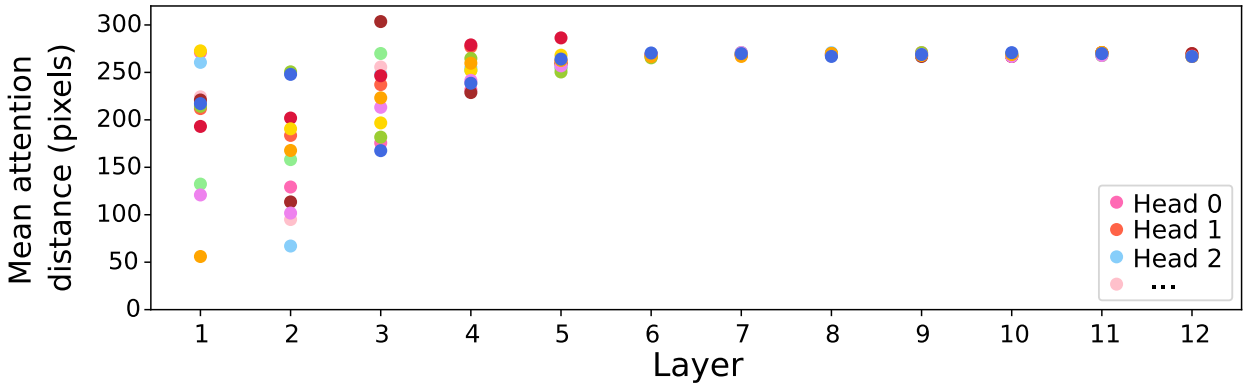


Figure 7: Size of attended area by head and model depth.

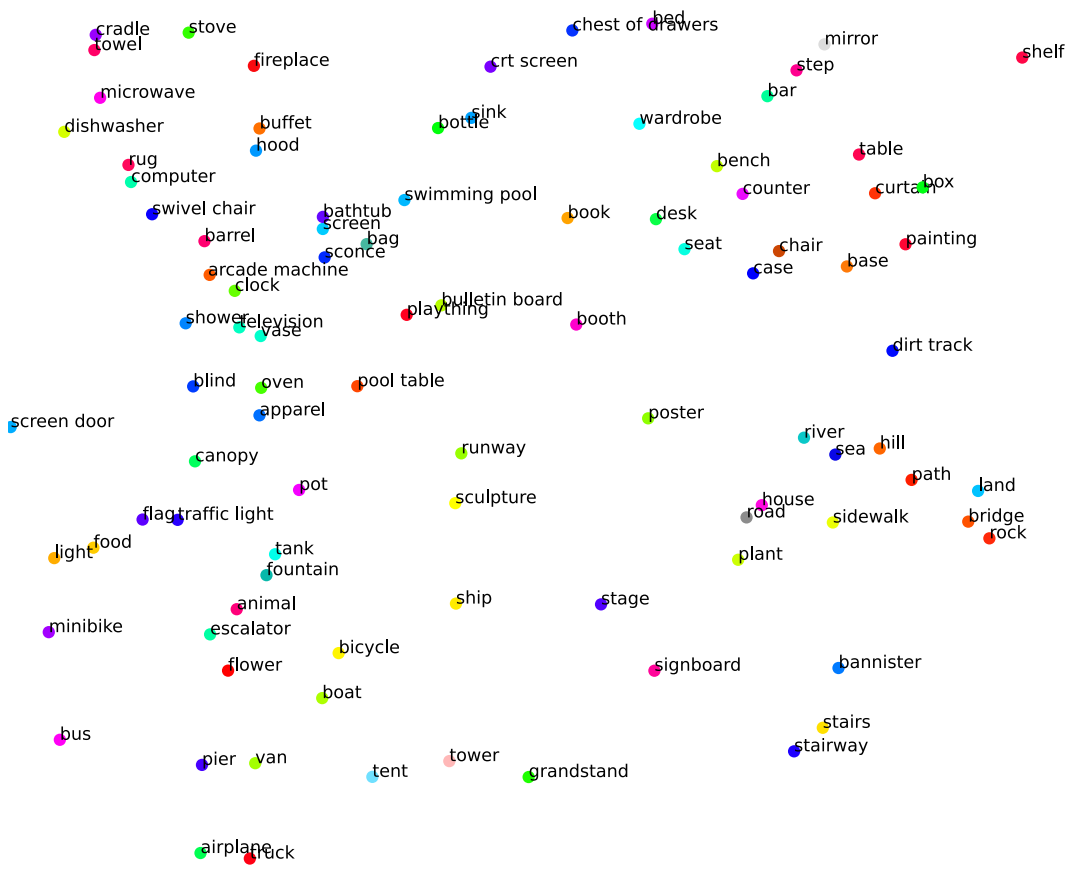


Figure 8: Singular value decompostion of the class embeddings learned with the mask transformer on ADE20K.

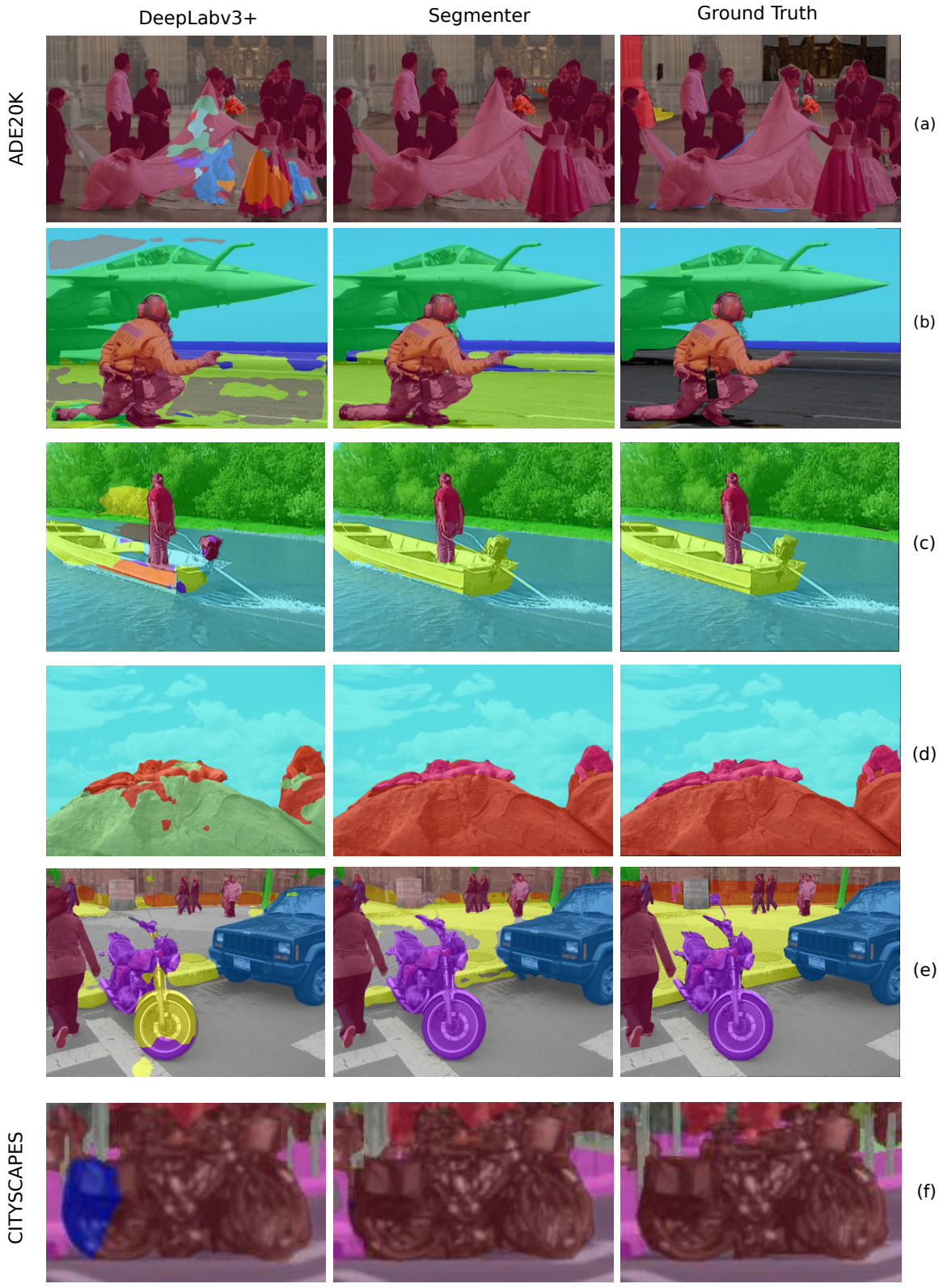


Figure 9: Segmentation maps where Seg-L-Mask/16 produces more coherent segmentation maps than DeepLabv3+ ResNeSt-101.

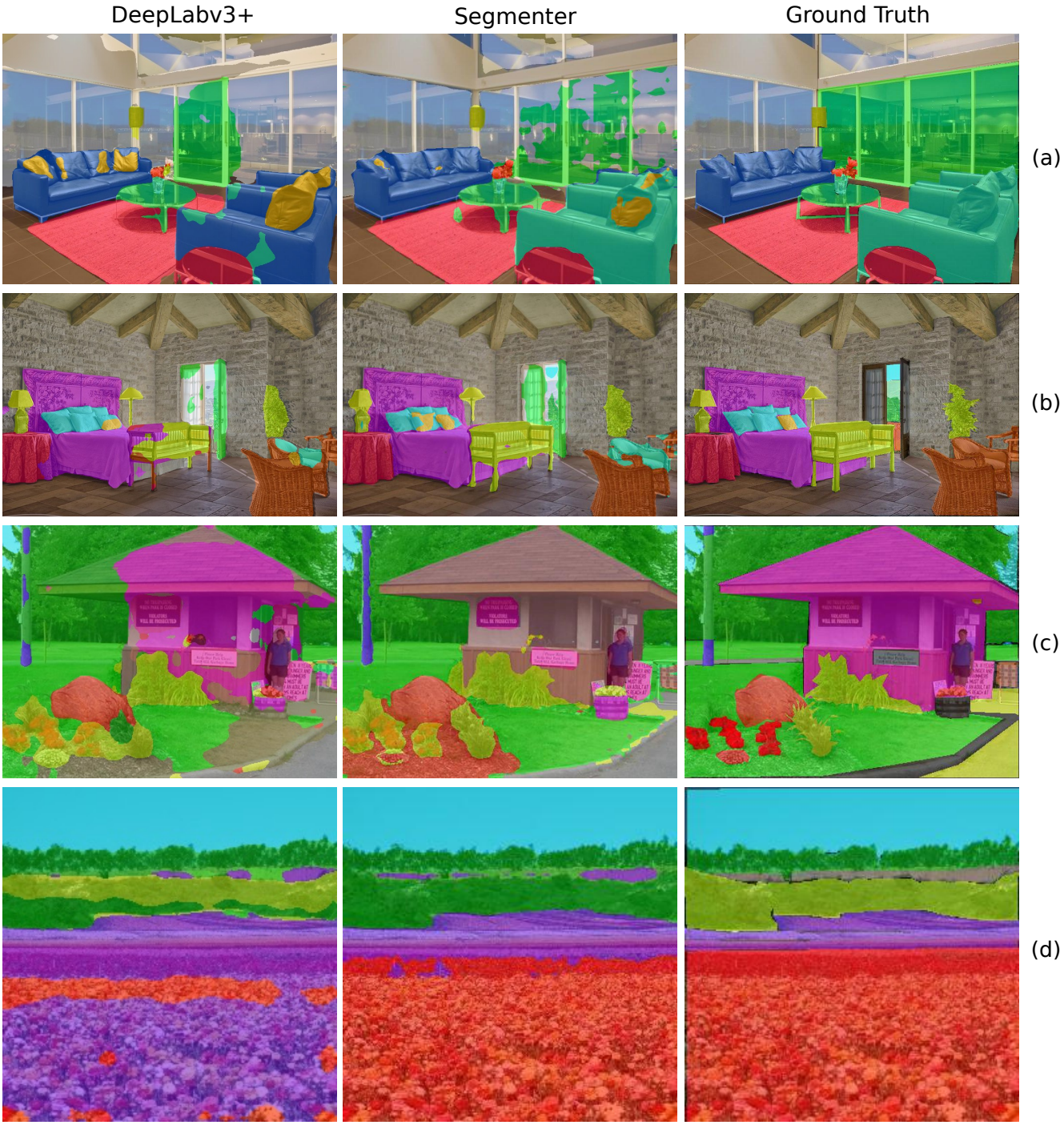


Figure 10: Examples for Seg-L-Mask/16 and DeepLabv3+ ResNeSt-101 on ADE20K, where elements which look very similar are confused.

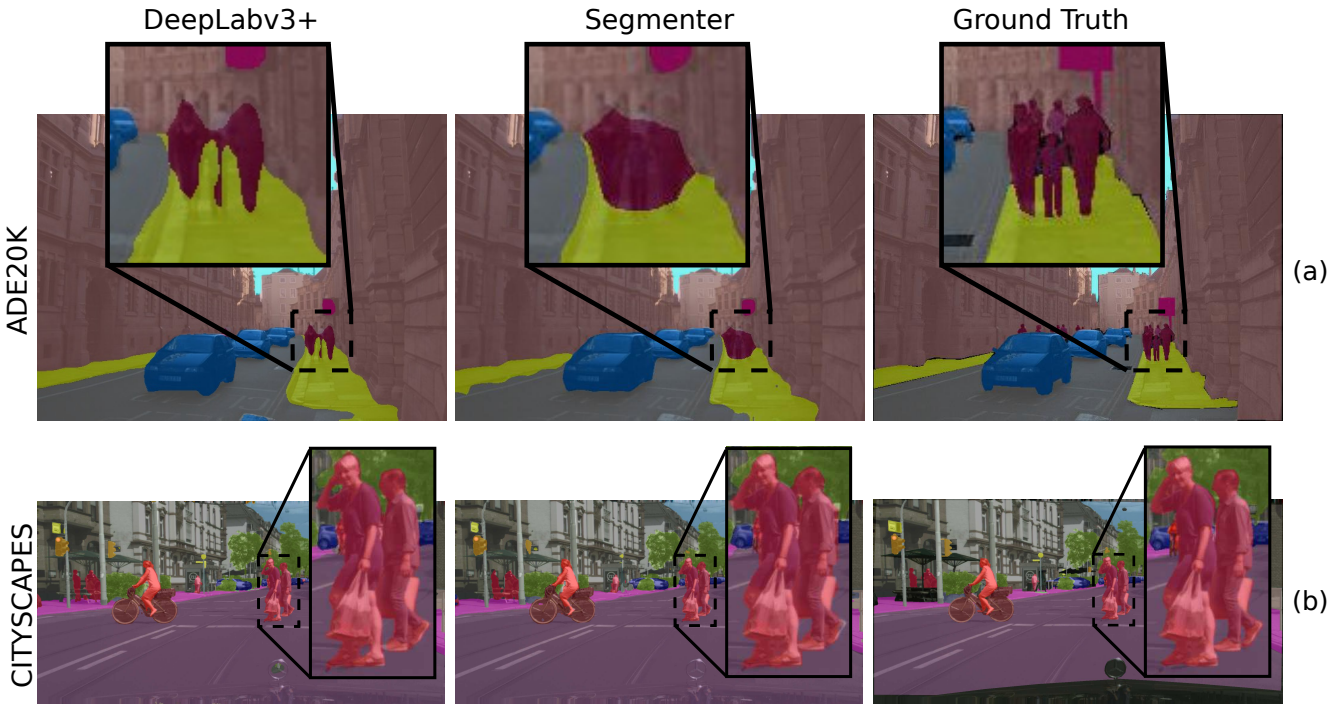


Figure 11: Comparison of Seg-L-Mask/16 with DeepLabV3+ ResNeSt-101 for images with near-by persons. We can observe that DeepLabV3+ localizes boundaries better.

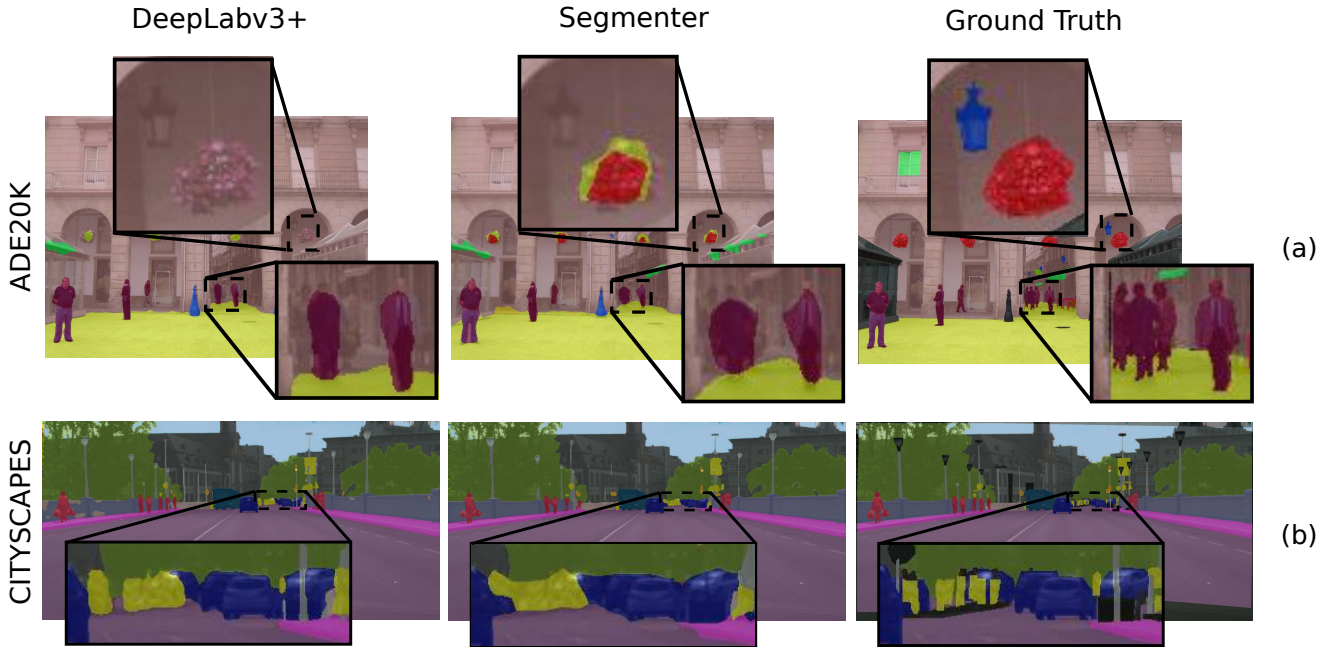


Figure 12: Failure cases of DeepLabV3+ ResNeSt-101 and Seg-L-Mask/16, for small instances such as (a) lamp, people, flowers and (b) cars, signals.