Supplementary Material: The Role of Context for Object Detection and Semantic Segmentation in the Wild

Roozbeh Mottaghi[1] Xianjie Chen[2] Xiaobai Liu[2] Nam-Gyu Cho[3] Seong-Whan Lee[3]
Sanja Fidler[4] Raquel Urtasun[4] Alan Yuille[2]
Stanford University[1] UCLA[2] Korea University[3] University of Toronto[4]

In this paper [6], we are interested in analyzing the effect of context in detection and segmentation approaches. Towards this goal, we label every pixel of the training and validation sets of the PASCAL VOC 2010 detection challenge with a semantic class. We selected PASCAL as our testbed as it has served as *the* benchmark for detection and segmentation in the community for years (over 600 citations and tens of teams competing in the challenges each year). Our analysis shows that our new dataset is much more challenging than existing ones (e.g., Barcelona [7], SUN [8], SIFT flow [5]), as it has higher class entropy, less pixels are labeled as "stuff" and instead belong to a wide variety of object categories beyond the 20 PASCAL object classes.

We analyze the ability of state-of-the-art methods [7, 1] to perform semantic segmentation of the most frequent classes, and show that approaches based on nearest neighbor (NN) retrieval are significantly outperformed by approaches based on bottom-up grouping, showing the variability of PASCAL images. We also study the performance of contextual models for object detection, and show that existing models have a hard time dealing with PASCAL imagery. In order to push forward the performance in this difficult scenario, we propose a novel deformable part-based model, which exploits both local context around each candidate detection as well as global context at the level of the scene. As contextual features we use class-specific segmentation features inspired by the success of segDPM [4]. We show that the model significantly helps in detecting objects at all scales and is particularly effective at tiny objects as well as extra-large ones.

The supplementary material includes the following items:

- Plots that show the statistics for location and frequency of context classes with respect to different sizes of objects.
- Additional successful and failure cases for detection with contextual information, comparing it with DPM [3]
- Additional successful and failure cases for segmentation with contextual information, comparing it with O2P [1]

Note that in a parallel paper [2] we also provide detailed annotations and analysis for object parts in PASCAL.

References

[1] J. Carreira, R. Caseiroa, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. 1, 9
[2] X. Chen, R. Mottaghi, X. Liu, N.-G. Cho, S. Fidler, and A. Y. Raquel Urtasun. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014. 1
[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010. 1, 7
[4] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation for top-down detection. In *CVPR*, 2013. 1
[5] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. In *CVPR*, 2009. 1
[6] R. Mottaghi, X. Chen, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 1
[7] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV*, 2010. 1
[8] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1

1. Location and Frequency Statistics of Context Classes

The frequency of contextual categories around the objects varies for different sizes of objects. In Figures 1–5, we show the frequency of each context class with respect to different object size percentiles. The statistics are computed within four boxes around the object (the same as four context parts that we had in the paper, but without deformation). The statistics represent the normalized number of pixels for each class. The normalization is done according to the total number of pixels that fall in the boxes of a particular direction.

There are some interesting trends. For instance, the amount of sky in the bottom region of airplanes increases as airplanes become smaller, which shows that small airplanes typically appear in the sky. Another example is that we see more sky pixels in the top region of buses compared to cars, which shows buses are taller than cars.

It is evident that the surroundings of objects have a very biased distribution, which should be exploited particularly when recognizing "difficult" / ambiguous object regions. For example, for tiny objects where little of the structure is visible, or for highly occluded objects, context should play key role in recognition.

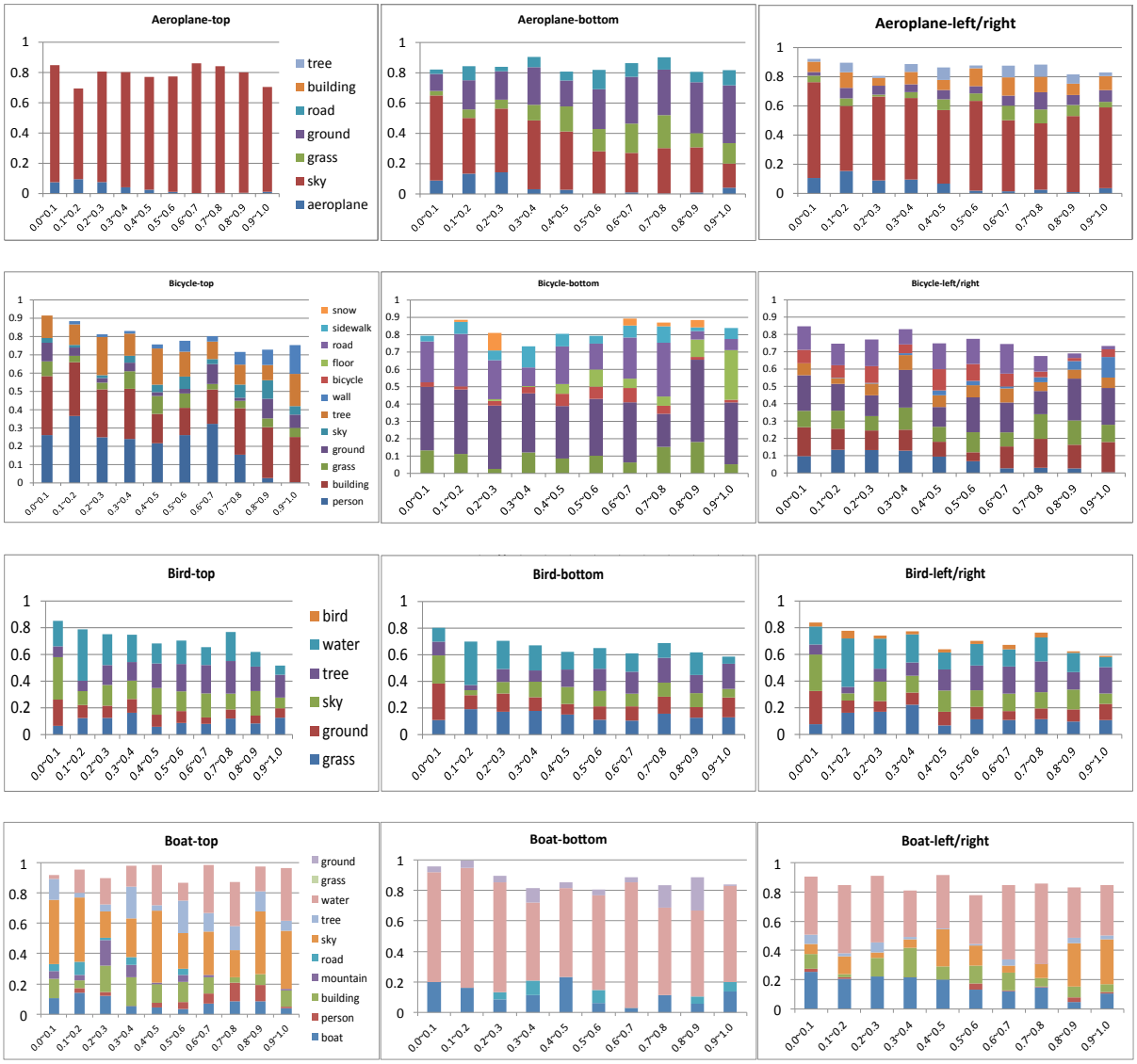

Figure 1. Pixel-wise frequency of context classes in top, bottom, and left/right contextual parts. The x-axis corresponds to size percentile and the y-axis represents the frequency of appearance. Only the most correlated classes are shown.

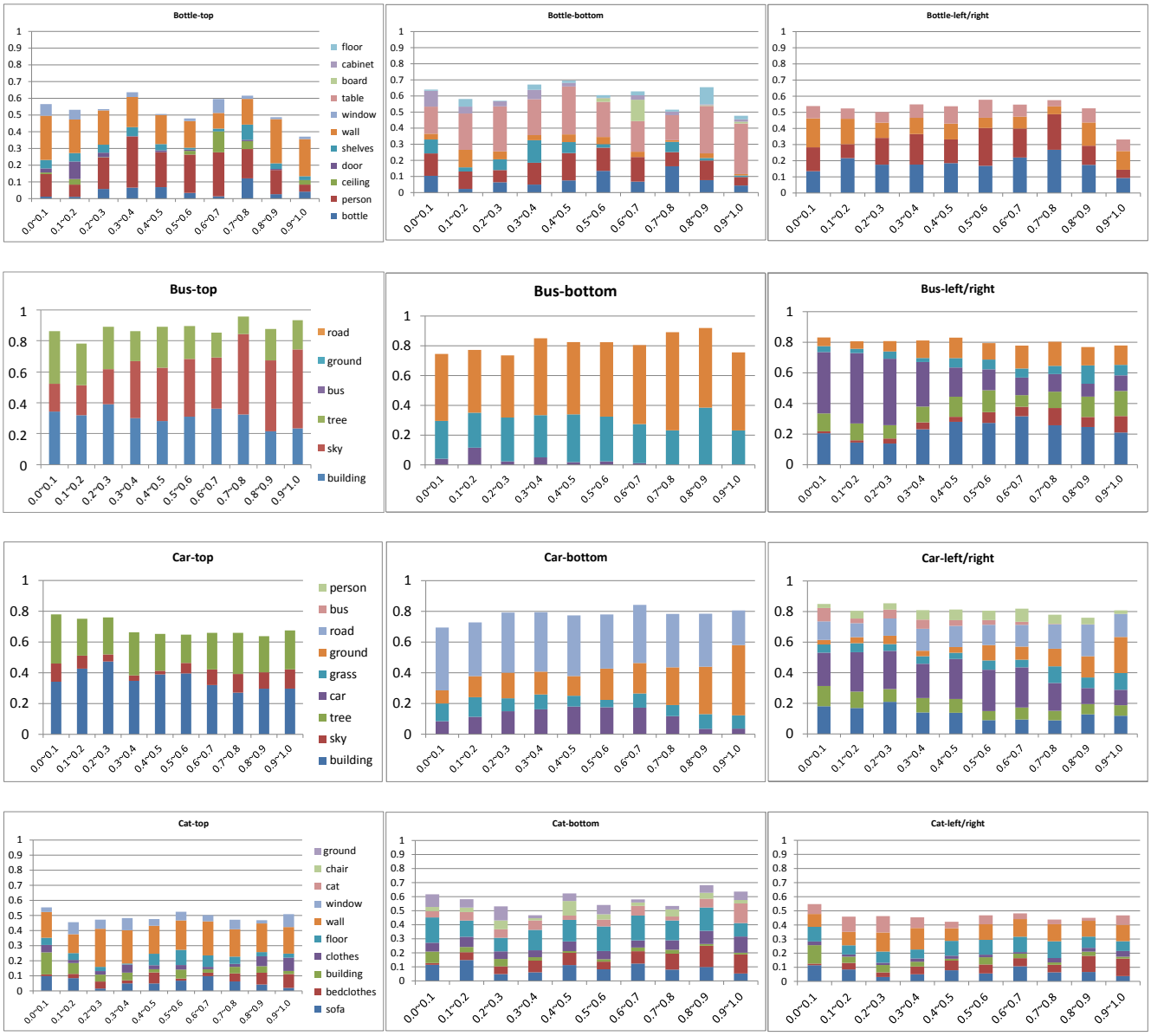

Figure 2. Pixel-wise frequency of context classes in top, bottom, and left/right contextual parts. The x-axis corresponds to size percentile and the y-axis represents the frequency of appearance. Only the most correlated classes are shown.

Figure 3. Pixel-wise frequency of context classes in top, bottom, and left/right contextual parts. The x-axis corresponds to size percentile and the y-axis represents the frequency of appearance. Only the most correlated classes are shown.

Figure 4. Pixel-wise frequency of context classes in top, bottom, and left/right contextual parts. The x-axis corresponds to size percentile and the y-axis represents the frequency of appearance. Only the most correlated classes are shown.

Figure 5. Pixel-wise frequency of context classes in top, bottom, and left/right contextual parts. The x-axis corresponds to size percentile and the y-axis represents the frequency of appearance. Only the most correlated classes are shown.

2. Additional Detection Results

In this section we show additional successful and failure cases for detection with our context model (Figures 6–8).

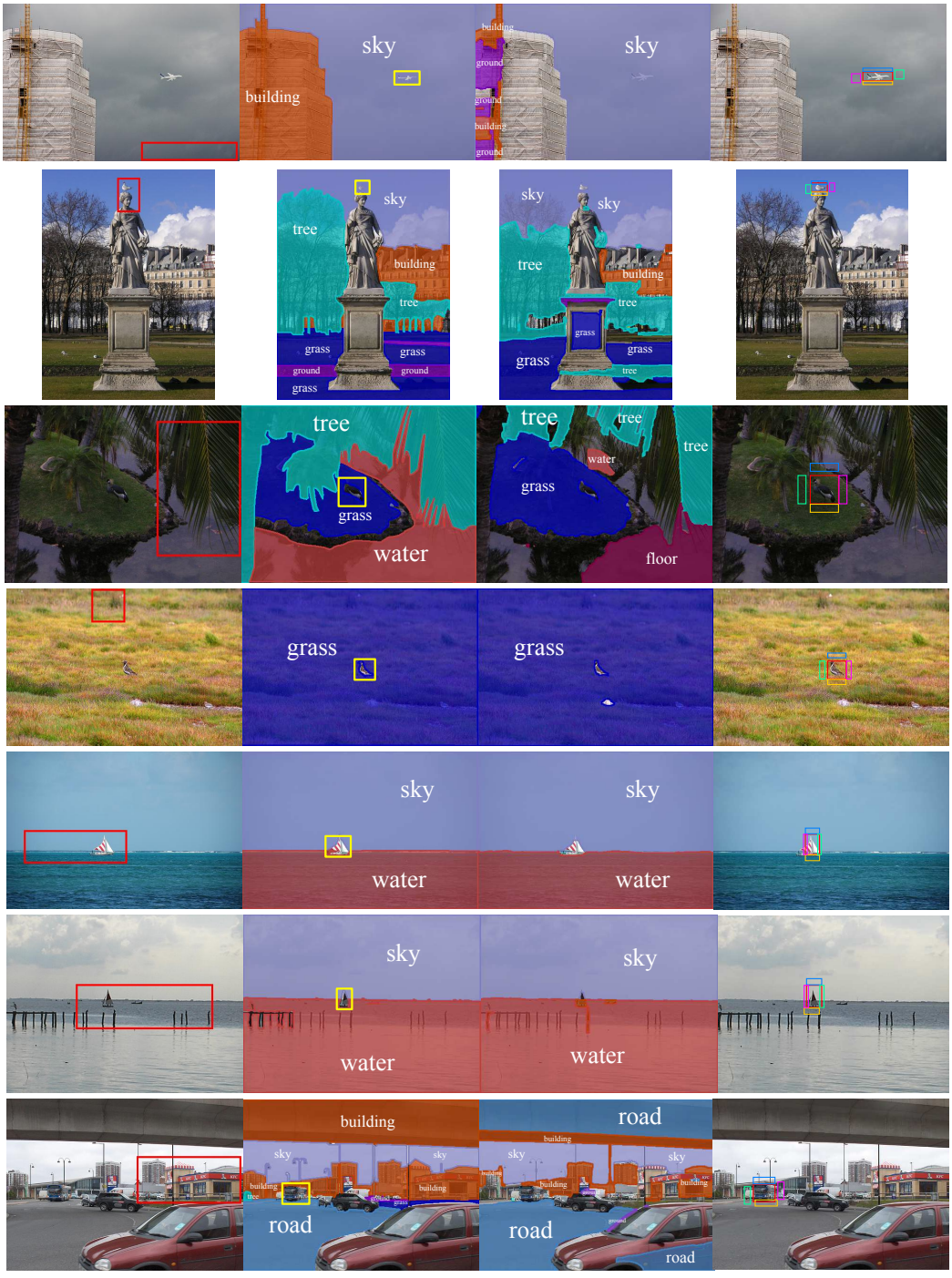

Figure 6. In the first column we show the top detection of DPM [3]. The second column shows groundtruth context labeling and groundtruth object box. Third column is the context prediction result. The last column is the result of our context-aware DPM. Inferred context boxes are also shown with different colors. The original 20 classes of PASCAL are not shown in prediction and groundtruth images.

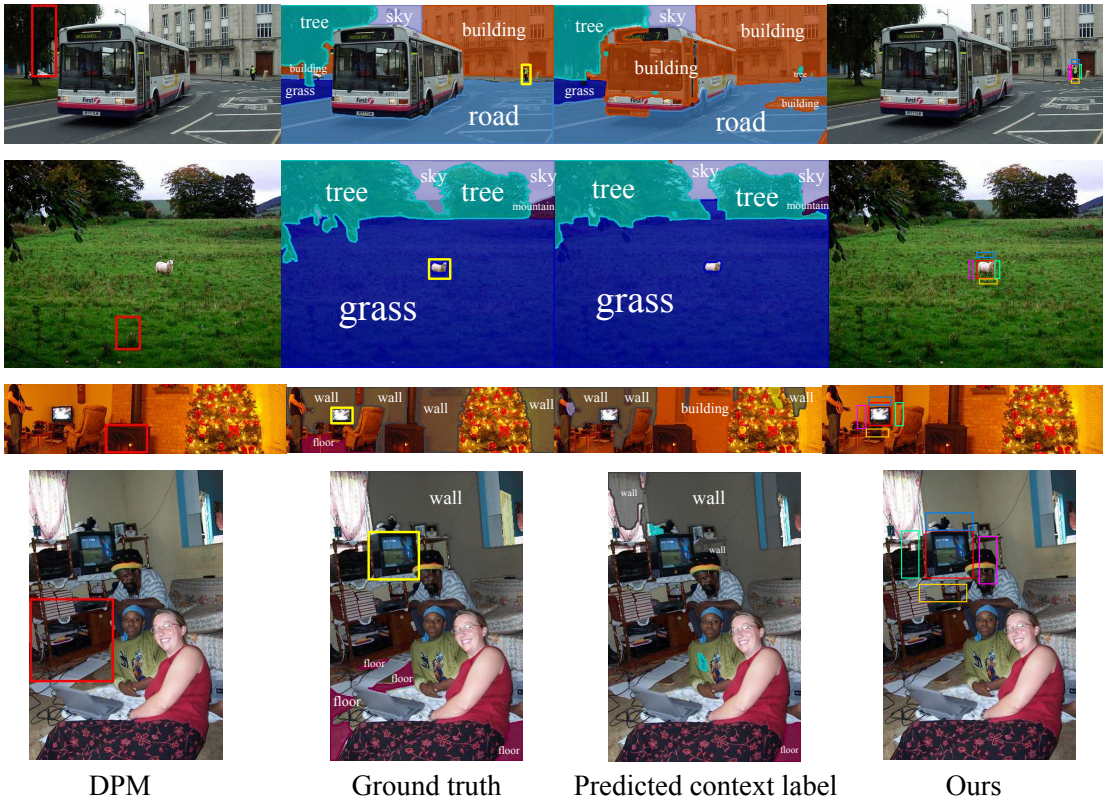

Figure 7. In the first column we show the top detection of DPM. The second column shows groundtruth context labeling and groundtruth object box. Third column is the context prediction result. The last column is the result of our context-aware DPM. Inferred context boxes are also shown with different colors. The original 20 classes of PASCAL are not shown in prediction and groundtruth images.

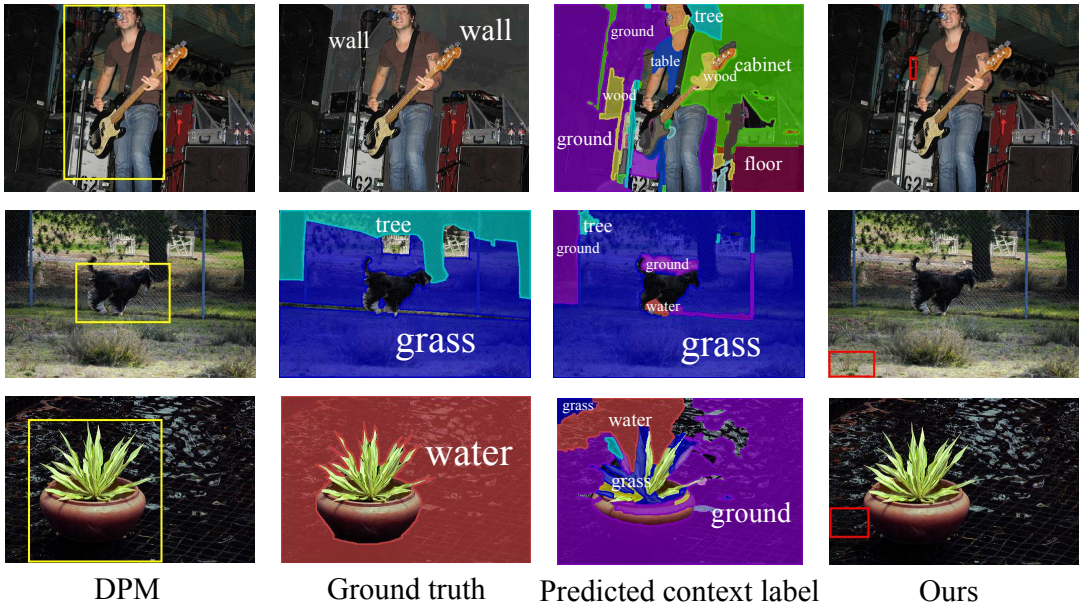

Figure 8. **Failure cases.** In the first column we show the top detection of DPM. The second column shows groundtruth context labeling. Third column is the context prediction result. The last column is the result of our context-aware DPM. The original 20 classes of PASCAL are not shown in prediction and groundtruth images.

3. Additional Segmentation Results

In this section we show examples where the context feature helps or hurts O_2P [1] segmentation (Figures 9–10).

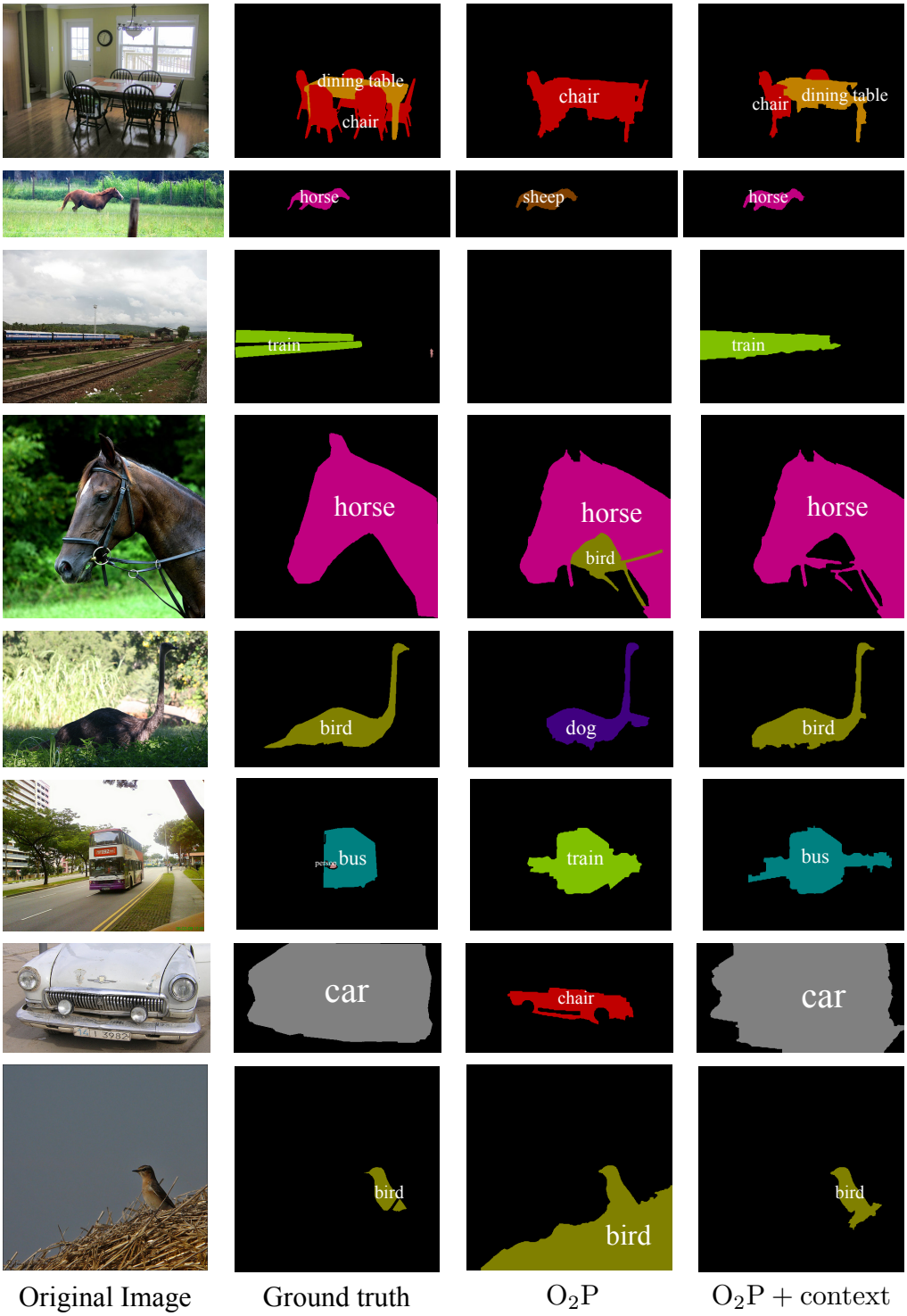

Figure 9. Examples for which the simple context feature provides improvement over O_2P.

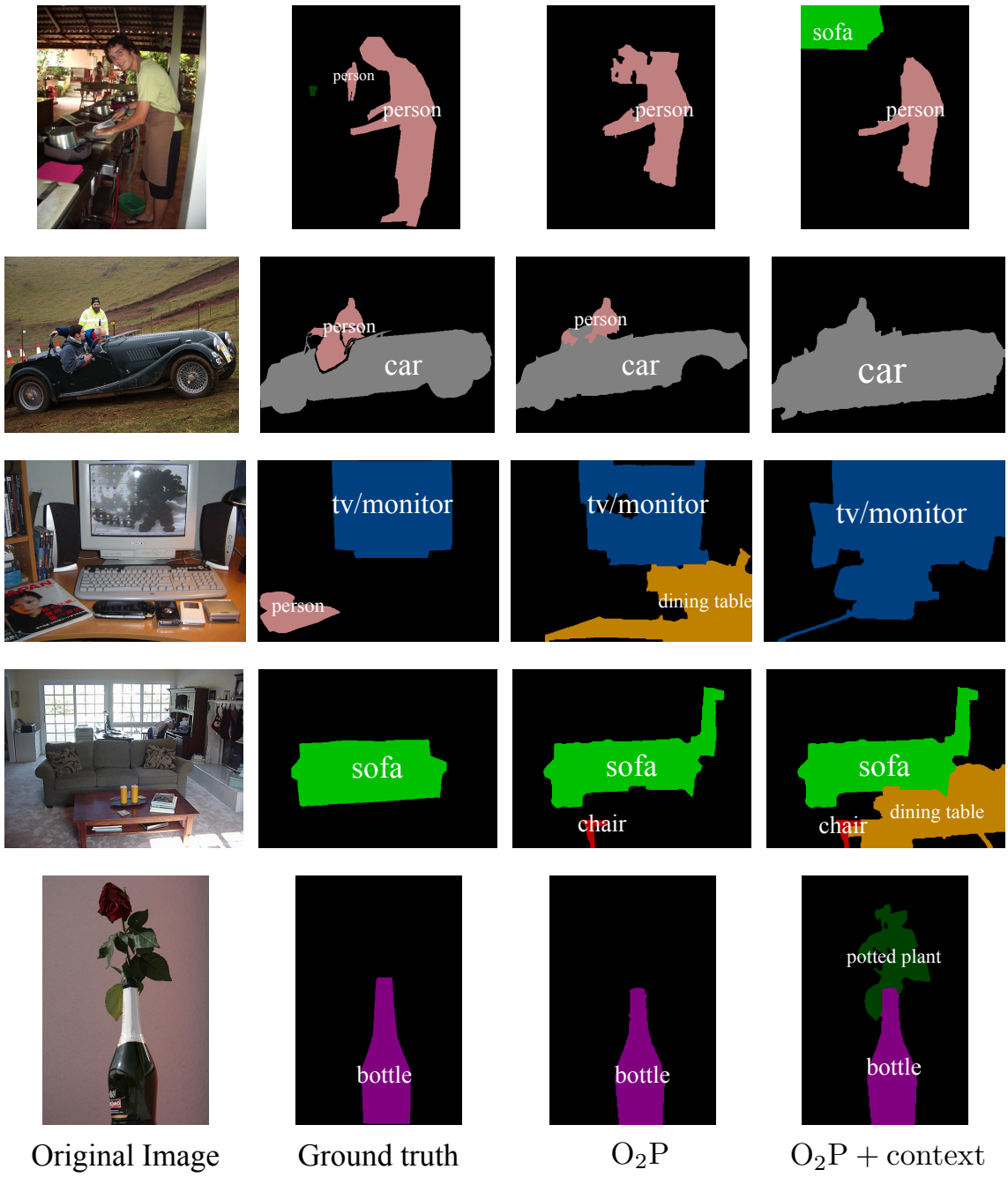

Figure 10. **Failure cases.** Examples that the context feature is misleading for O_2P.