Supplementary Material for "Subcategory-aware Convolutional Neural Networks for Object Proposals and Detection"

Yu Xiang[1], Wongun Choi[2], Yuanqing Lin[3], and Silvio Savarese[4]
[1]University of Washington, [2]NEC Laboratories America, Inc., [3]Baidu, Inc., [4]Stanford University
yuxiang@cs.washington.edu, wongun@nec-labs.com, linyuanqing@baidu.com, ssilvio@stanford.edu

1. Additional Qualitative Results

We present additional qualitative examples obtained from our detection framework on the KITTI detection benchmark [2] and the PASCAL3D+ dataset [8] in this supplementary material.

For car in KITTI and the 12 rigid categories in PASCAL3D+, we use 3D Voxel Patterns (3DVPs) [7] as subcategories in our region proposal network and our detection network. For pedestrian and cyclist in KITTI, we cluster objects according to their orientations to obtain their subcategories. After detecting the objects with bounding boxes, we transfer the segmentation masks from 3DVPs (i.e., segmentation mask of the cluster center in each 3DVP) to the detected objects according to the subcategory classification results. As a result, our method is able to segment the detected objects. Using the provided camera matrices in KITTI, we are able to back-project the detected objects into 3D, so as to localize them in the 3D space. Fig. 1, Fig. 2, Fig. 3 and Fig. 4 present 2D detection and 3D localization results on the KITTI test set, where object detections with scores larger than 0.5 are shown. Fig. 5 and Fig. 6 present detection results on the PASCAL3D+ test set, where object detections with scores larger than 0.7 are shown.

2. Running Time

We implemented our detection framework in Caffe [3], and conducted experiments in the environment with an Intel Xeon GPU 2.8GHz and a NVIDIA GeForce GTX TITAN X Graphics Card. Table 1 presents the running time of our detection framework on KITTI and PASCAL (PASCAL3D+ and PASCAL VOC 2007). For region proposal, we generated around 2,000 regions per image on the three datasets. Since we used more image scales on KITTI than PASCAL and the size of KITTI images is also larger, it takes 2.3 seconds to process one KITTI image, while 1.2 seconds is needed to process one PASCAL image in our experimental setting. We also present the running time of Faster R-CNN [5] in Table 1 for comparison.

	Region Proposal	Detection	Total
Ours			
KITTI (AlexNet [4])	1.5s	0.8s	2.3s
PASCAL (VGG16 [6])	0.8s	0.4s	1.2s
Faster R-CNN [5]			
KITTI (AlexNet [4])	1.1s	0.9s	2.0s
PASCAL (VGG16 [6])	0.3s	0.4s	0.7s

Table 1. Running time of our detection framework on KITTI [2] and PASCAL (PASCAL3D+ [8] and PASCAL VOC 2007 [1]).

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 1

[2] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 1

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 1

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 1

[5] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1

[6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[7] Y. Xiang, W. Choi, Y. Lin, and S. Savarese. Data-driven 3d voxel patterns for object category recognition. In *CVPR*, pages 1903–1911, 2015. 1

[8] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, pages 75–82, 2014. 1

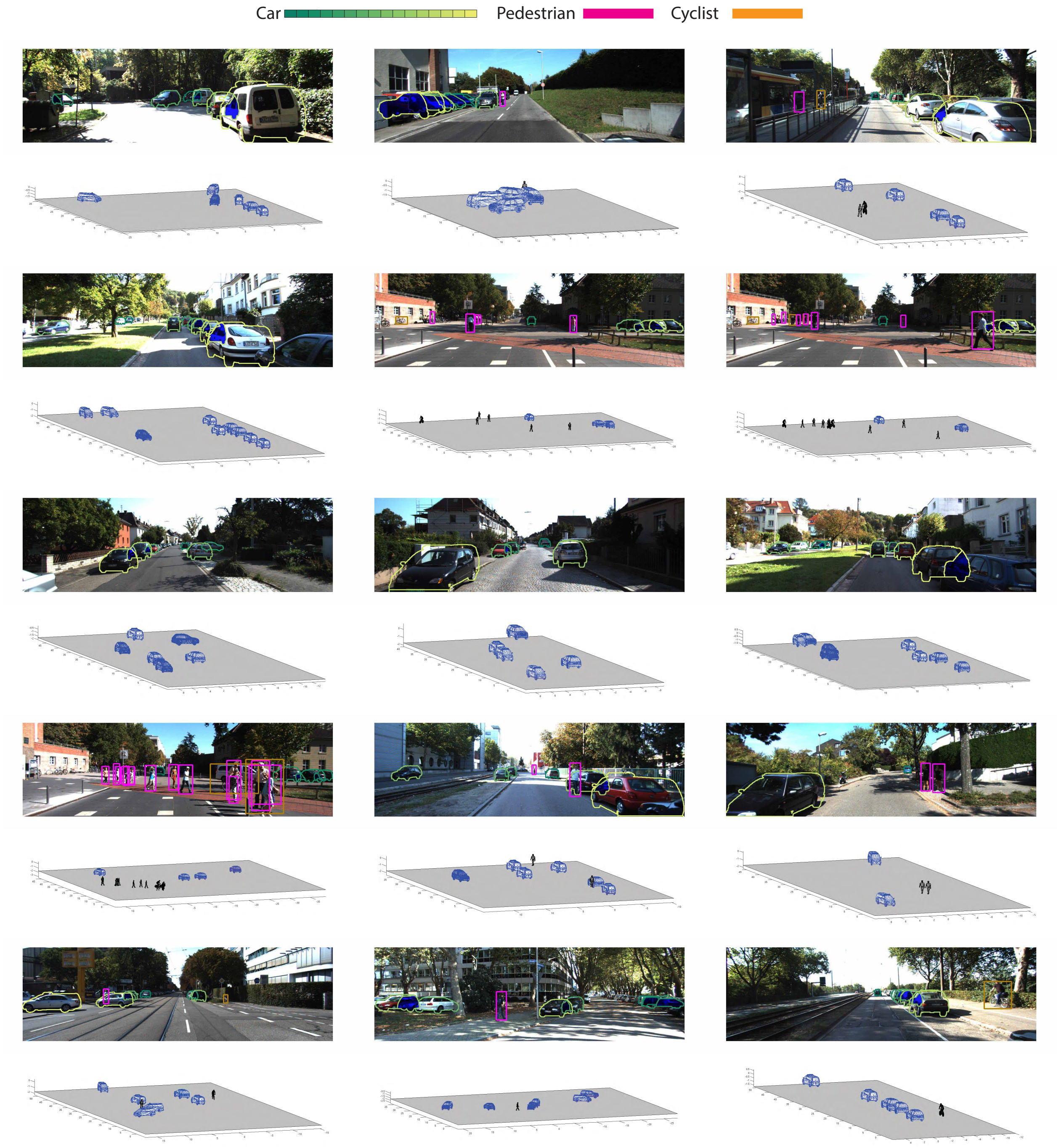

Figure 1. 2D detection and 3D localization results on the KITTI test set. Detections with scores larger than 0.5 are shown. Blue regions in the images are the estimated occluded areas.

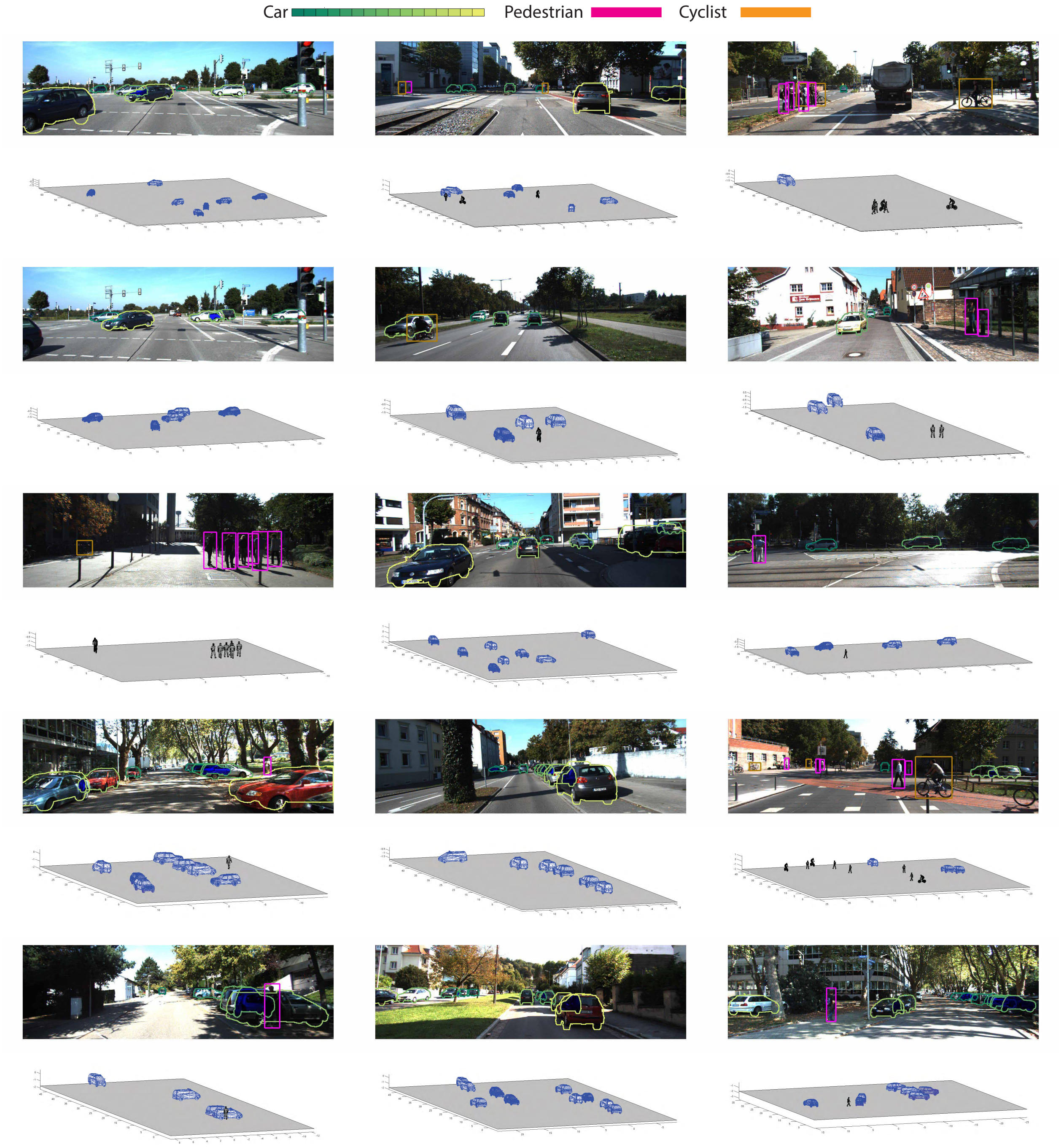

Figure 2. 2D detection and 3D localization results on the KITTI test set. Detections with scores larger than 0.5 are shown. Blue regions in the images are the estimated occluded areas.

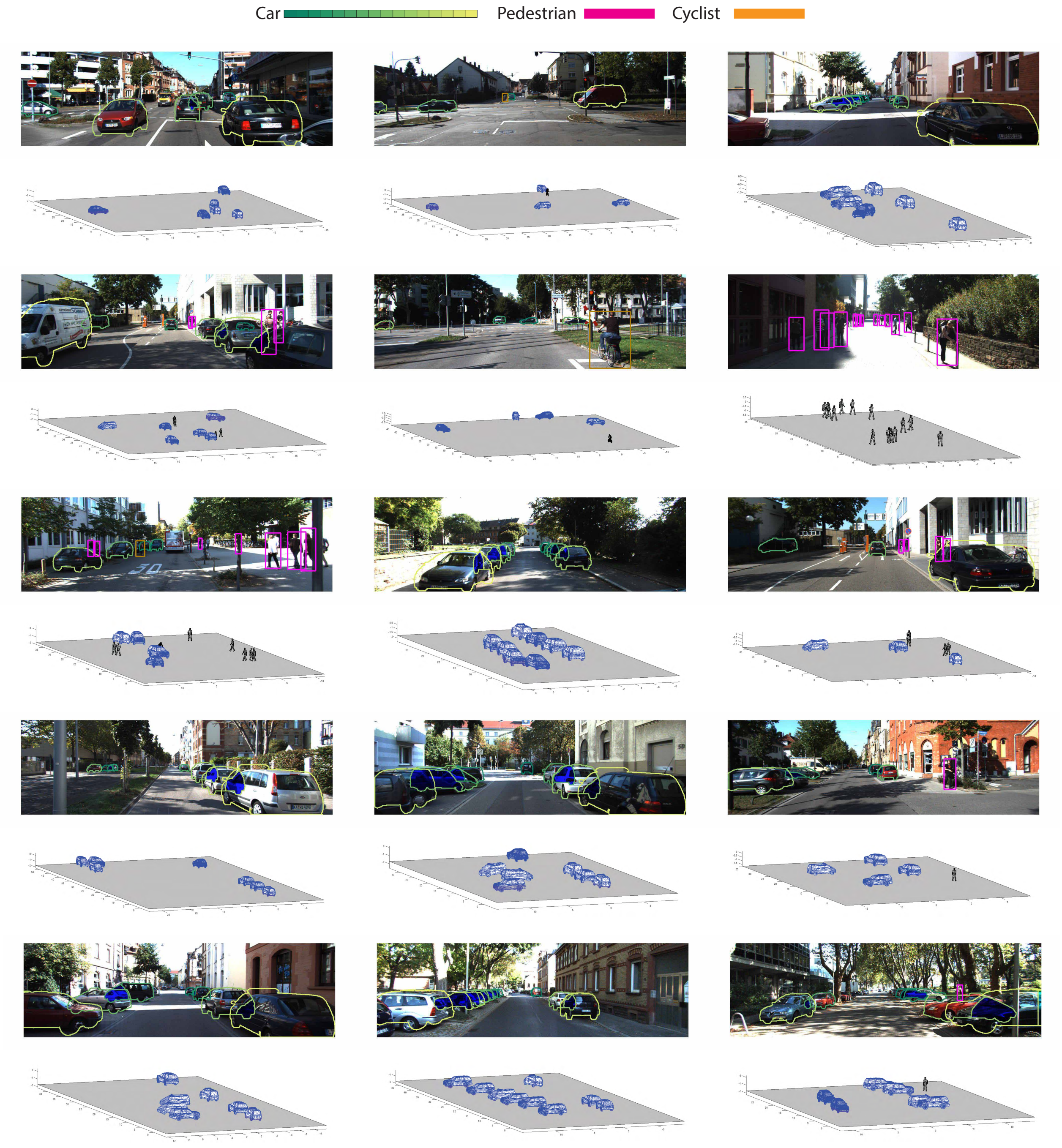

Figure 3. 2D detection and 3D localization results on the KITTI test set. Detections with scores larger than 0.5 are shown. Blue regions in the images are the estimated occluded areas.

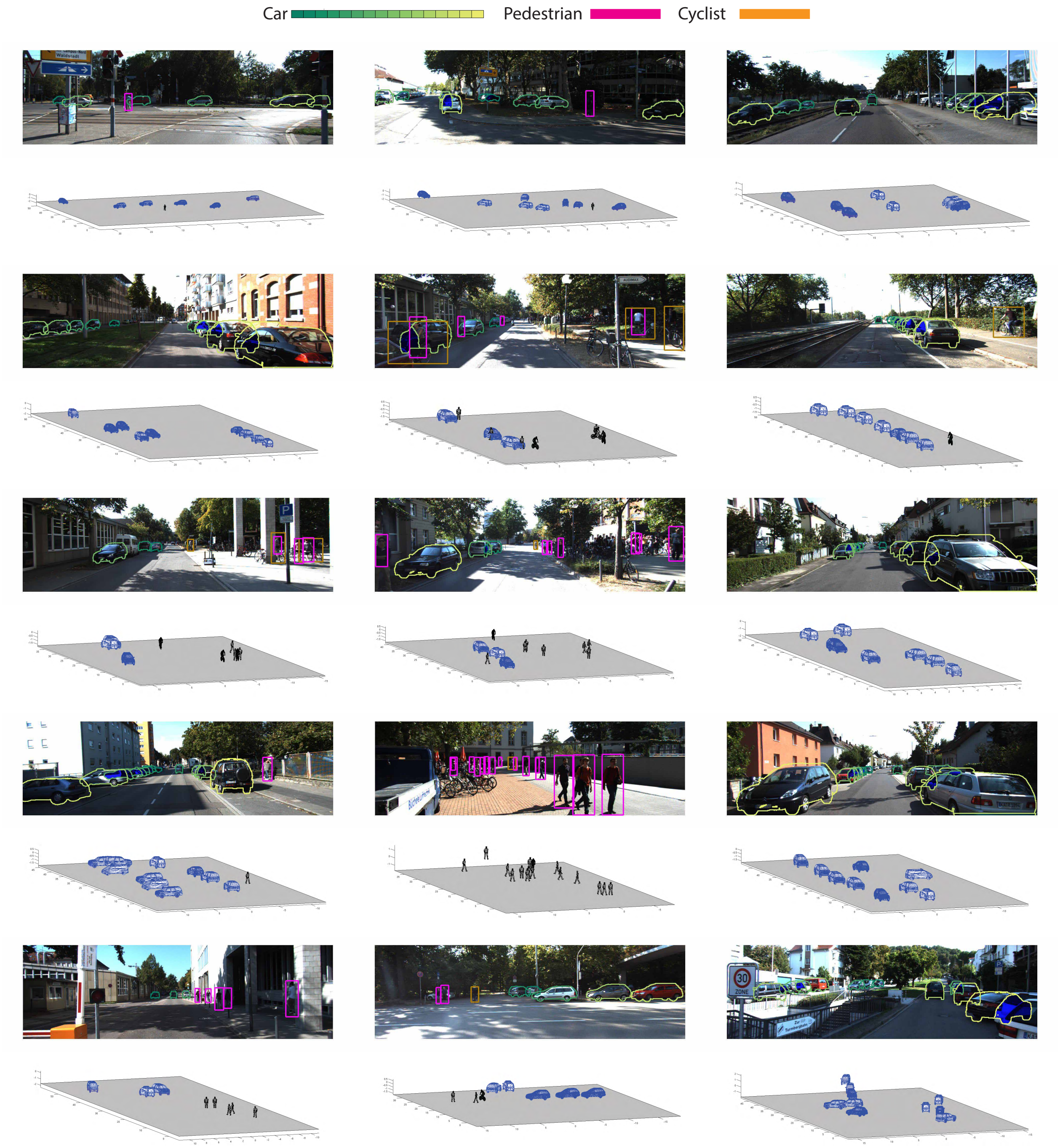

Figure 4. 2D detection and 3D localization results on the KITTI test set. Detections with scores larger than 0.5 are shown. Blue regions in the images are the estimated occluded areas.

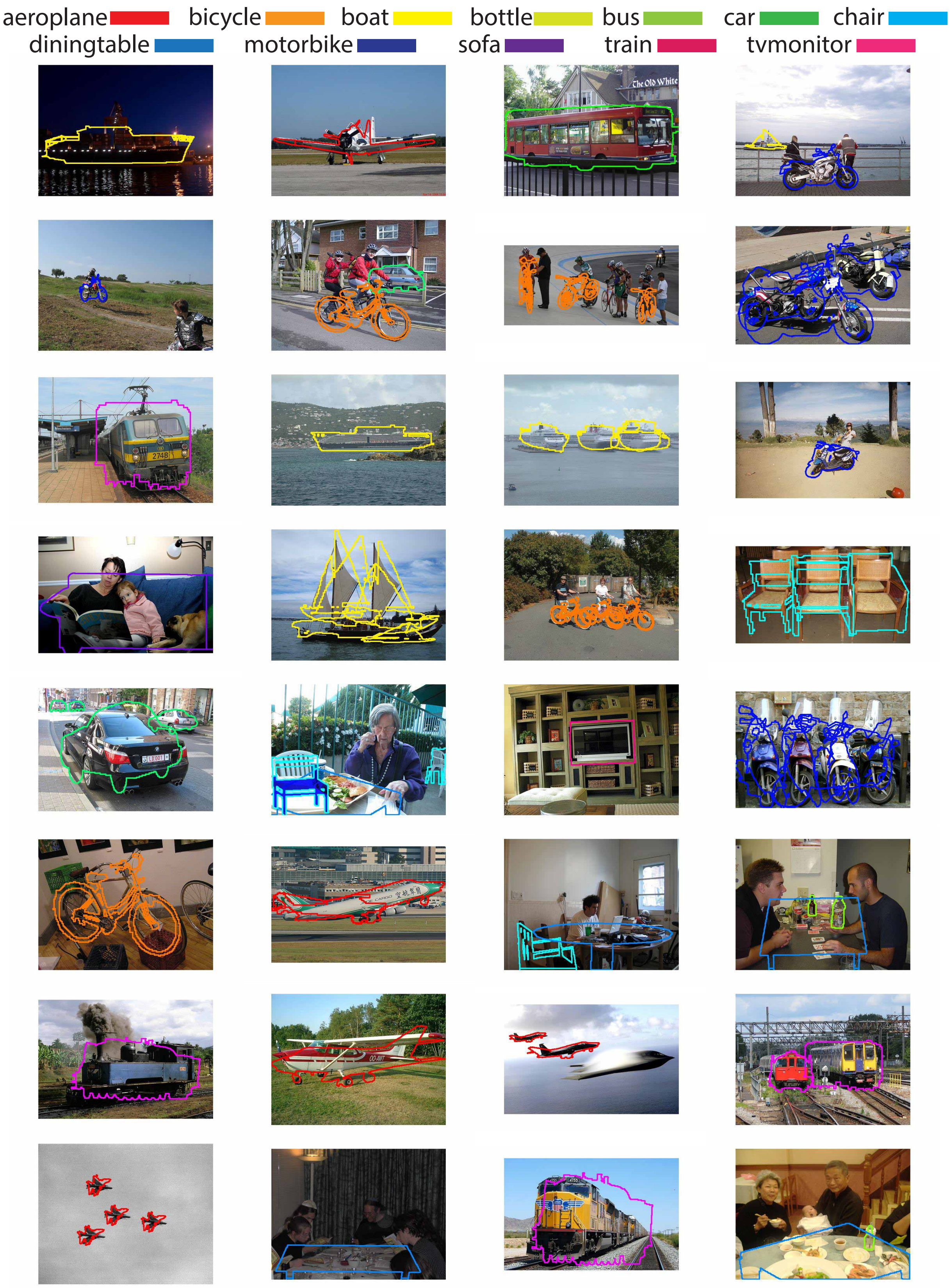

Figure 5. 2D detection results on the PASCAL3D+ test set. Detections with scores larger than 0.7 are shown.

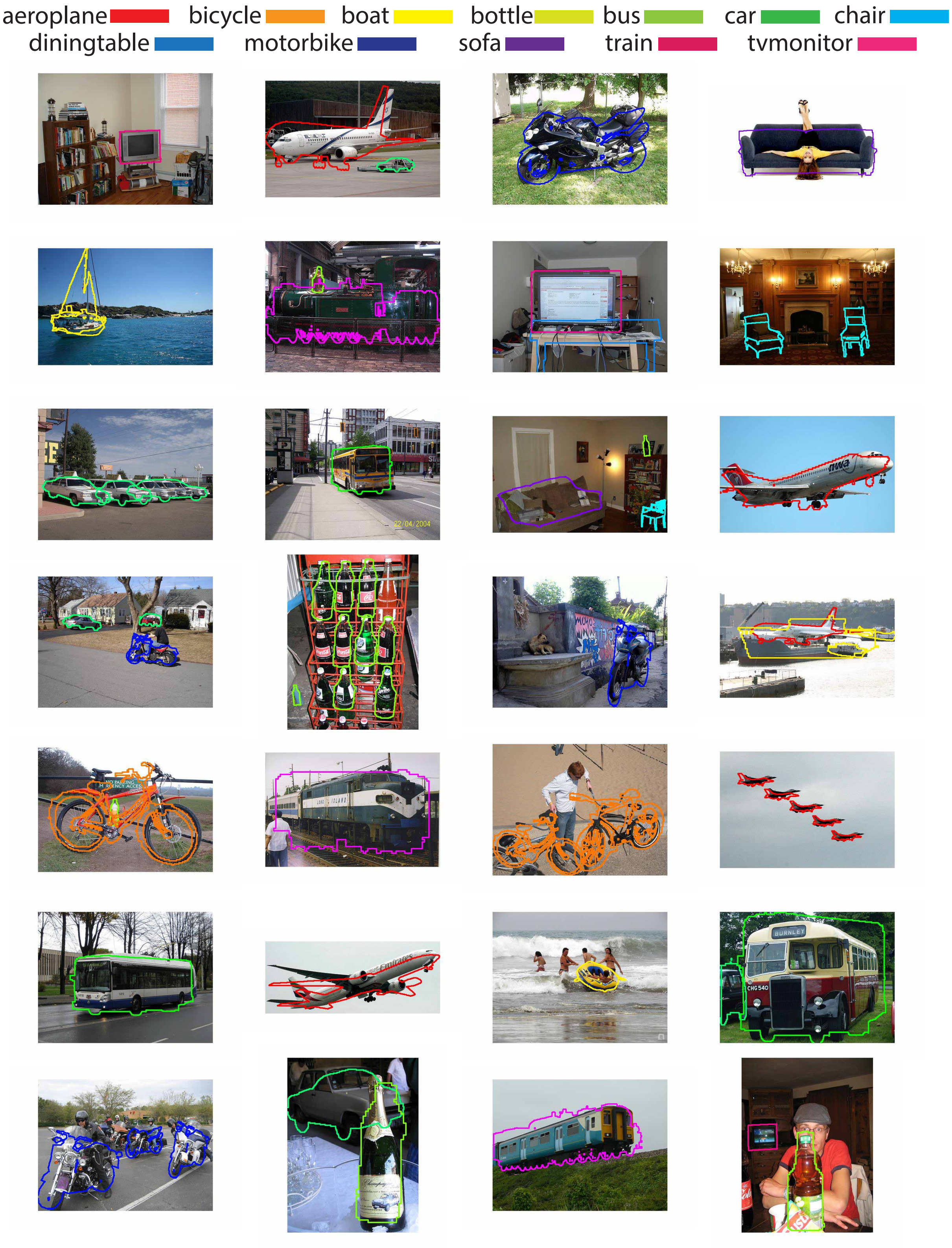

Figure 6. 2D detection results on the PASCAL3D+ test set. Detections with scores larger than 0.7 are shown.