Supplementary Material for "Data-Driven 3D Voxel Patterns for Object Category Recognition"

Yu Xiang[1,2], Wongun Choi[3], Yuanqing Lin[3], and Silvio Savarese[1]
[1]Stanford University, [2]University of Michigan at Ann Arbor, [3]NEC Laboratories America, Inc.
yuxiang@umich.edu, {wongun, ylin}@nec-labs.com, ssilvio@stanford.edu

We present the implementation details of our object category recognition framework and additional qualitative examples on the KITTI dataset [5] and the OutdoorScene dataset [8] in this supplementary material.

1. Implementation Details

1.1. Voxelization

In building the 3D voxel exemplars, we voxelize a 3D CAD model into a distribution of 3D voxels. Since 3D CAD models from the web repositories, such as the Trimble 3D Warehouse [1], are usually irregular and not water-tight. We employ the volumetric depth map fusion technique, which is widely used in dense 3D reconstruction in the literature [7], to build the voxel representation of a 3D CAD model. Fig. 1 illustrates our voxelization process. We first render depth images of a CAD model from different viewpoints (Fig. 1(a)). In our implementation, we render from 8 azimuths and 6 elevations, which produces 48 depth images. Then we fuse these depth images to obtain a 3D point cloud on the surface of the object (Fig. 1(b)). Finally, we voxelize the 3D space and determine which voxels are inside or outside the object using the surface point cloud (Fig. 1(c)). We experimented with different sizes of the 3D voxel space. There is a tradeoff between computational efficiency and representation power according to different sizes of 3D voxel space. We found that a 50 × 50 × 50 voxel space works well in our experiments.

1.2. 3D Clustering

We discover 3D Voxel Patterns (3DVPs) by clustering 3D voxel exemplars in a uniform 3D space. We first review the similarity metric between two exemplars defined in our paper. A 3D voxel exemplar is represented by a feature vector $\mathbf{x}$ with dimension N^3, where N denotes the size of the 3D voxel space. The elements of the feature vector takes values from a finite set $\mathcal{S} = \{0, 1, 2, 3, 4\}$, which encodes the visibility of the voxels, i.e., 0 for empty voxels, 1 for visible voxels, 2 for self-occluded voxels, 3 for voxels occluded by other objects, and 4 for truncated voxels. The

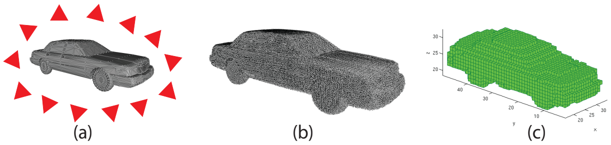

Figure 1. Illustration of volumetric depth map fusion to voxelize a CAD model. (a) We render depth images of the CAD from different viewpoints. (b) We fuse all the depth images to obtain the 3D point cloud on the surface the object. (c) Using the surface points, we voxlize the 3D space and determine which voxels are inside or outside the object.

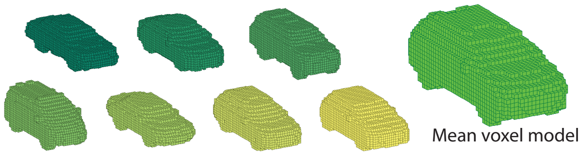

Figure 2. The mean voxel car model obtained by averaging 7 different 3D CAD model of cars. We simply aggregate all the occupied voxels from all the CAD models, and retain voxels shared by at least K models in the mean model, where $K = 2$ in our implementation.

similarity metric between two feature vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ of two 3D voxel exemplars is defined as:

$$s(\mathbf{x_1}, \mathbf{x_2}) = \frac{|\mathcal{S}|}{N^3} \sum_{i=1}^{N^3} \mathbb{1}(x_1^i = x_2^i) \cdot w(x_1^i), \quad \text{s.t.,} \sum_{i=0}^{|\mathcal{S}|-1} w(i) = 1, \qquad (1)$$

where x_1^i and x_2^i are the ith element of $\mathbf{x_1}$ and $\mathbf{x_2}$ respectively, $\mathbb{1}$ is the indicator function, and $w(i)$ is the weight for voxel status i.

The definition in Eq. (1) is general such that the weights can be designed for different applications. In our implementation, for object categories with small intra-class variations, such as cars, we propose to use a mean voxel representation as show in Fig. 2. Besides using the closest 3D CAD models, all the 3D voxel exemplars are also represented with the mean voxel model which is only used in the 3D clustering process. Specifically, we assign visibility labels to the mean voxel model for each exemplar according to its 2D segmentation mask in the same way as we build the 3D voxel exemplar. As a result, we marginalize the shape variation in the 3D clustering process, i.e., the voxel space is reduced to all the occupied voxels of the mean model and $\mathcal{S} = \{1, 2, 3, 4\}$. So the 3D clustering generates patterns which are consistent in terms of 3D object pose, occlusion and truncation. In our implementation, we simply define all the weights for voxel status to be $1/4$. Before applying the clustering algorithm, we do left-right flipping for each object instance in order to double the number of 3D voxel exemplars in training.

1.3. Training 3DVP Detectors

In training an ACF [2] detector for a 3DVP, we use all the image patches in the cluster of the 3DVP as positive examples, and negative examples are harvested from the positive images. For car detection in the KITTI dataset [5], a negative bounding box is used if its overlap with any positive bounding box is less than 60%. Note that car detection in KITTI requires 70% bounding box overlap with the ground truth annotation. After training all the 3DVP detectors, we can apply them to test images. As aslo noted in [6], we find out that it is not necessary to carefully calibrate the detection scores among the ACF detectors.

2. Additional Qualitative Results

2.1. 3DVPs from the KITTI Dataset

Fig. 3, Fig. 4 and Fig. 5 show all the 227 3DVPs we built from the KITTI training set. These 3DVPs are obtained by clustering 57,224 3D voxel exemplars with the affinity propagation algorithm [4]. Among the 227 3DVPs, 91 3DVPs are fully visible, 121 3DVPs are partially occluded and 15 3DVPs are truncated. As we can see from the figures, 3DVPs capture various viewpoints, occlusion patterns and truncation patterns of the object category.

2.2. Results on the KITTI Dataset

Fig. 6 and Fig. 8 show additional qualitative results for car recognition on our validation split of the KITTI dataset, where we compare our method w/wo occlusion reasoning and DPM [3] in terms of 2D recognition and 3D localization. As we can see, severe false alarms are removed with occlusion reasoning. Fig. 10 and Fig. 11 show additional qualitative results on the KITTI test set. Our method is able to recognize detailed 2D/3D properties of the objects.

2.3. Results on the OutdoorScene Dataset

Fig. 12 and Fig. 13 show qualitative results for car recognition on the OutdoorScene dataset [8], where detections at 1 false positive per image (fppi) are displayed for each image. Note that we directly apply the 3DVP detectors trained on the KITTI dataset to the OutdoorScene dataset. As we can see from these qualitative results, our 3DVP detectors can generalize to different scenarios, such as city and parking lots scenes.

References

[1] Trimble 3d warehouse. http://3dwarehouse.sketchup.com. 1

[2] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *TPAMI*, 2014. 2

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010. 2, 6, 8

[4] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007. 2

[5] 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, 2

[6] E. Ohn-Bar and M. M. Trivedi. Fast and robust object detection using visual subcategories. In *CVPRW*, pages 179–184, 2014. 2

[7] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, volume 1, pages 519–528, 2006. 1

[8] Y. Xiang and S. Savarese. Object detection by 3d aspectlets and occlusion reasoning. In *ICCVW*, pages 530–537, 2013. 1, 2

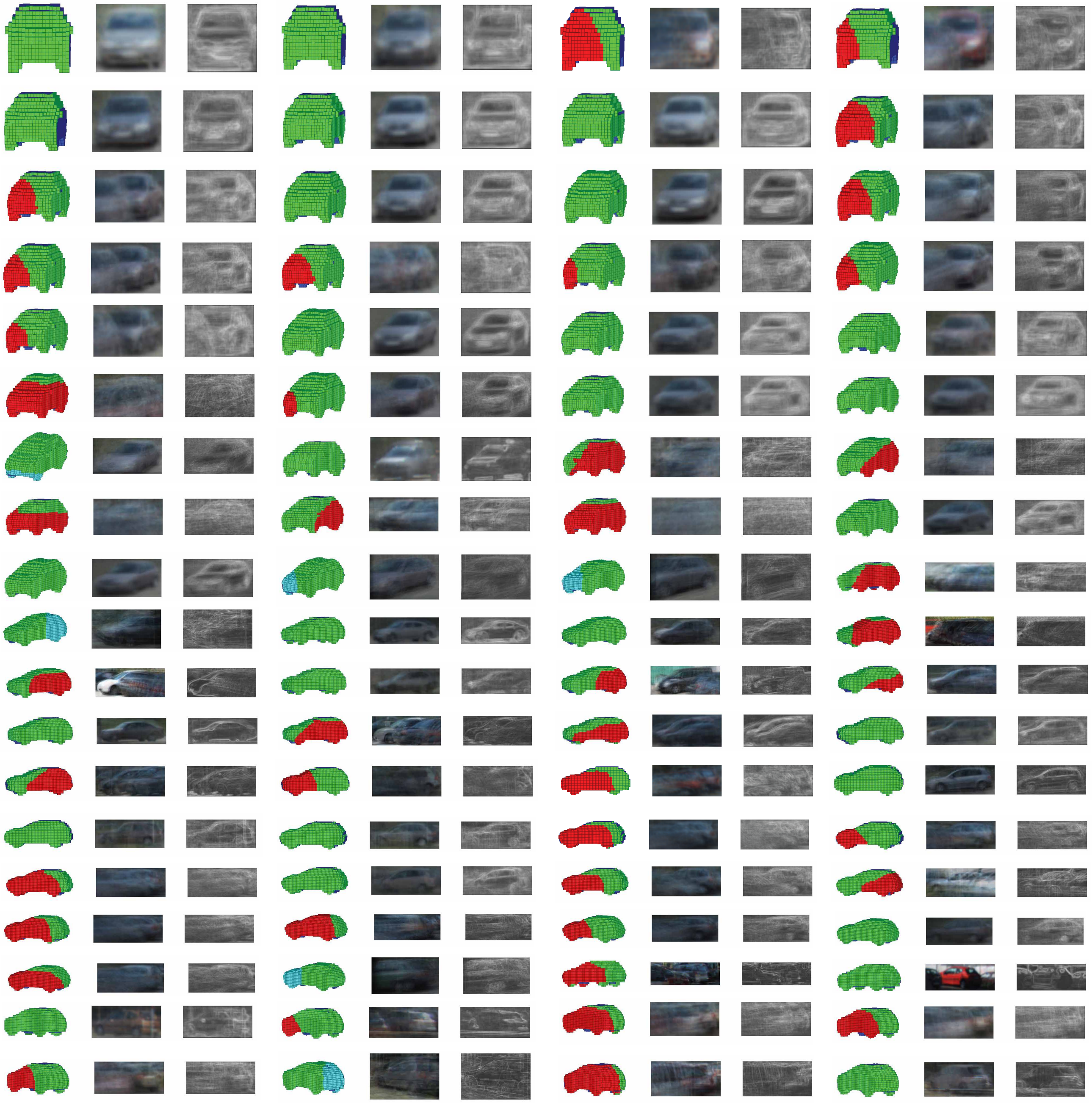

Figure 3. Visualization of the first 76 3DVPs among the 227 3DVPs we built from the KITTI training set. We show the 3D mean voxel model of the cluster center, the average RGB image, and the average gradient image of each 3DVP. Green, red and cyan voxels are visible, occluded and truncated respectively.

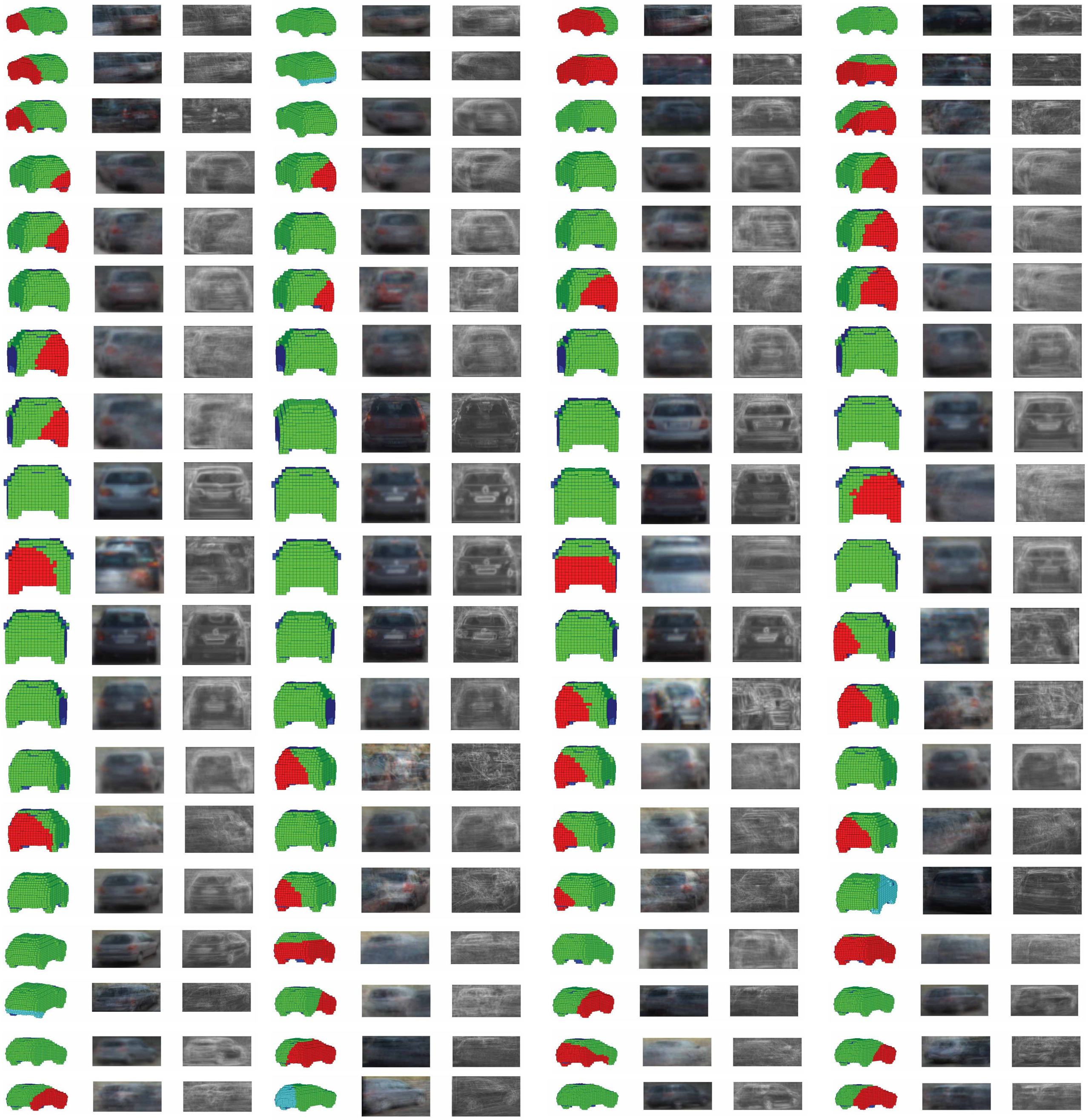

Figure 4. Visualization of the second 76 3DVPs among the 227 3DVPs we built from the KITTI training set. We show the 3D mean voxel model of the cluster center, the average RGB image, and the average gradient image of each 3DVP. Green, red and cyan voxels are visible, occluded and truncated respectively.

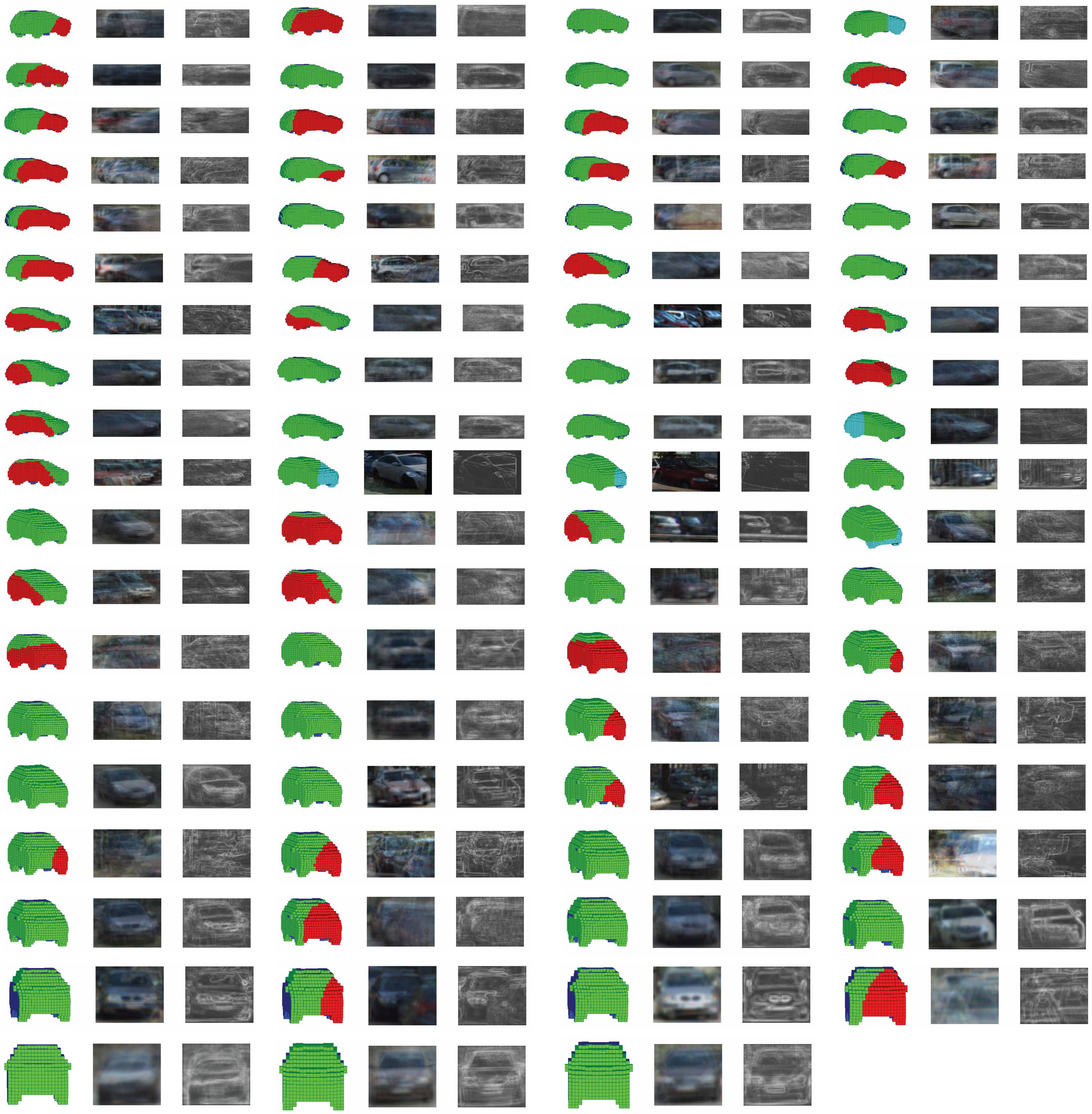

Figure 5. Visualization of the last 75 3DVPs among the 227 3DVPs we built from the KITTI training set. We show the 3D mean voxel model of the cluster center, the average RGB image, and the average gradient image of each 3DVP. Green, red and cyan voxels are visible, occluded and truncated respectively.

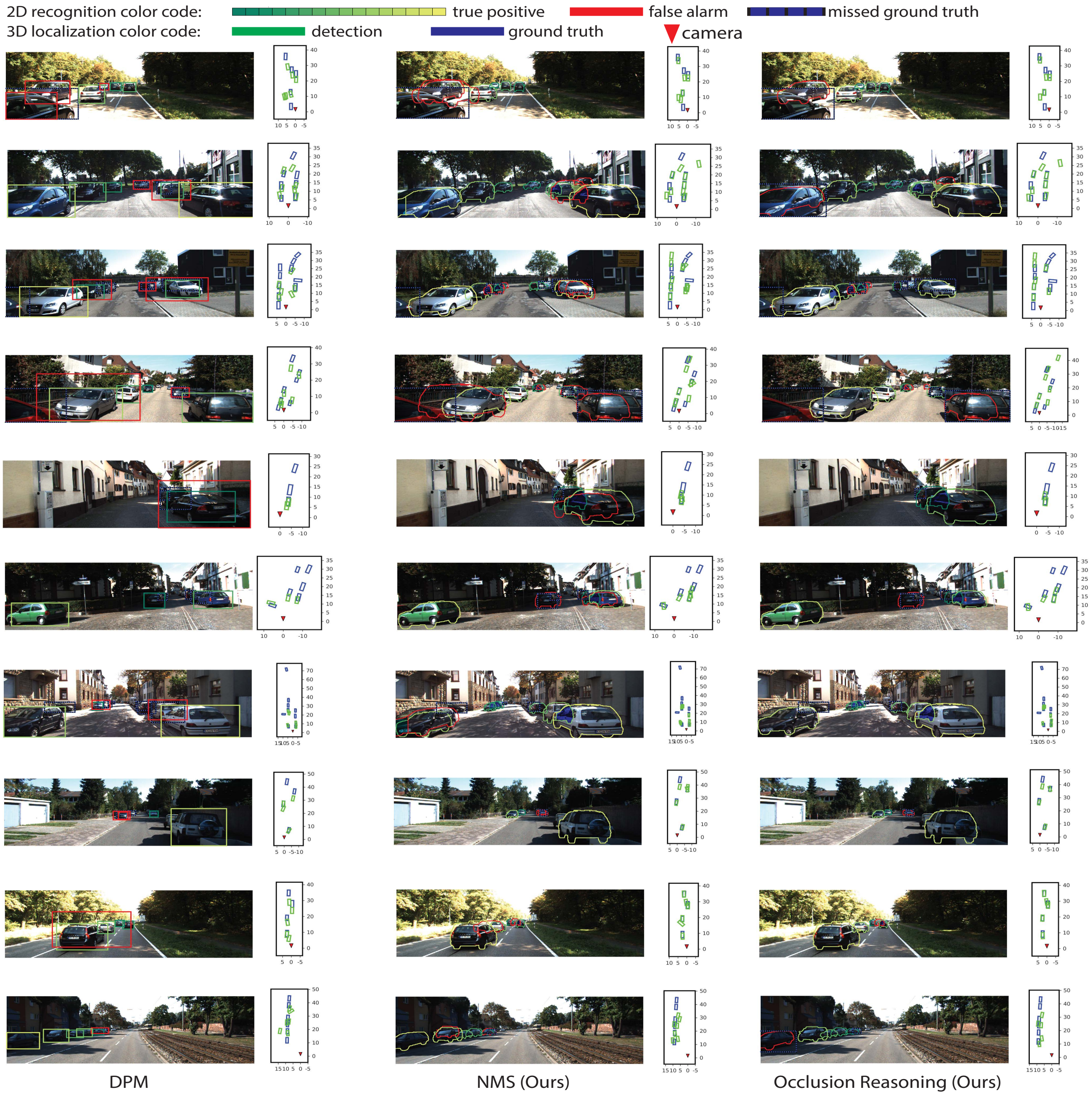

Figure 6. Car recognition results on the KITTI validation set. We compare our method w/wo occlusion reasoning and DPM [3]. Detections at 1 false positive per image (fppi) for the three methods are shown. Blue regions in the images are the estimated occluded areas. Note that severe false alarms in NMS disappear with occlusion reasoning. Please see Fig. 7 for the zoomed in 3D localization results.

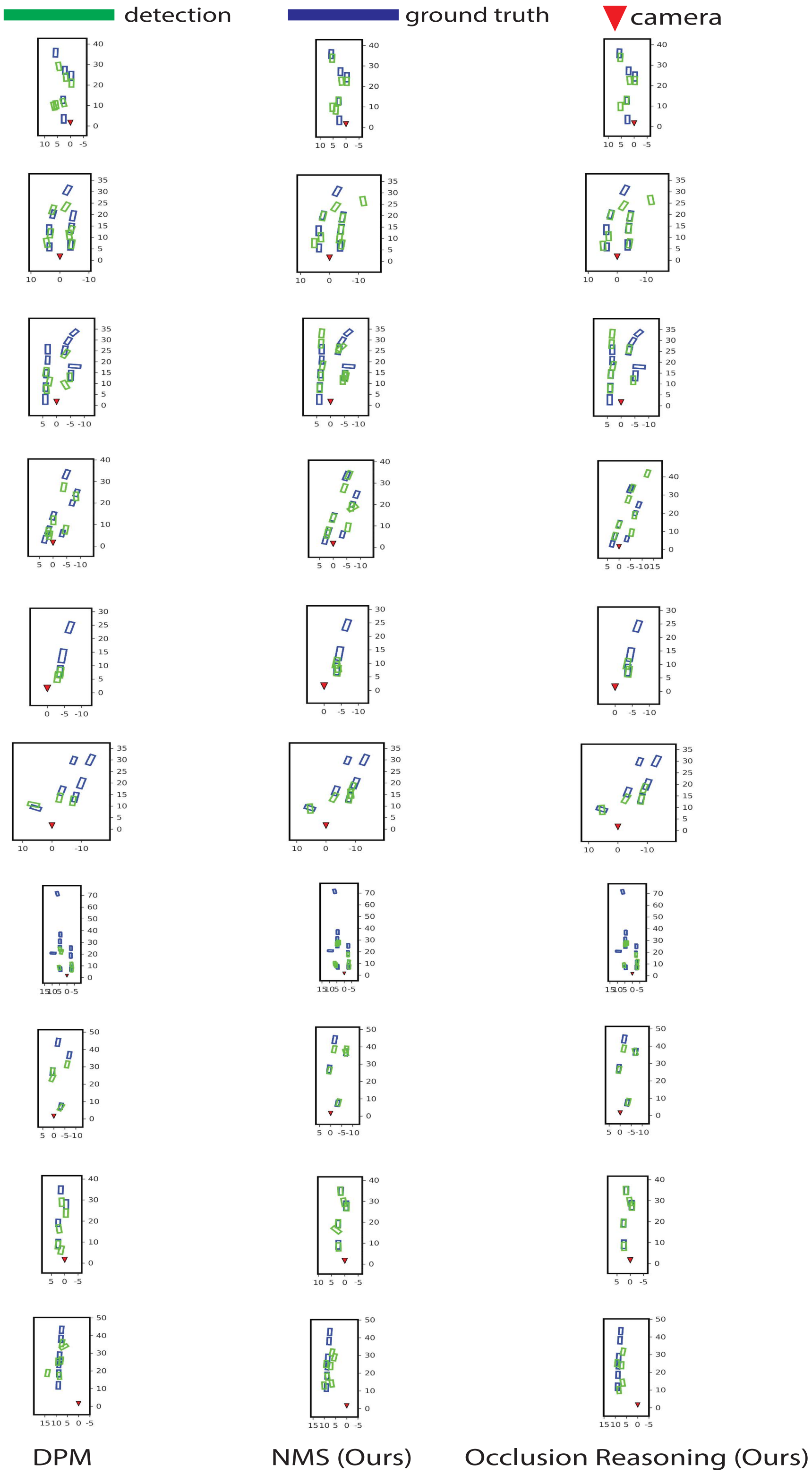

Figure 7. Zoomed in version of the 3D localization results in Fig. 6.

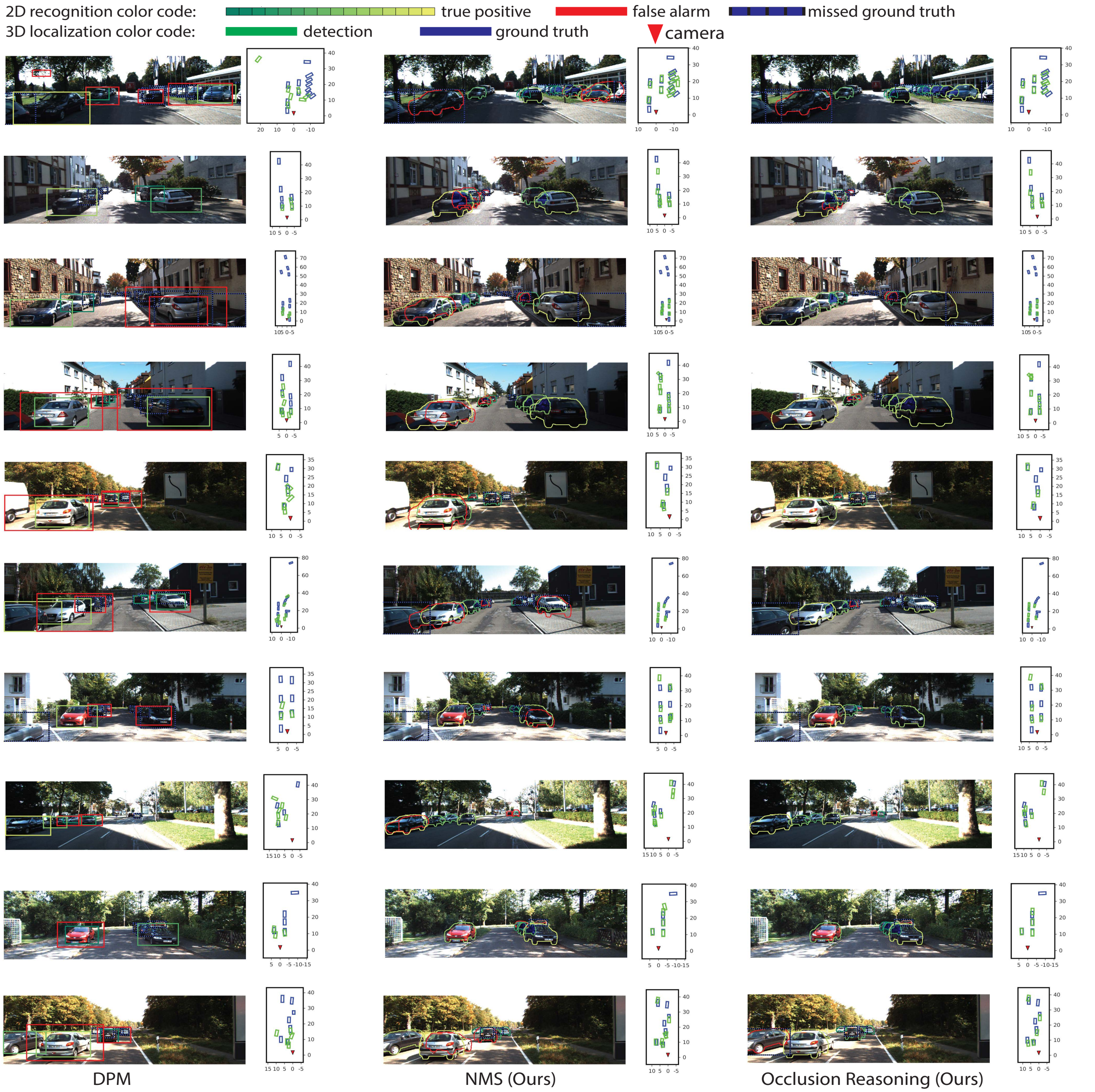

Figure 8. Car recognition results on the KITTI validation set. We compare our method w/wo occlusion reasoning and DPM [3]. Detections at 1 false positive per image (fppi) for the three methods are shown. Blue regions in the images are the estimated occluded areas. Note that severe false alarms in NMS disappear with occlusion reasoning. Please see Fig. 9 for the zoomed in 3D localization results.

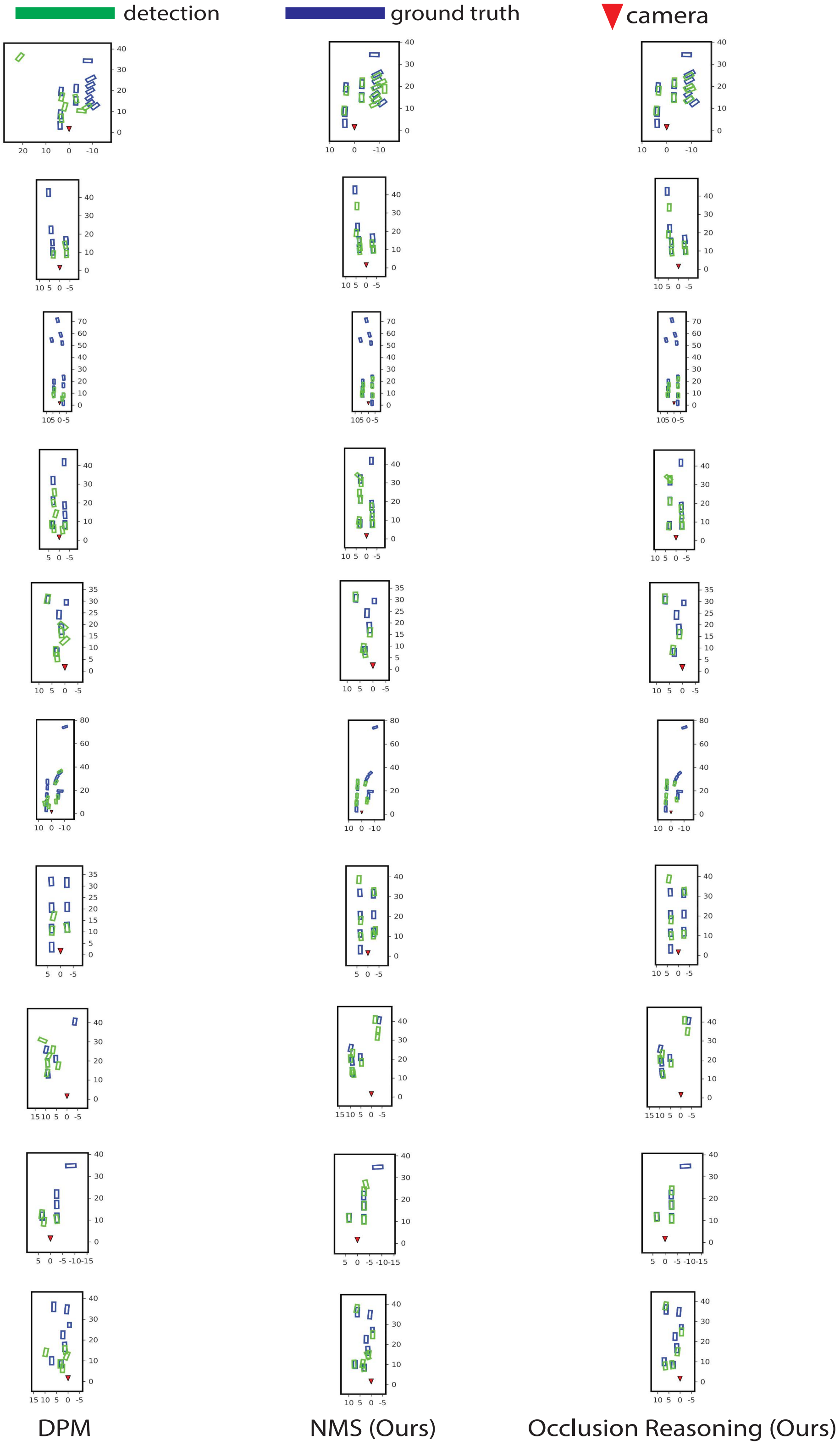

Figure 9. Zoomed in version of the 3D localization results in Fig. 8.

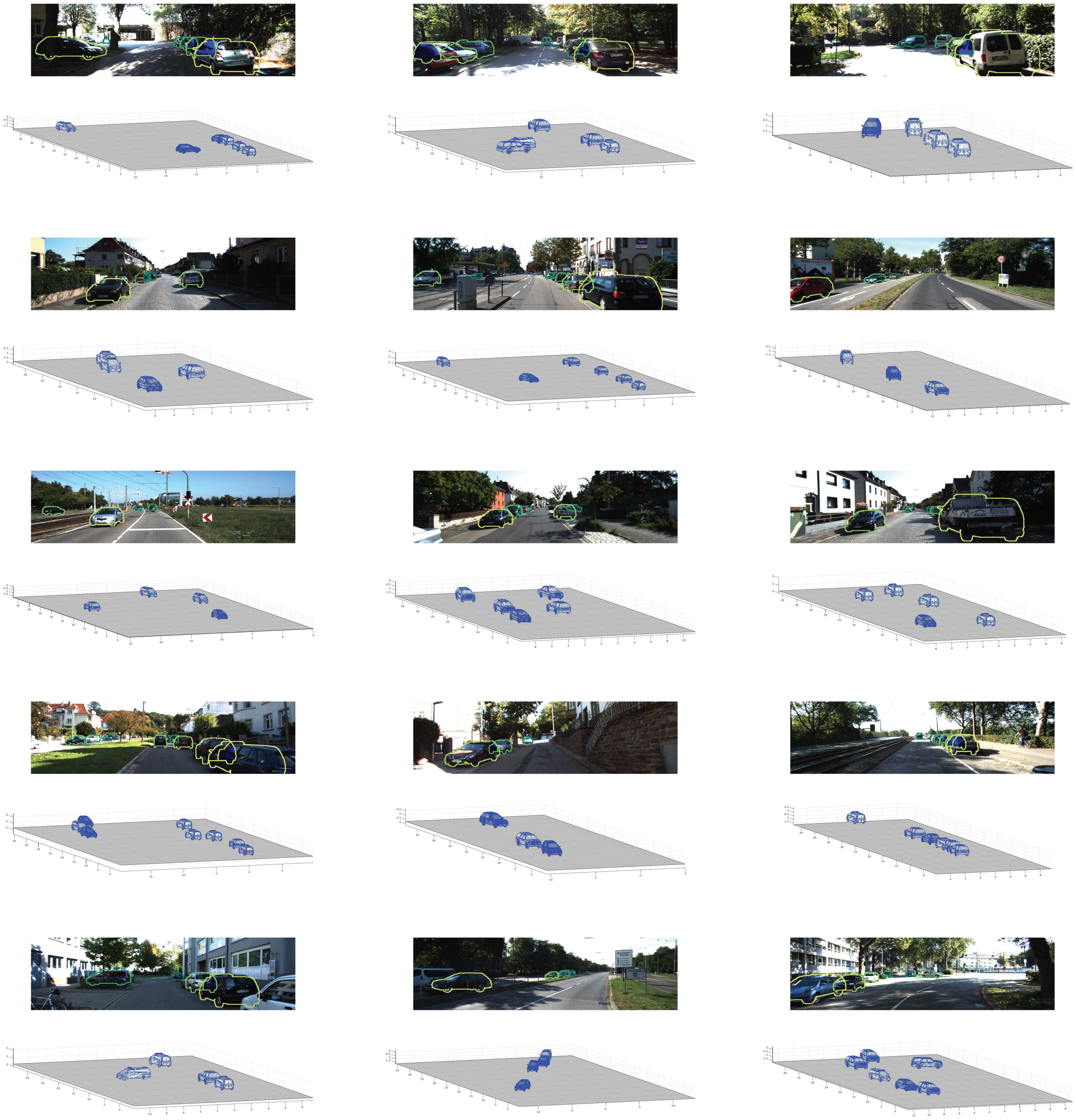

Figure 10. 2D recognition and 3D localization results on the KITTI test set. Detections at 1 false positive per image (fppi) are shown. Blue regions in the images are the estimated occluded areas.

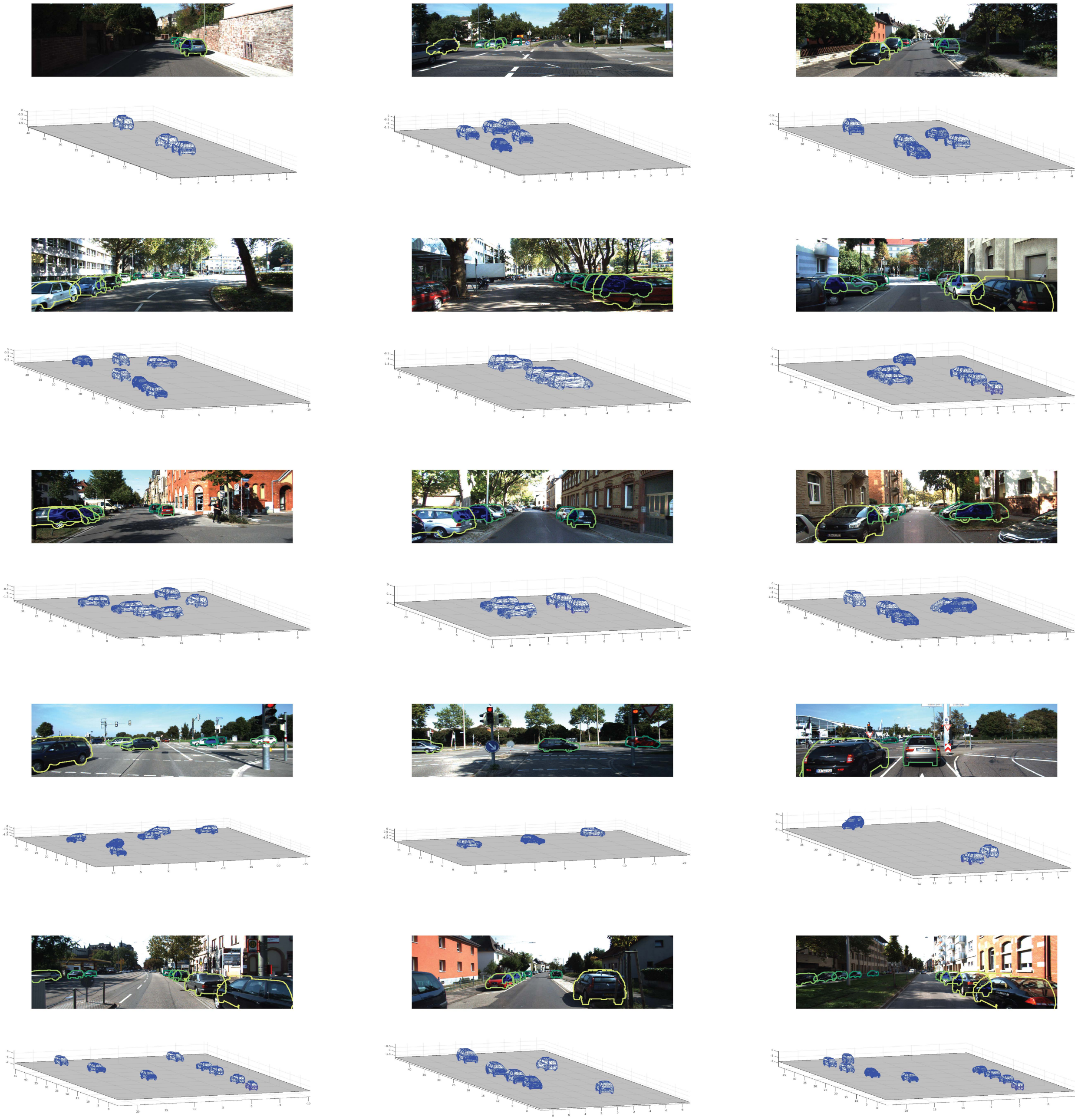

Figure 11. 2D recognition and 3D localization results on the KITTI test set. Detections at 1 false positive per image (fppi) are shown. Blue regions in the images are the estimated occluded areas.

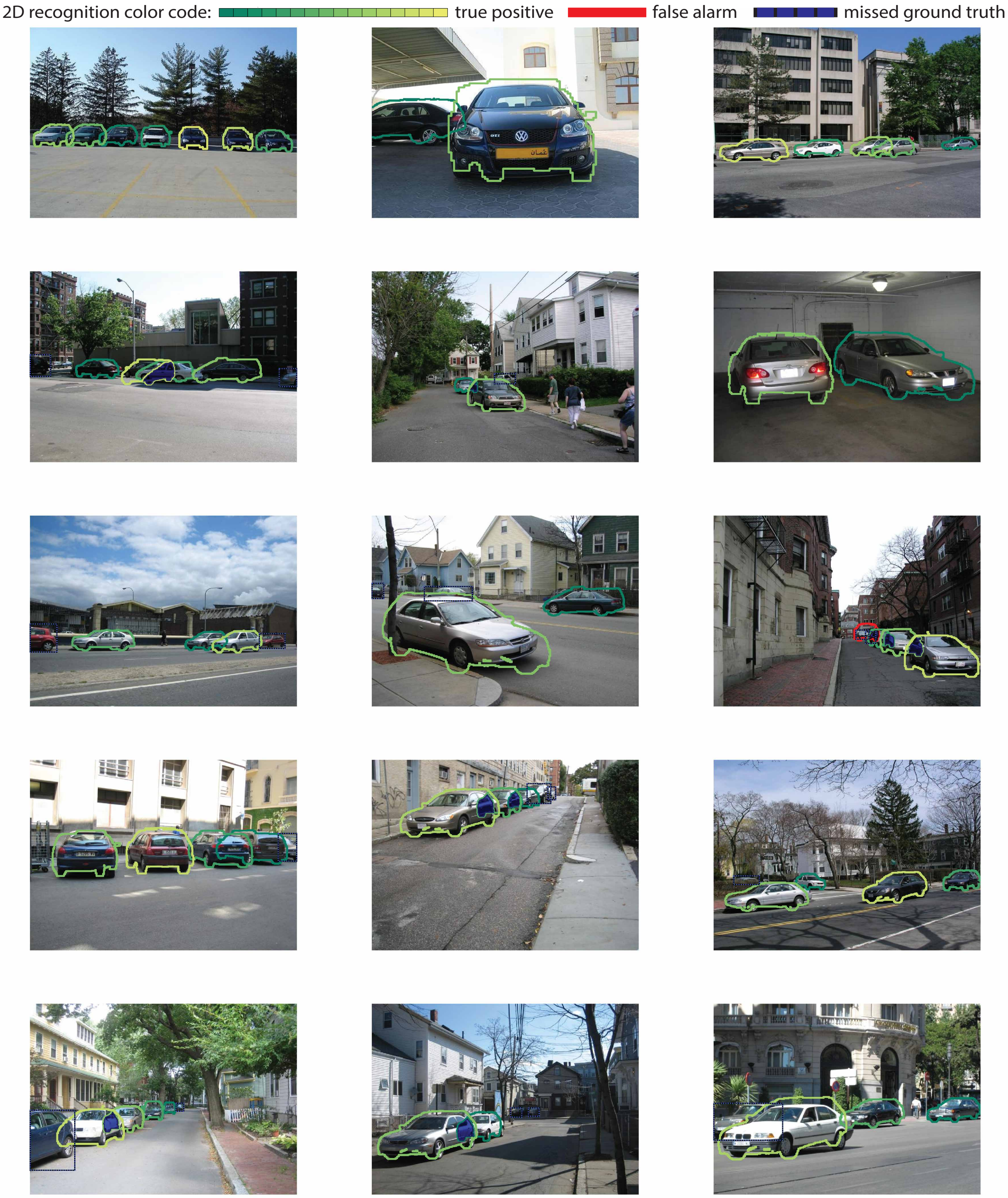

Figure 12. 2D recognition results on the OutdoorScene dataset. Detections at 1 false positive per image (fppi) are shown. Blue regions in the images are the estimated occluded areas.

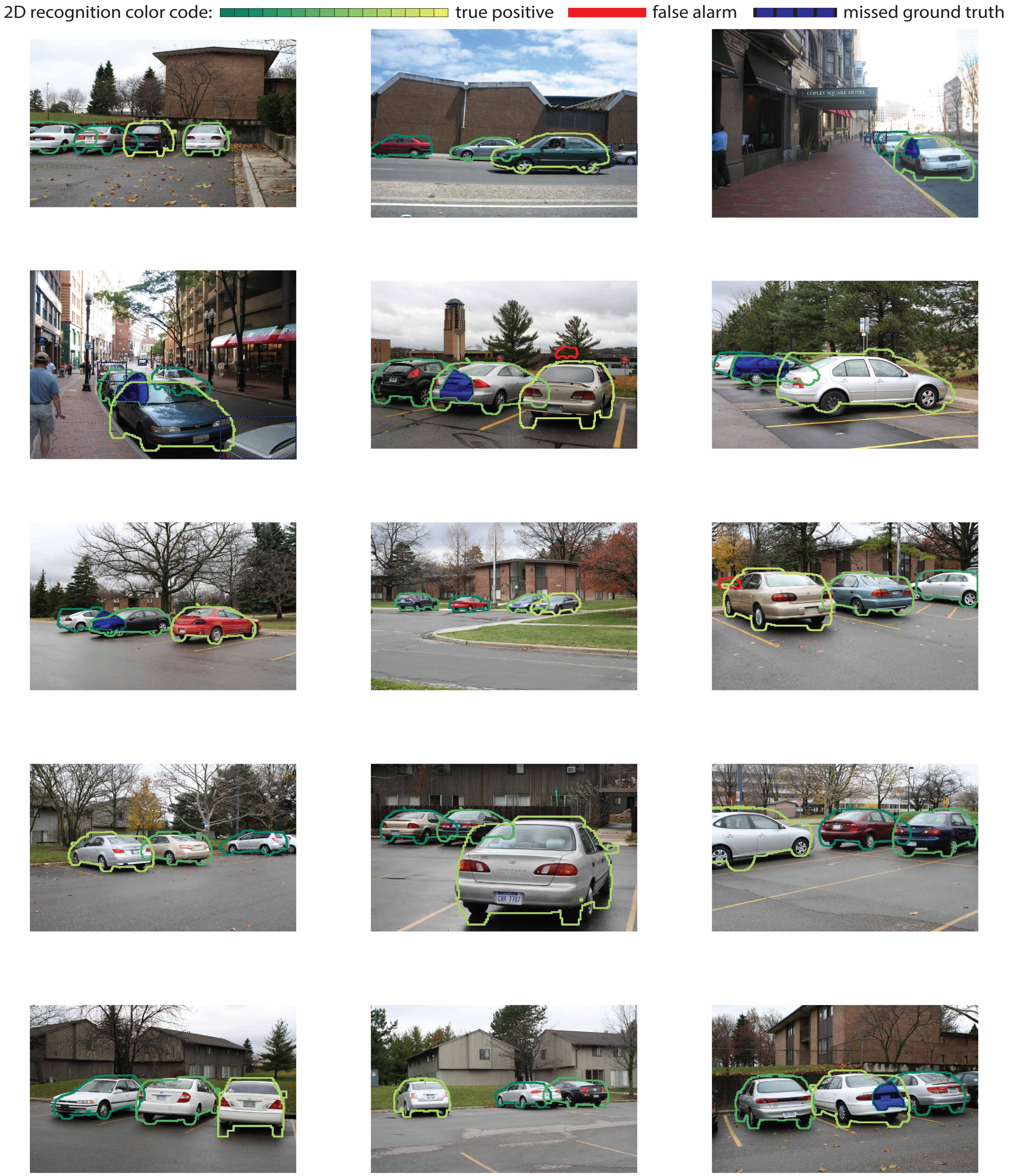

Figure 13. 2D recognition results on the OutdoorScene dataset. Detections at 1 false positive per image (fppi) are shown. Blue regions in the images are the estimated occluded areas.