3D Reconstruction Meets Semantics – Reconstruction Challenge 2017

Torsten Sattler, Radim Tylecek, Thomas Brox, Marc Pollefeys, Robert B. Fisher

Abstract

Part of the workshop is a challenge on combining 3D and semantic information in complex scenes. To this end, a challenging outdoor dataset, captured by a robot driving through a semantically-rich garden that contains fine geometric details, was released. A multi-camera rig is mounted on top of the robot, enabling the use of both stereo and motion stereo information. Precise ground truth for the 3D structure of the garden has been obtained with a laser scanner and accurate pose estimates for the robot are available as well. Ground truth semantic labels and ground truth depth from a laser scan are used for benchmarking the quality of the 3D reconstructions.

1. Description

Given a set of images and their known camera poses, the goal of the challenge is to create a semantically annotated 3D model of the scene. To this end, it will be necessary to compute depth maps for the images and then fuse them together (potentially while incorporating information from the semantics) into a single 3D model.

We have provided the following data for the challenge[1]:

- A set of **training** sequences consisting of
  - calibrated images with their camera poses,
  - ground truth semantic annotations for a subset of these images,
  - a semantically annotated 3D point cloud depicting the area of the training sequence.
- A **testing** sequence consisting of calibrated images with their camera poses.

Both training and testing data are available from the git repository `https://gitlab.inf.ed.ac.uk/3DRMS/Challenge2017`, where also details on the file formats can be found.

[1] `http://trimbot2020.webhosting.rug.nl/events/3drms/challenge`

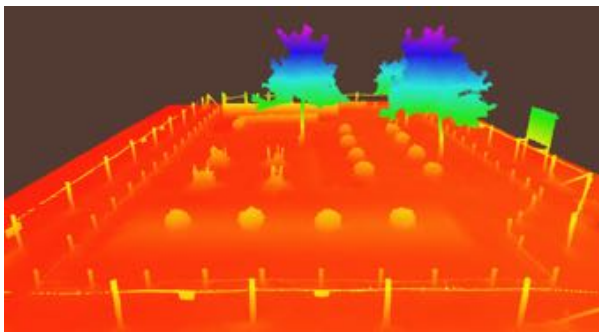

Figure 1. Point cloud of the entire garden (height-colored).

2. Garden Dataset

The dataset for the the 3DRMS challenge was collected in a test garden at Wageningen University Research Campus, Netherlands, which was built specifically for experimentation in robotic gardening.

Four scenarios of robot driving around different parts of the garden (Fig. 2) were used: `around_hedge (17)`, `boxwood_row (57)`, `boxwood_slope (23)` and `around_garden (124)`. The first three were designated entirely for training, the last one was split between testing and training (`around_garden_roses (11)`). The numbers in brackets indicate the sequence length in frames.

2.1. Calibrated Images

Image streams from four cameras (0,1,2,3) were provided. Fig. 3 shows these are mounted in a pairwise setup, the pair 0-1 is oriented to the front and the pair 2-3 to the right side of the robot vehicle. Resolution of the images is 752x480 (WVGA), cameras 0 and 2 are color while cameras 1 and 3 are greyscale (but sharper). All images were undistorted with the intrinsic camera parameters[2].

The camera poses were estimated with COLMAP [3] and manually aligned to the coordinate system of the laser point cloud.

[2] Calibration was performed with Kalibr toolbox, `https://github.com/ethz-asl/kalibr`.

Figure 2. Trajectories of the captured scenarios, training (yellow) and test (purple) sequences.

2.2. Semantic Image Annotations

The set of classes we distinguish in the images contains 9 labels (color code in brackets):

- Grass (light green)
- Ground (brown)
- Pavement (grey)
- Hedge (ochre)
- Topiary (cyan)
- Rose (red)
- Obstacle (blue)
- Tree (dark green)
- Background (black)

Manual pixel-wise annotations (Fig. 4) are provided for frames in cameras 0 and 2.

2.3. Semantic Point Cloud

The geometry of the scene was acquired by *Leica ScanStation P15*, accuracy of 3 mm at 40 m. Its native output merged from 20 individual scans (Fig. 1) was subsampled with a spatial filter to achieve a minimal distance between two points of 10 mm, which becomes the effective accuracy of the GT. For some dynamic parts, like leaves and branches, the accuracy can be further reduced due to movement by the wind, etc.

Semantic labels were assigned to the points with multiple 3D bounding boxes drawn around individual components of the point cloud belonging to the garden objects or terrain. Ultimately the point cloud was split into segments corresponding to train and test sequences as shown in Fig. 5.

3. Evaluation

We have evaluated the quality of the 3D meshes based on the *completeness* of the reconstruction, i.e., how much of the ground truth is covered, the *accuracy* of the reconstruction, i.e., how accurately the 3D mesh models the scene, and the *semantic quality* of the mesh, i.e., how close the semantics of the mesh are to the ground truth.

3.1. Compared Methods

In addition to the two submitted results we have also compared to current state-of-the-art methods in both reconstruction [3] and classification [1] tasks.

SnapNet-R (Moras) [2] A variant of the SnapNet deep net, with better semantic segmentation of 2D & 3D data. New views are synthesized from the 3D point cloud, which can be used by 2D semantic labeling and segmentation, thus boosting performance from the fusion of the resulting multiple labels. The fact that pixels are semantically labeled is used to constrain false correspondences when constructing 3D point clouds.

Taguchi [6] Semantic 3D reconstruction using depth and label fusion.

COLMAP [3] (3D Reconstruction baseline) A general-purpose Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipeline with a graphical and command-line interface. It offers a wide range of features for reconstruction of ordered and unordered image collections.

SegNet [1] (Semantic baseline) For comparison with the 2D state-of-the-art a SegNet architecture [1] is adapted for the given garden semantics. The network was trained on 20k synthetic garden images and then fine-tuned with the challenge training set.

3.2. 3D Geometry Reconstruction: Accuracy & Completeness

We have followed the usual evaluation methodology described in [5]. In particular, *accuracy* is distance d (in m) such that 90% of the reconstruction is within d of the ground truth mesh and *completeness* is the percent of points in the GT point cloud that are within 5 cm of the reconstruction.

The distances between the reconstruction and GT are calculated using a point-to-mesh metric for completeness and vertex-to-point for accuracy. The faces of submitted meshes were subdivided to have a same maximum edge length. The

difference between the evaluated results are shown in Figures 6,7,8, which all use the same color scale for accuracy or completeness. Cold colors indicate well reconstructed segments while hot colors indicate hallucinated surface (accuracy) or missing parts (completeness).

The evaluation was limited to the space delimited in XY by the perimeter of the test area and in Z to 1 m high section above the ground, ie. the tree-tops were excluded. Following [4] we also plot cumulative histograms of distances in Fig. 9.

3.3. Semantic Classification: Quality

The accuracy of semantic labels assigned to vertices or faces of the 3D model (Fig. 14) was evaluated by its projection to all test images with known poses. Only the pixels corresponding to the 3D test part (as specified in the previous section) were considered. The rest of the image was masked out and ignored, as in Fig. 10.

Visual comparison of the results in a selected frame is given in Figures 11,12,13. In the error mask the red pixels indicate incorrectly classified pixels, grey were correct and black were not evaluated. Quantitative results are presented by confusion matrices for all images in the test set in Fig. 15, where semantic accuracy is the percentage of correctly predicted pixels across all test images.

3.4. Results

The quantitative comparison in all three performance categories is given in the following table:

Method	*Accuracy*	*Completeness*	*Semantic*
Taguchi [6]	**0.101 m**	71.1 %	**82.2 %**
SnapNet-R [2]	0.198 m	**83.3 %**	69.3 %
Colmap [3]	*0.022 m*	*85.3 %*	
SegNet [1]			*82.2 %*

Table 1. Comparison of submitted results (top rows) with baselines (bottom). Semantic quality is the ratio of correctly predicted pixels in the test part of images.

The baseline Structure-from-Motion method [3] outperformed the challenge participants by a large margin in accuracy while obtaining similar completeness of SnapNet-R [2]. The SnapNet-R method of Moras et al. achieves that completeness level only at the cost of significantly lower accuracy, corresponding to the large amount of hallucinated surfaces visible in the reconstruction result.

Compared to that the method of Taguchi and Feng [6] appears to be rather conservative, with less complete but semantically more consistent mesh, as observed in Fig. 14, resulting in the same performance the as deep convolutional network [1].

4. Conclusion

In summary the workshop challenge competitors did not fully leverage the joint semantic & 3D information, as both independent 2D and 3D baseline methods we compared with performed the same or better in quantitative terms.

Acknowledgements

The garden dataset was prepared within EU project TrimBot2020. The semantic baseline results were provided by Hoang-An Le (University of Amsterdam).

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2, 3

[2] Joris Guerry, Alexandre Boulch, Bertrand Le Saux, Julien Moras, Aurelien Plyer, and David Filliat. SnapNet-R: consistent 3D multi-view semantic labeling for robotics. In *Proc. ICCV Workshops (3DRMS)*, 2017. 2, 3

[3] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. CVPR*, 2016. 1, 2, 3

[4] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proc. CVPR*, 2017. 3

[5] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. CVPR*, CVPR '06, pages 519–528, Washington, DC, USA, 2006. IEEE Computer Society. 2

[6] Yuichi Taguchi and Chen Feng. Semantic 3D reconstruction using depth and label fusion. In *3DRMS Workshop Challenge, ICCV*, 2017. 2, 3

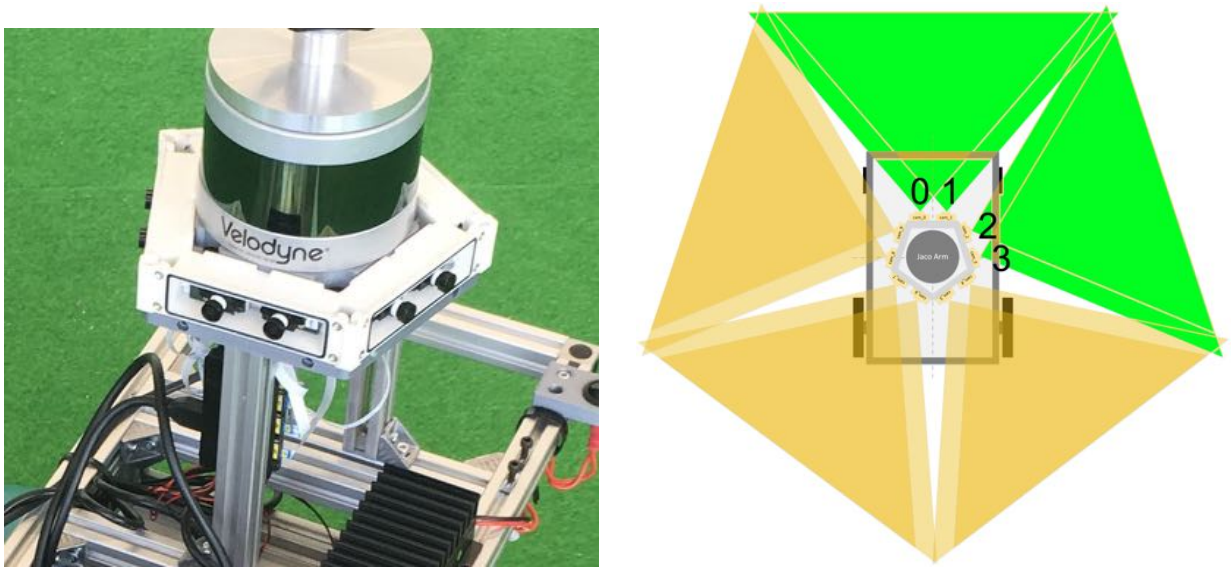

Figure 3. Pentagonal camera rig mounted on the robot (left). First four cameras were included in the challenge data (right, green).

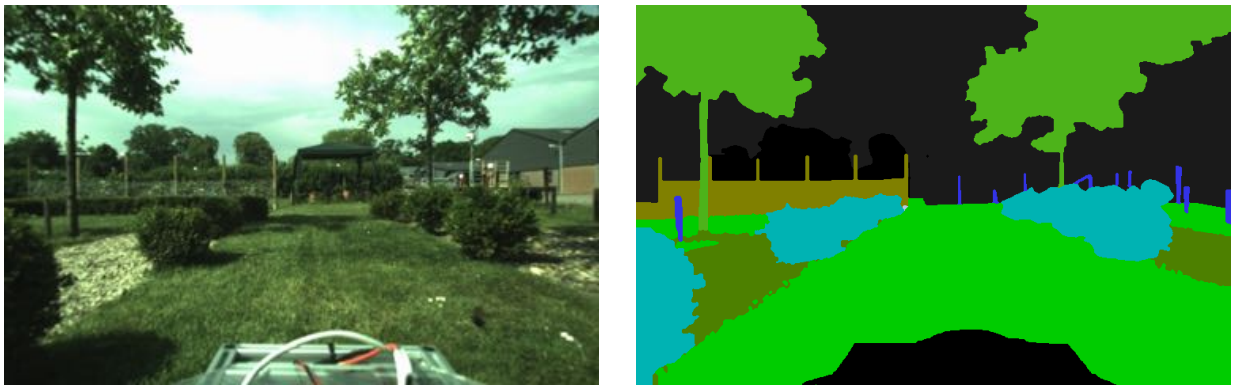

Figure 4. Undistorted image from camera 0 (left) and its semantic annotation (right).

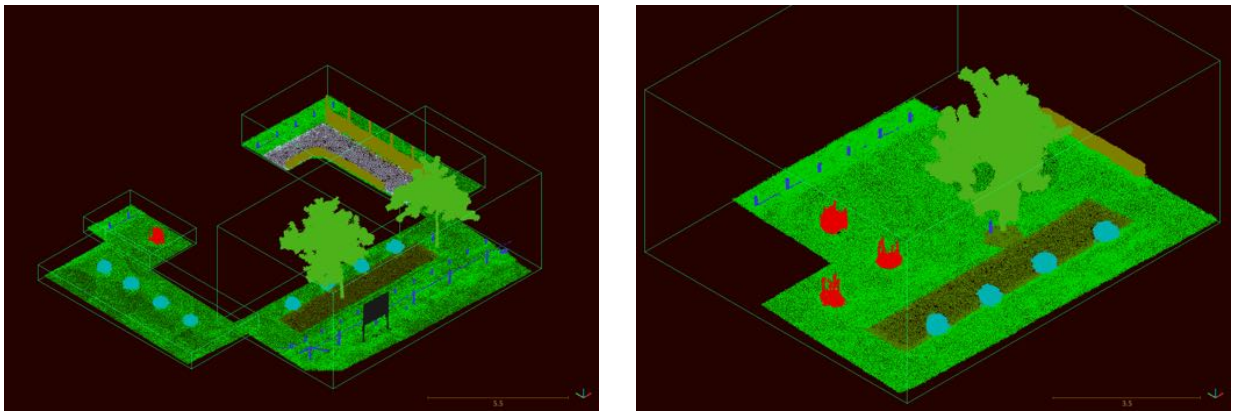

Figure 5. Semantic point cloud of the entire garden with color-coded object classes. Left: 4 training parts. Right: test part.

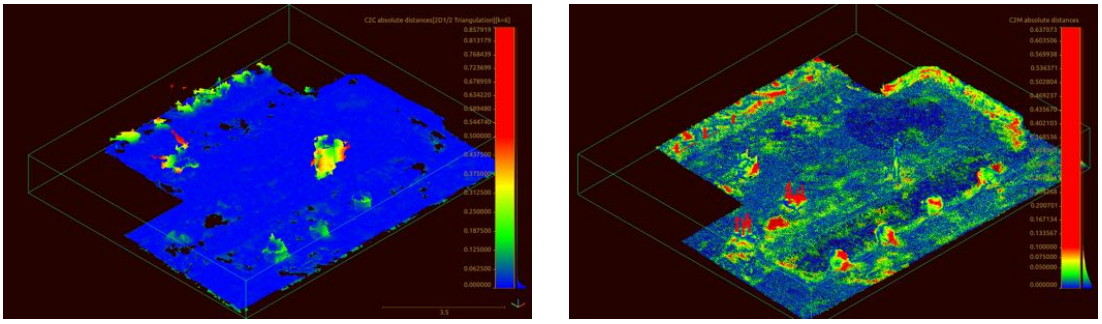

Figure 6. Accuracy (left) and completeness (right) of Taguchi's reconstruction.

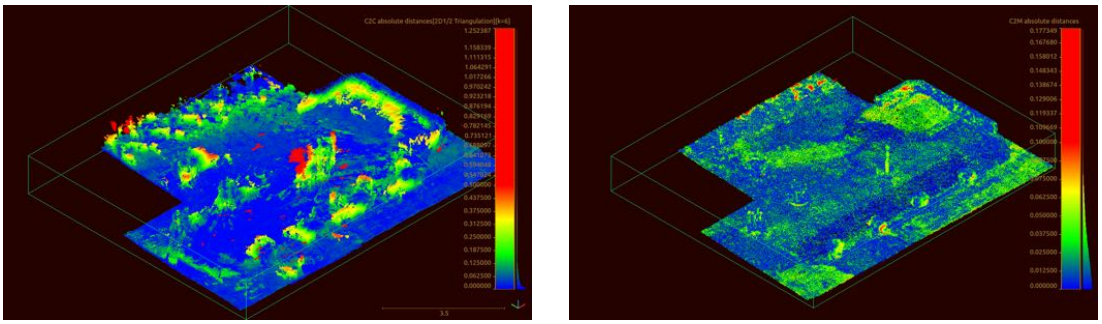

Figure 7. Accuracy (left) and completeness (right) of SnapNet-R reconstruction.

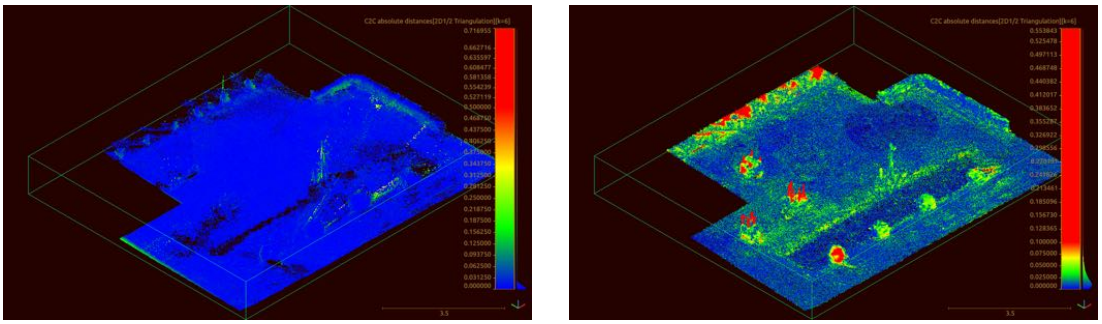

Figure 8. Accuracy (left) and completeness (right) of COLMAP reconstruction.

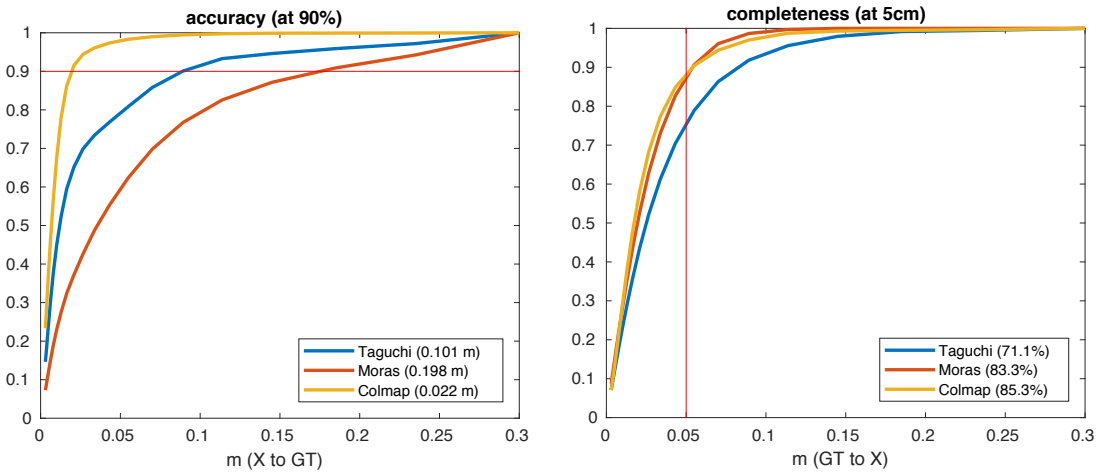

Figure 9. Accuracy and completeness of evaluated 3D reconstructions.

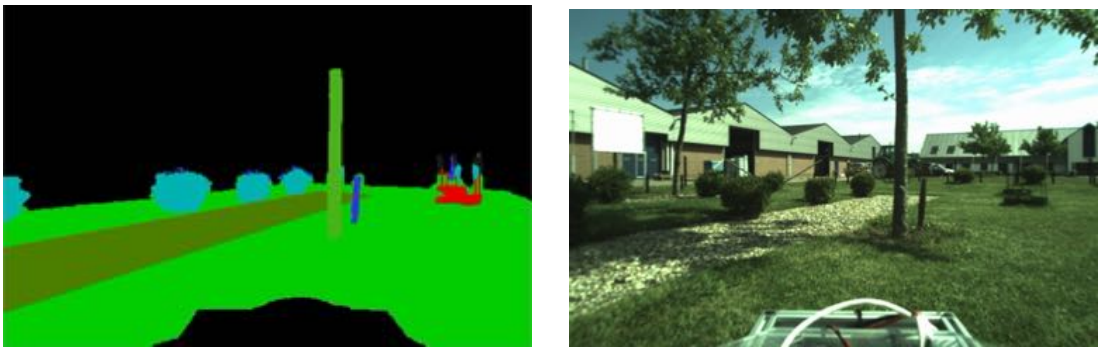

Figure 10. Masked GT annotation (left) of a sample test frame (right).

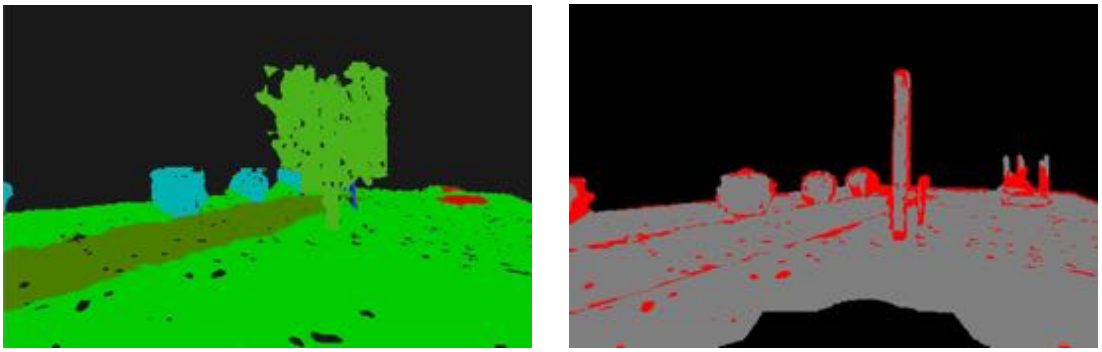

Figure 11. Projection of Taguchi's reconstruction (left) and its error mask (right).

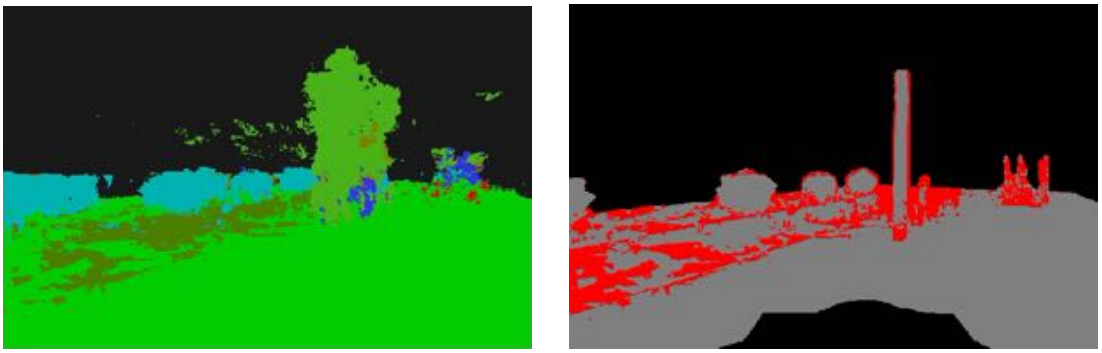

Figure 12. Projection of SnapNet-R reconstruction (left) and its error mask (right).

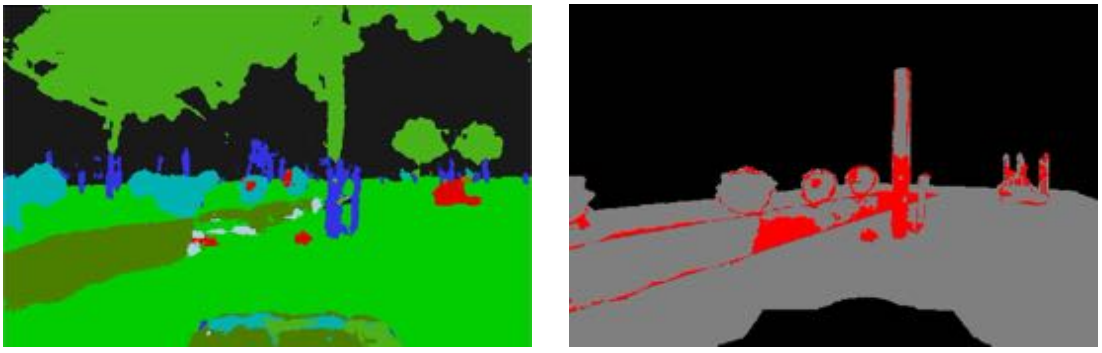

Figure 13. SegNet classification (left) and its error mask (right).

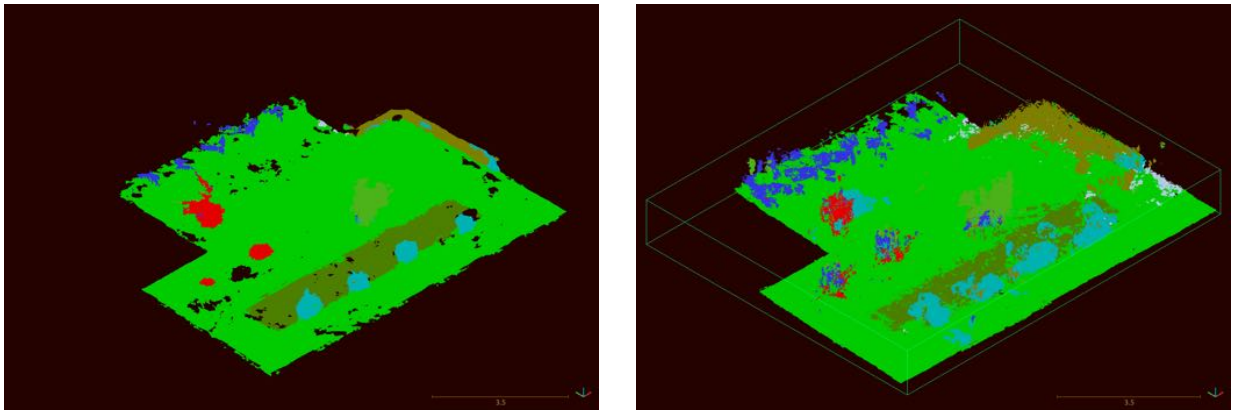

Figure 14. Semantic mesh of Taguchi (left) and SnapNet-R (right).

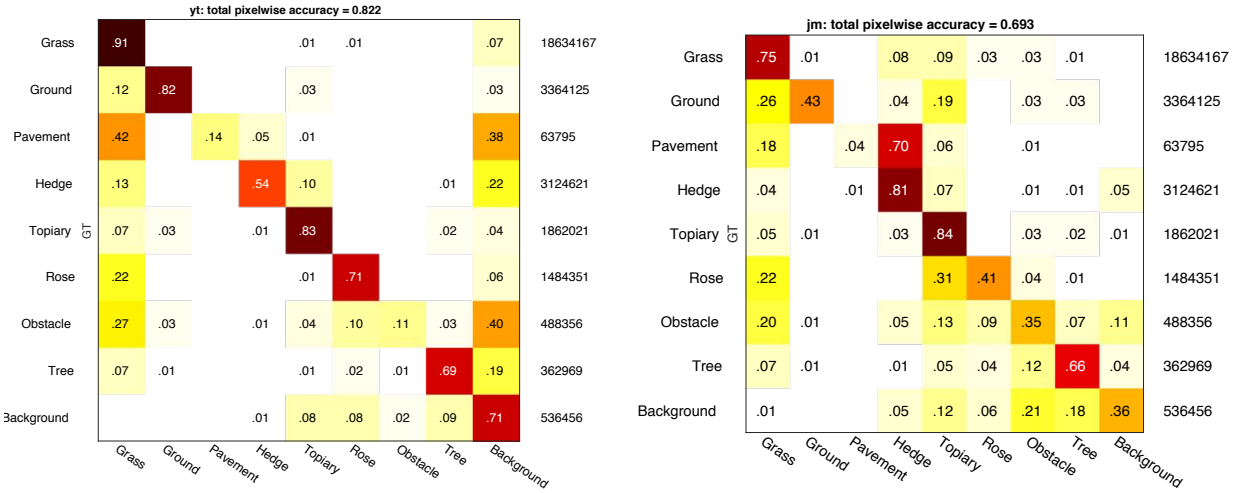

Figure 15. Confusion matrices for submissions of Taguchi (left) and SnapNet-R (right).

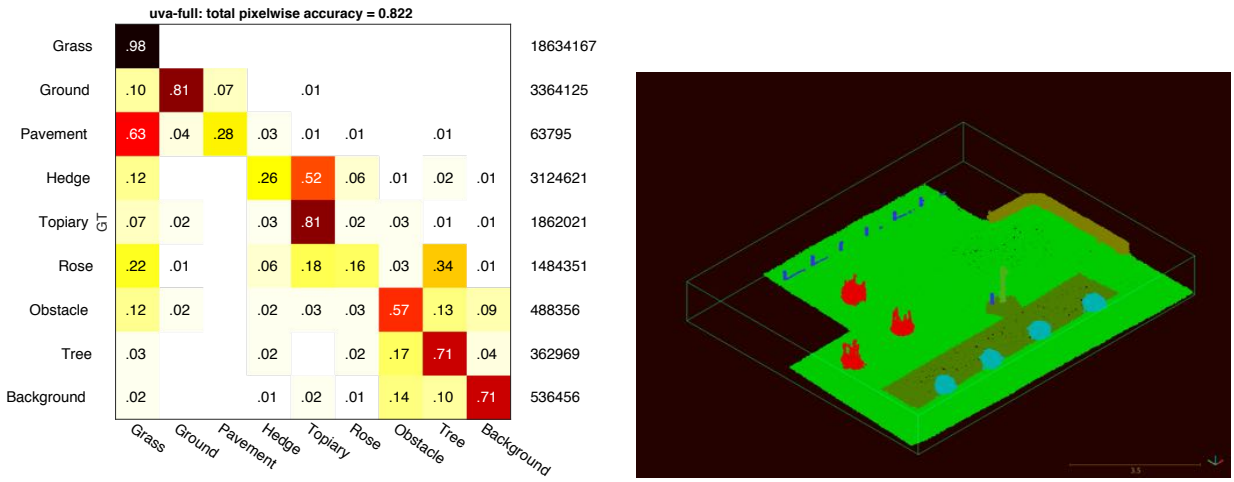

Figure 16. Confusion matrix for SegNet (left) and the 1m high test section of GT model (right).