Supplementary Material
A Probabilistic Framework for Real-time 3D Segmentation using Spatial, Temporal, and Semantic Cues

David Held, Devin Guillory, Brice Rebsamen, Sebastian Thrun, Silvio Savarese
Computer Science Department, Stanford University

I. Evaluation metric

To evaluate our segmentation, we assign a best-matching segment to each ground-truth bounding box. For each ground-truth bounding box gt, we find the set of non-ground points within this box, C_{gt}. For each segment s, let C_s be the set of points that belong to this segment. We then find the best-matching segment to this ground-truth bounding box by computing

$$s = \arg\max_{s'} |C_{s'} \cap C_{gt}| \tag{1}$$

The best-matching segment is then assigned to this ground-truth bounding box for the evaluation metric described in our paper, as well as for the metric described below.

Some previous works have evaluated 3D segmentation using the intersection-over-union metric on 3D points [5]. Note that our method segments the entire scene, as opposed to the method of Wang et al. [5], so the evaluation metric from Wang et al. [5] does not directly apply. However, we could modify the intersection-over-union metric [5] as follows: we can compute the fraction of ground-truth bounding boxes which have an intersection-over-union score less than a threshold τ_{IOU}, as

$$E_{IOU} = \frac{1}{N} \sum_{gt} \mathbb{1}\Big(\frac{|C_s \cap C_{gt}|}{|C_s \cup C_{gt}|} < \tau_{IOU}\Big), \tag{2}$$

where $\mathbb{1}$ is an indicator function that is equal to 1 if the input is true and 0 otherwise. In our experiments, we set $\tau_{IOU} = 0.5$.

However, we find that the intersection-over-union evaluation metric [5] is non-ideal for autonomous driving because this score penalizes undersegmentation errors more than oversegmentation errors. For example, suppose a person is undersegmented together with a large building; the intersection-over-union score will be extremely low. On the other hand, suppose that the person is instead oversegmented into two pieces. The intersection-over-union score for the larger segment will often still be above our threshold and thus this oversegmentation will not be penalized. In practice, we find that optimizing our hyperparameters for the intersection-over-union metric causes the number of undersegmentation errors to decrease while increasing the number of oversegmentation errors.

Regardless, if we use the intersection-over-union metric of Equation 2, we get 9% segmentation errors, compared to the best baseline performance of 16% [4]. When analyzing segments within 15 m of the ego-vehicle, the improvement is even more dramatic: we get 6% segmentation errors, compared to a baseline of 17% [4], thus reducing the absolute number of errors by 62%. Nearby objects are most important for autonomous driving because these are the objects that will contribute the most to immediate path-planning decisions. Thus, even though this metric is non-ideal, we still show that our approach can outperform the baseline methods when evaluated using this metric.

II. Parameters

As mentioned in the main text, we evaluate our segmentation method on the KITTI tracking dataset [1, 2, 3]. This dataset consists of a total of 21 sequences. We use sequences 0001 and 0013 to train our method and select parameters and the remaining 19 sequences for testing and evaluation.

We choose the parameters for our method using a grid-search on the training set, and the resulting parameter values are listed in Table I. The parameter $p(N_t)$ is the prior probability that a set of points obtained from our initial segmentation belongs to just one object. The parameter $p(z_t^p | a_t = \emptyset)$ is the probability of observing a set of points at a given position given that these points don't match to any previous segment. The parameter $p_1(z_t | N_t, a_t, z_{1\ldots t-1})$ is the probability of observing segment s_t as a single segment from our initial segmentation, given that s_t is only one object.

The parameter $p(N_t | a_t, N_{t-1})$ is the probability that a segment from the previous frame that represents a single object still represents a single object in the current frame. The number of objects can change if, for example, a person dismounts a bicycle or exits a car or if there was a previous undersegmentation error. The parameter $p(\neg N_t | a_t, \neg N_{t-1})$ is the probability that more than one segment from the previous frame represents more than one object in the current frame. The number of objects can change if a person mounts a bicycle or enters a car or due to a previous oversegmentation error.

The parameter τ_s is a threshold that we use for temporal splitting, i.e. we try to perform temporal splitting with any segment s_{t-1} for which $p(z_t|a_t, z_{1...t-1}) > \tau_s$. The parameter t_0 is the frame number to begin using temporal and semantic information. Prior to that, our classification and velocity estimates are assumed to be too inaccurate to use as cues for segmentation. The parameter μ_S is used by our shape probability distribution when the class of the object is unknown, and $\mu_{S,c}$ is used when the class is known. The parameter μ_D is also used by our shape probability distribution and was computed by fitting a distribution over distances between pairs of objects in our training set (as opposed to the other parameters which were chosen using cross-validation on our training set).

The parameters $\mu_{i,c}$, $\sigma_{i,c}$, k_1, and k_2 are used in equations 18 and 19 for the volumetric shape distribution. The parameter k_3 is used for spatial splitting, ensuring that all points in each segment are at least $k_3\, r$ from all points in a neighboring segment, based on the sensor resolution r.

The parameter n_{min} is the minimum number of points for a segment to be created using temporal splitting. The parameter T is the length of the past history that we use for the recursive computation from equation 6 (recall that we perform this computation as-needed in a lazy manner). The 2 sequences that were used in choosing these parameters are distinct from the 19 sequences that were used for testing.

Parameter	Value
$p(N_t)$	0.99
$p(z_t^p \| a_t = \emptyset)$	0.05
$p_1(z_t \| N_t, a_t, z_{1...t-1})$	0.6
$p(N_t \| a_t, N_{t-1})$	0.6
$p(\neg N_t \| a_t, \neg N_{t-1})$	0.999
τ_s	0.5
t_0	2
μ_S	0.03 m
$\mu_{S,c}$, c = car	0.15 m
μ_D	7.4 m
$\mu_{i,c}$, i = length, c = person	0.6 m
$\sigma_{i,c}$, i = length, c = person	0.4 m
$\mu_{i,c}$, i = width, c = person	0.4 m
$\sigma_{i,c}$, i = width, c = person	0.1 m
$\mu_{i,c}$, i = length, c = bike	1.5 m
$\sigma_{i,c}$, i = length, c = bike	0.5 m
$\mu_{i,c}$, i = width, c = bike	1.5 m
$\sigma_{i,c}$, i = width, c = bike	1 m
$\mu_{i,c}$, i = length, c = car	3 m
$\sigma_{i,c}$, i = length, c = car	2.5 m
$\mu_{i,c}$, i = width, c = car	2 m
$\sigma_{i,c}$, i = width, c = car	0.1 m
k_1	2
k_2	10
k_3	4
n_{min}	10
T	10

TABLE I
PARAMETER VALUES, CHOSEN USING OUR TRAINING SET.

REFERENCES

[1] Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2013.

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[4] Alex Teichman, Jesse Levinson, and Sebastian Thrun. Towards 3d object recognition via classification of arbitrary object tracks. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 4034–4041. IEEE, 2011.

[5] Dominic Zeng Wang, Ingmar Posner, and Paul Newman. What could move? finding cars, pedestrians and bicyclists in 3d laser data. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 4038–4044. IEEE, 2012.