Monocular Visual Object 3D Localization in Road Scenes

Yizhou Wang[1], Yen-Ting Huang[2], Jenq-Neng Hwang[1]
[1] Information Processing Lab, University of Washington, Seattle, WA
[2] Pervasive AI Research Labs, National ChengChi University, Taipei, Taiwan

INTRODUCTION

Problems to Address:

- Accurately and robustly 3D localize objects corresponding to camera.
- Only use a monocular camera setup on an autonomous vehicle.

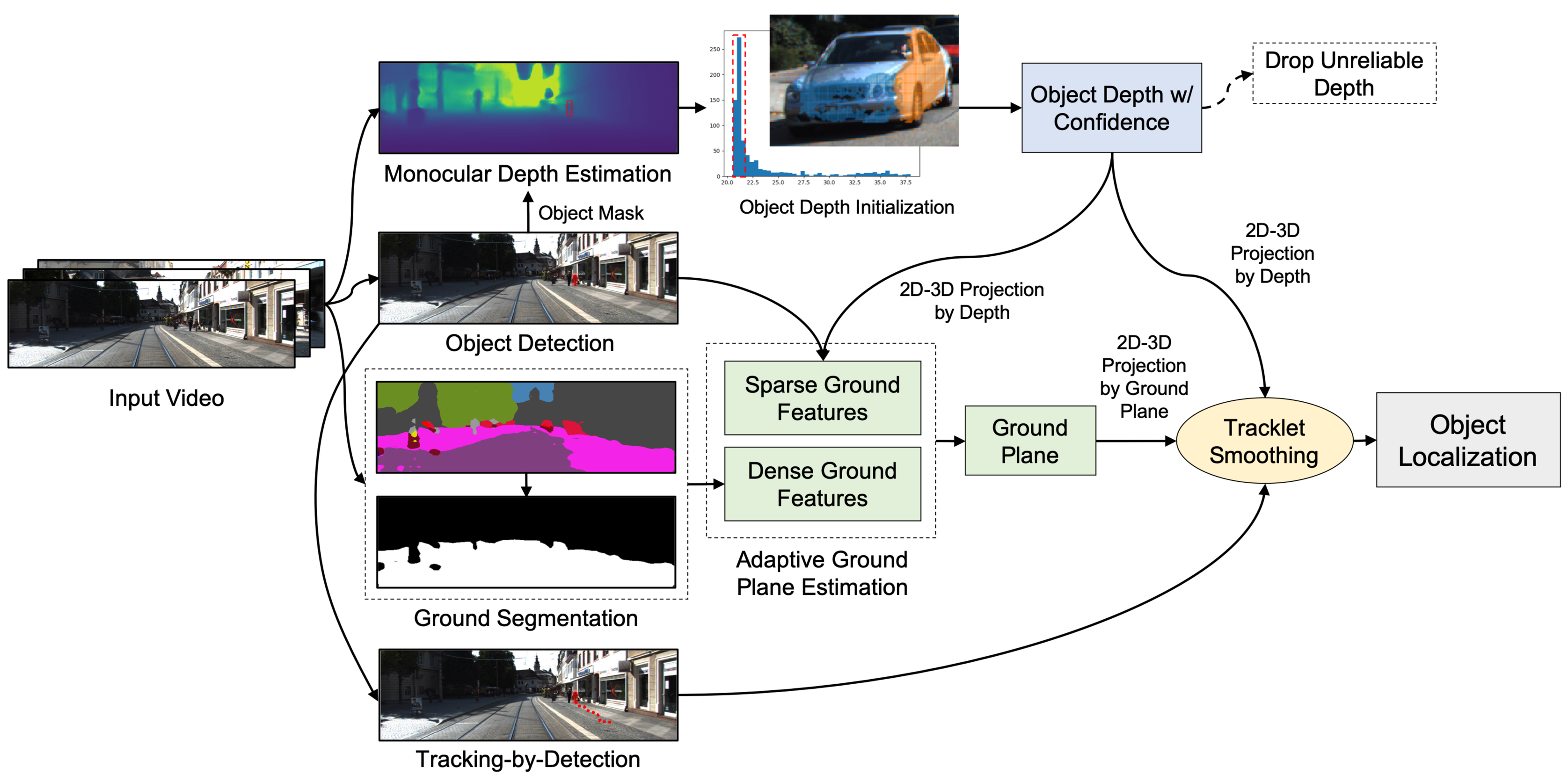
Challenges:

- Obtaining 3D information is ill-posed for monocular cameras.
- Objects are usually occluded in the autonomous driving view.
- Lack of general 3D localization methods which is applicable for different kinds of objects, e.g., cars, pedestrians, cyclists, etc.

Contributions:

- A accurate and robust monocular object 3D localization framework.
- Generalized: Applicable for common moving objects in road scenes.
- Competitive: Input depthmap can be replaced by other equivalent depth sensors, e.g., LiDAR, depth camera and RADAR.

PROPOSED SYSTEM



Object Depth Initialization:

- Depth estimation + instance segmentation + depth histogram analysis.

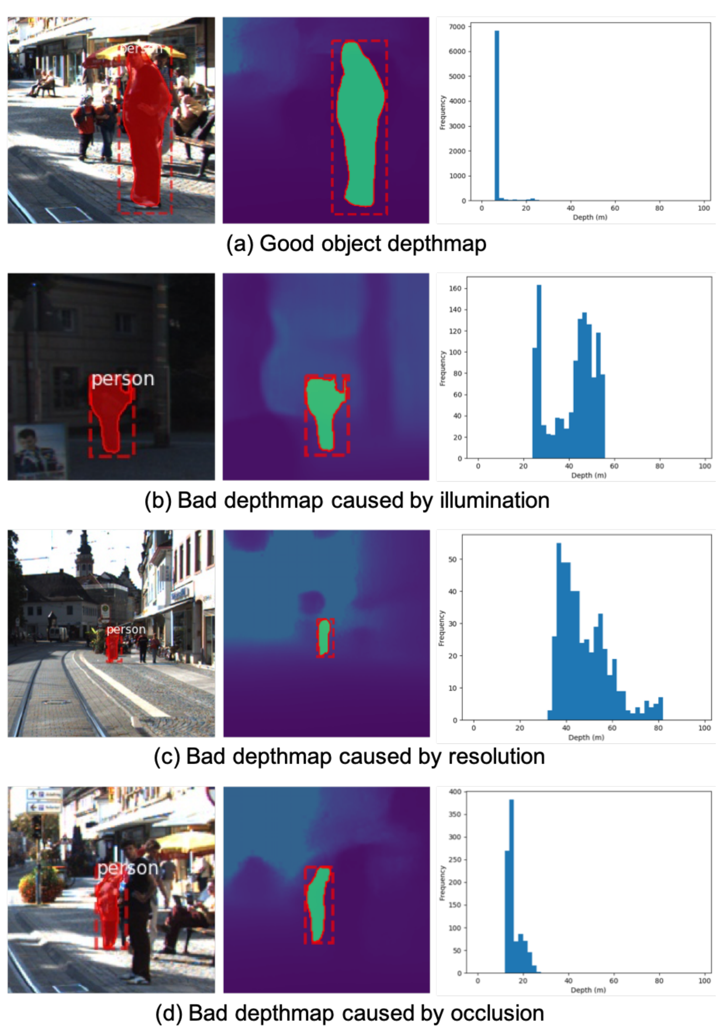
Adaptive ground plane estimation:

- Sparse & dense ground features + augmented RANSAC.

Object tracklet smoothing:

- Multi-object tracking + moving split + weighted Huber regression.

Object Depth Initialization

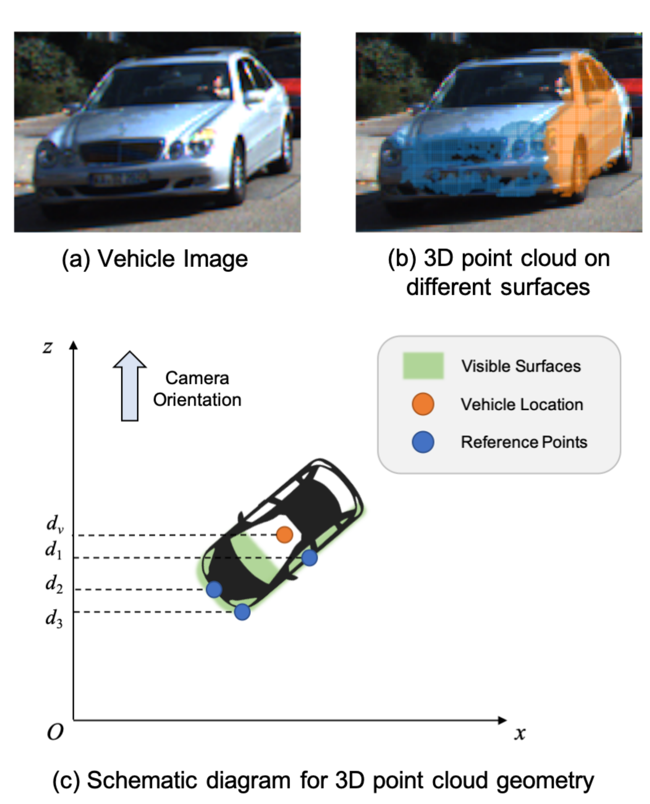


(a) Good object depthmap

(b) Bad depthmap caused by illumination

(c) Bad depthmap caused by resolution

(d) Bad depthmap caused by occlusion



(a) Vehicle Image

(b) 3D point cloud on different surfaces

(c) Schematic diagram for 3D point cloud geometry

- Monocular depth estimation [1] to get a dense depthmap.
- Use Mask R-CNN [2] to get object masks.
- Object depth – proposal bins in depth histogram $\mathcal{H}$: $d_{obj} = \frac{1}{|PB|} \sum_{d_i \in PB} d_i$.
- Object depth confidence: $c_{obj} = \left(1 - \frac{\sigma}{\mu}\right) \cdot \frac{|PB|}{|\mathcal{H}|}$.

OR

- Separate point cloud into two surfaces by SLIC [3].
- Calculate vehicle depth: $d_v = d_1 + (d_2 - d_3)$.

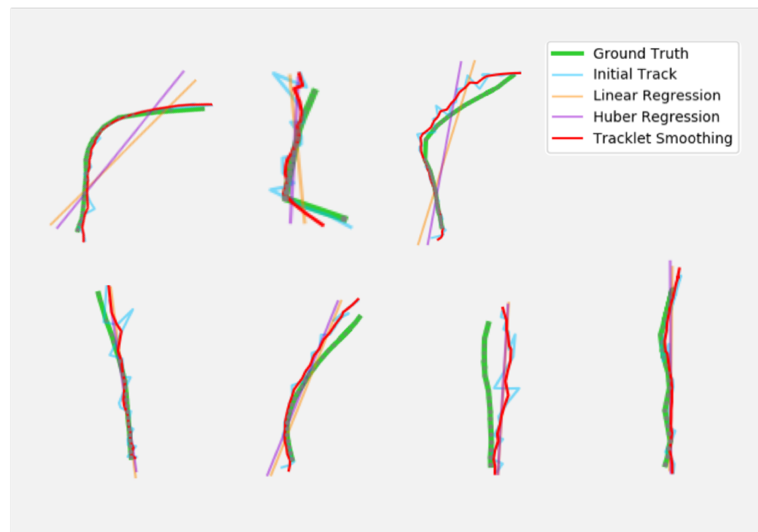
Adaptive Ground Plane Estimation

- Dense Features: 3D points from semantic segmentation [4].
- Sparse Ground Features: Object 3D bottom-center points.
- Augmented RANSAC to jointly consider the contributions from dense and sparse ground features.

Object Tracklet Smoothing

- Multi-object tracking [5] to obtain association among bounding boxes.
- Generate object 3D trajectories.
- Moving split the trajectories into short tracklets and apply weighted Huber regression to each tracklet.

$$\rho(r_i) = \begin{cases} c_i \cdot r_i^2 & \text{if } |r_i| \le \eta, \\ c_i \cdot \eta(2|r_i| - \eta) & \text{if } |r_i| > \eta. \end{cases}$$



EXPERIMENTS

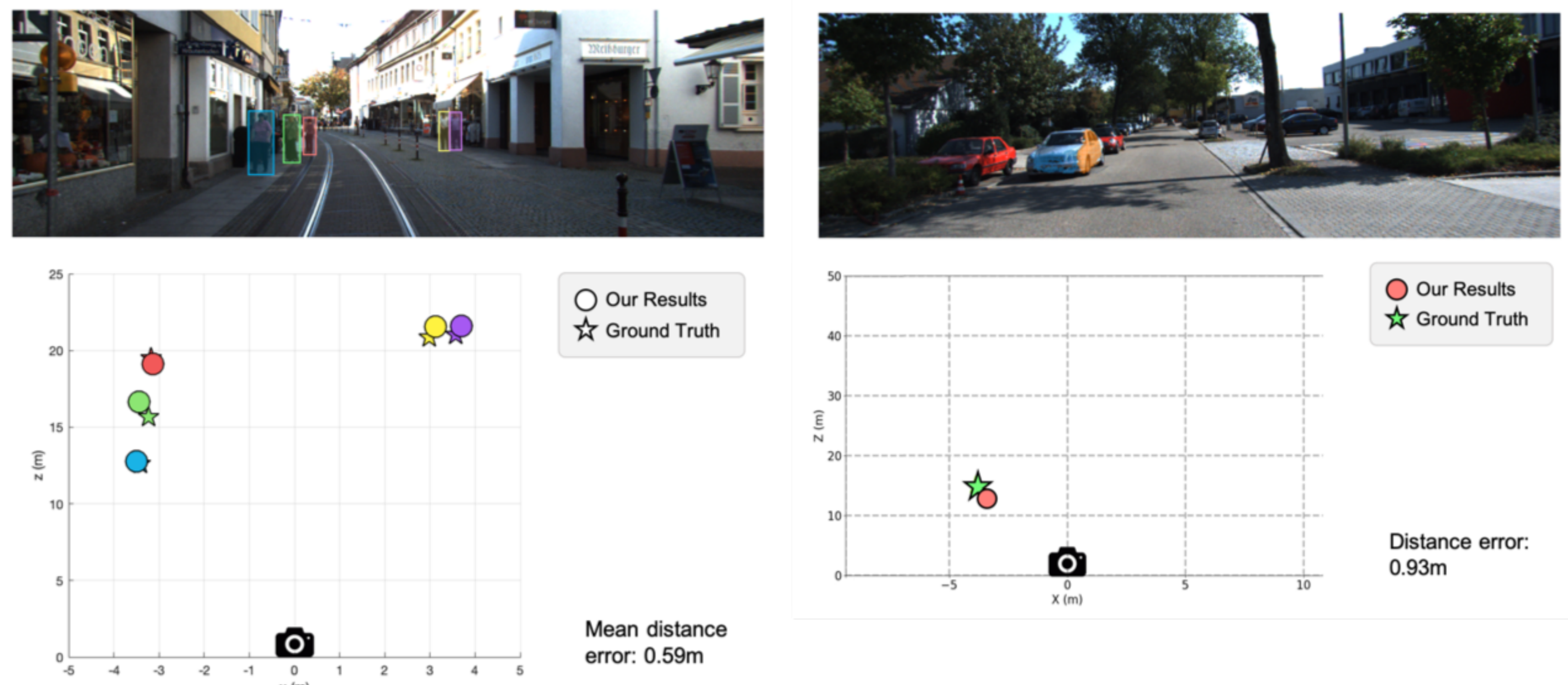


Table 1: Mean localization error (standard deviation) for pedestrians compared with some vehicle localization methods.

Methods	Overall (m)	≤ 15m	≤ 30m	> 30m	Running speed
Murthy et al. [27]	2.61 (±2.23)	1.59 (±0.96)	2.52 (±2.16)	4.30 (±2.83)	–
Ansari et al. [2]	1.00 (±0.77)	0.67 (±0.50)	0.94 (±0.69)	2.19 (±1.18)	–
Ansari et al. (Opt) [2]	0.86 (±0.87)	0.55 (±0.50)	0.79 (±0.79)	2.16 (±1.18)	–
Ours (DHist)	0.79 (±0.75)	0.43 (±0.31)	0.76 (±0.73)	2.78 (±2.01)	6.1 FPS
Ours (DHist+AGPE)	0.74 (±0.64)	0.43 (±0.31)	0.71 (±0.63)	2.39 (±1.61)	3.3 FPS
Ours (DHist+TS)	0.73 (±0.62)	**0.40 (±0.30)**	0.71 (±0.61)	2.15 (±1.32)	2.6 FPS
Ours (DHist+AGPE+TS)	**0.69 (±0.51)**	0.42 (±0.33)	**0.68 (±0.53)**	**1.22 (±0.74)**	2.0 FPS

Table 2: Vehicle 3D localization results based on qualified vehicle surface detection on KITTI Sequence 0009.

Methods	Mean localization error (m)
Depth histogram	1.19 (±0.90)
3D point cloud	0.83 (±0.73)

Table 3: Mean ground normal error for different ground plane estimation methods.

Methods	Ground normal error (deg)
HMM [6]	4.10
GroundNet [20]	0.96
Ours (DGPE)	0.79
Ours (SGPE)	0.89
Ours (AGPE)	**0.74**

REFERENCES

[1] Godard, Clément, et al. "Unsupervised monocular depth estimation with left-right consistency." *CVPR*. 2017.

[2] He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.

[3] Achanta, Radhakrishna, et al. "SLIC superpixels compared to state-of-the-art superpixel methods." *TPAMI*. 2012.

[4] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." *TPAMI*. 2017.

[5] Wang, Gaoang, et al. "Exploit the connectivity: Multi-object tracking with trackletnet." *ACM Multimedia,* 2019.

Project Website