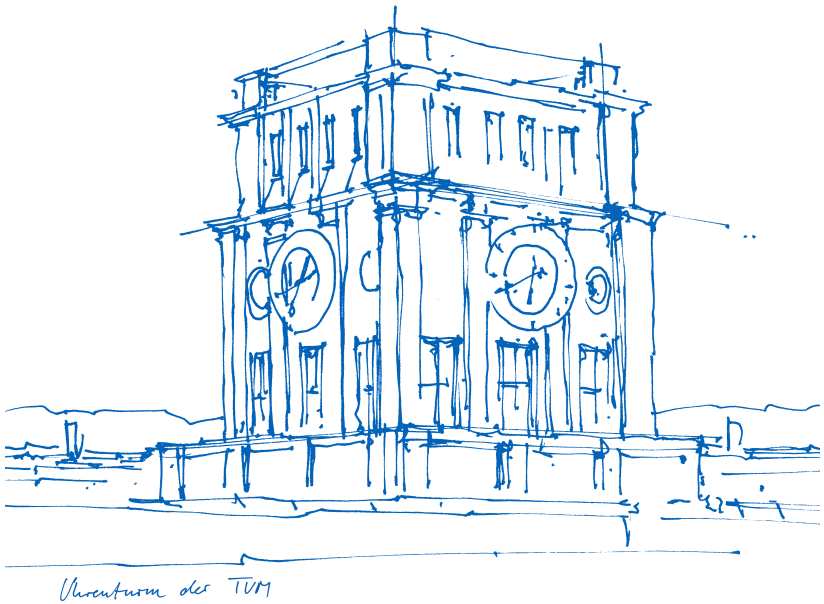
Incorporating Large Vocabulary Object Detection and Tracking into Visual SLAM

Maximilian Kempa

Master‘s Thesis in Robotics, Cognition, Intelligence



Technical University of Munich

Department of Informatics

Chair of Computer Vision and Artificial Intelligence

14 January 2021

Introduction

Motivation

- Good performance:
  - SLAM
  - Object detection and tracking

→ Development of combined systems

- Most systems: Focus on car and pedestrian

Goal

Estimate
- camera trajectory
- trajectory of surrounding objects of **many** classes

Related Work

Object Detection

CenterNet [1]:

- Center point of bounding boxes
- Regresses the bounding box size as offset

Object Tracking

CenterTrack [2]:

- Displacement of center point between current and previous frame

Related Work

Simultaneous Localization and Mapping (SLAM)

- Semantic Methods
- **Simultaneous tracking of ego motion and moving objects**

ORB-SLAM [6]:

- Open-source solution for visual monocular SLAM
- ORB-SLAM2 [7] enhances ORB-SLAM [6] to stereo images

Related Work

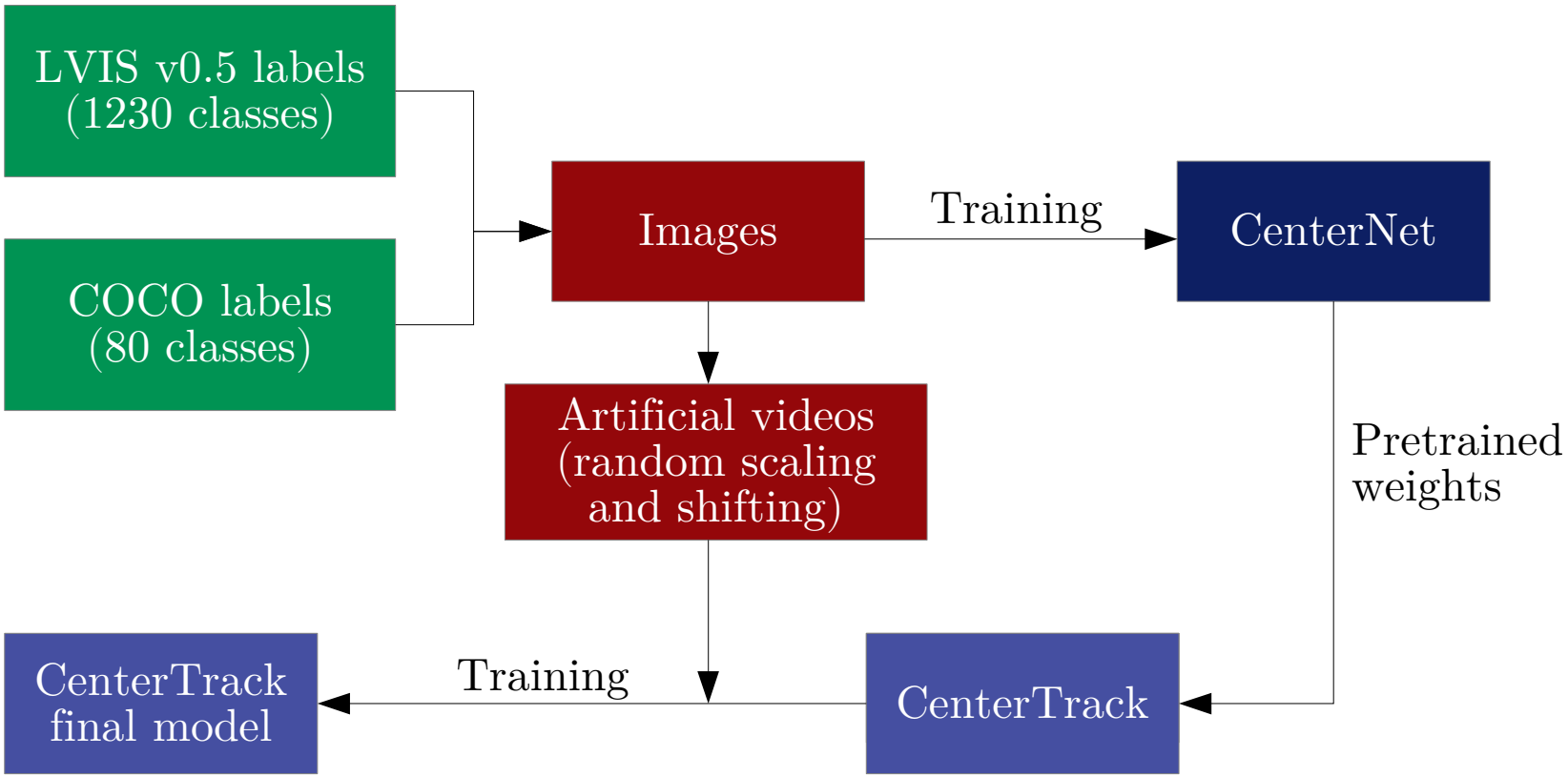
COCO
- Instance Segmentation
- 80 classes

LVIS (Large Vocabulary Instance Segmentation)
- Instance Segmentation
- 1230 classes

KITTI
- Odometry
- 2D and 3D Object Tracking
- 2 classes

Object Detection and Tracking

Training Procedure Overview



Object Detection and Tracking

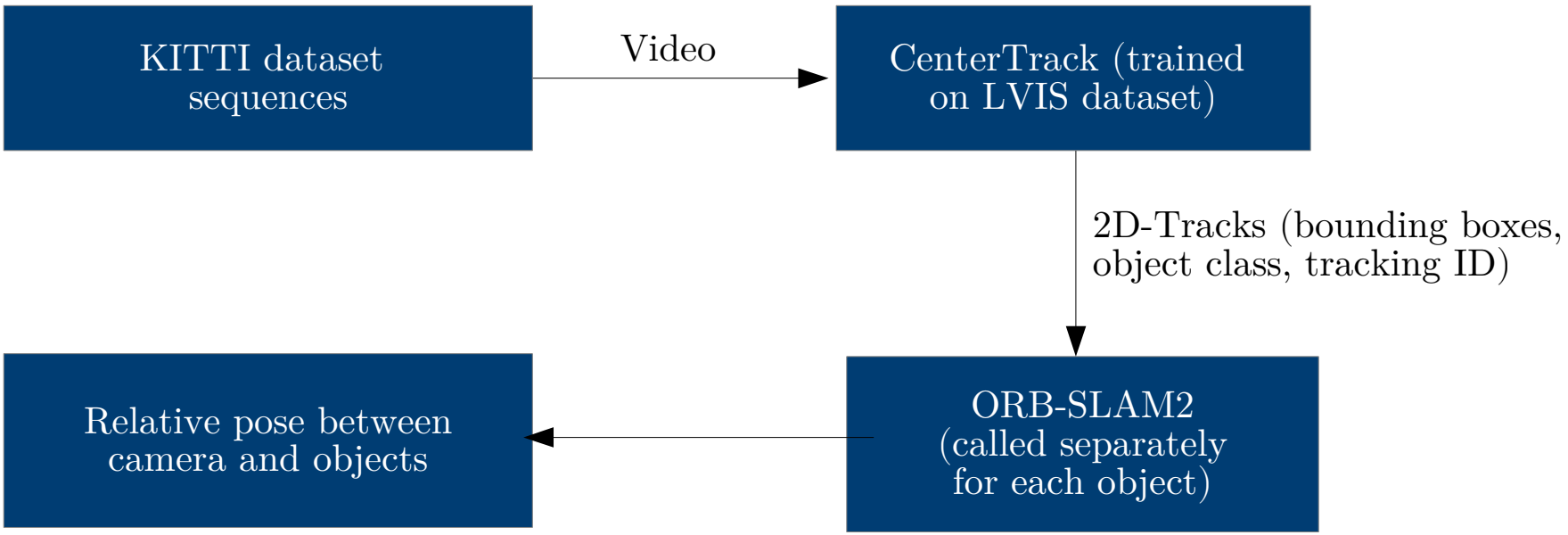
Hyperparameter Tuning and Evaluation of Tracker Network

- Multiple object tracking accuracy (MOTA) score [9]

- Statistical evaluation of track length

Incorporating 2D Object Tracking into SLAM

System Overview



Incorporating 2D Object Tracking into SLAM

3D Object Tracking Algorithm (Modified ORB-SLAM2)

- Extract ORB features inside the 2D bounding box of the object
- Match stereo keypoints
- Optimize transformation from camera to object frame
- Recover object trajectories in world frame

Evaluation

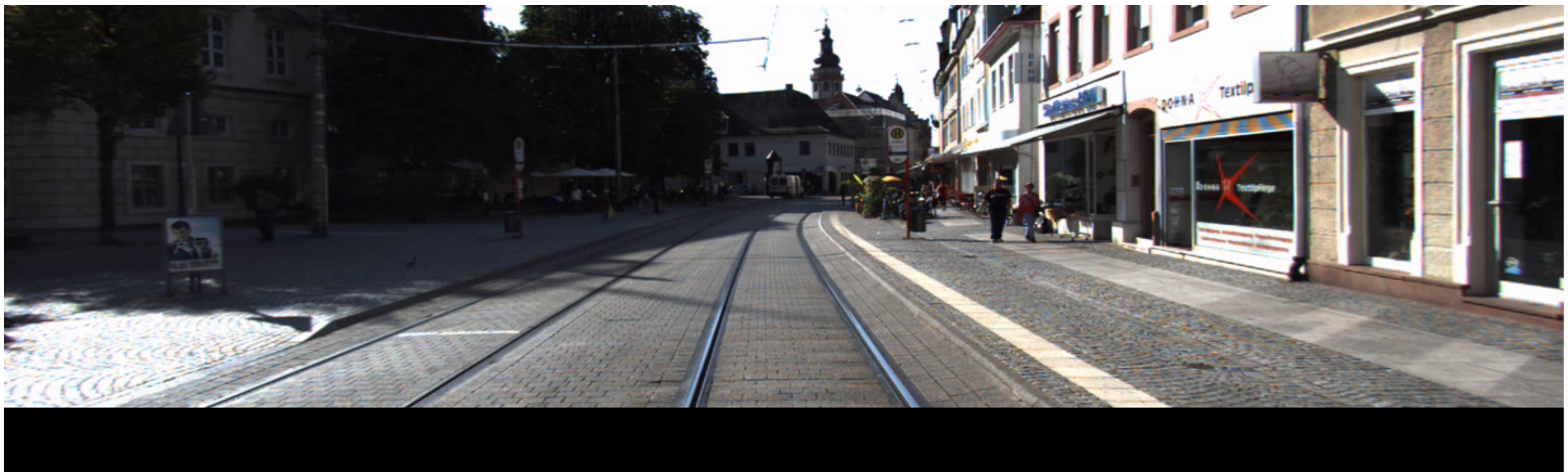
Qualitative Evaluation 3D Object Tracking

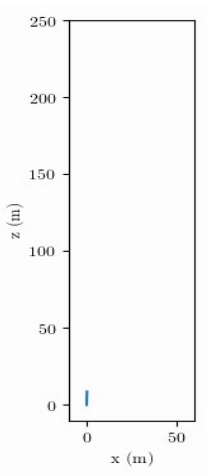
Most frequently tracked classes in 3D (without car and pedestrian)

- Bicycle
- Traffic Light
- Bus
- Street Sign
- Pot (flower pot)

Evaluation

Qualitative Evaluation 3D Object Tracking





Video playback speed: 0.5x

Evaluation

Qualitative Evaluation 3D Object Tracking

Small bounding boxes are difficult to track.

Class	Mean of $\sqrt{w_{bbox} * h_{bbox}}$	Recall $Recall = \frac{N_{3D}}{N_{2D}}$
Taillight	19.44	11.30%
License Plate	32	23.08%
Traffic Light	33.66	26.91%
Bag	39.52	31.71%

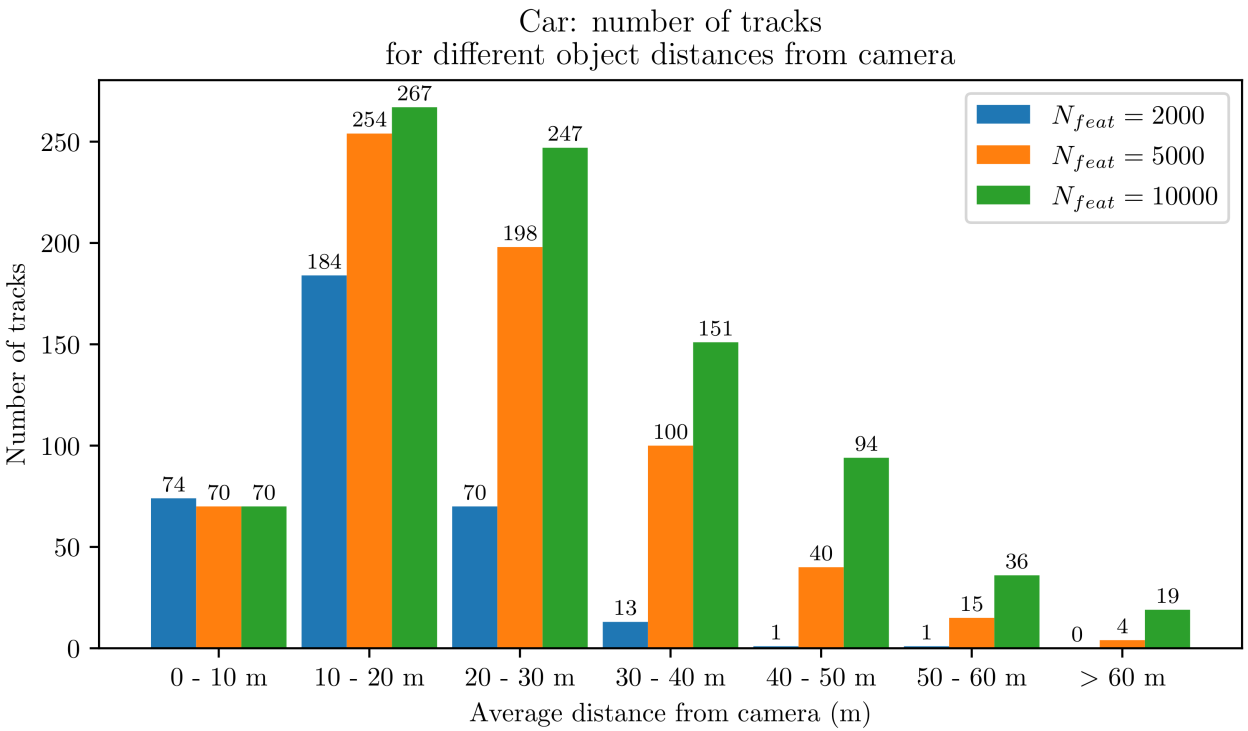
Evaluation

Quantitative Evaluation 3D Object Tracking

- 3D tracks with a minimum length of 5 frames
- N_{feat} = maximum number of ORB features
- Increasing N_{feat} → more features inside the 2D object bounding boxes

Evaluation

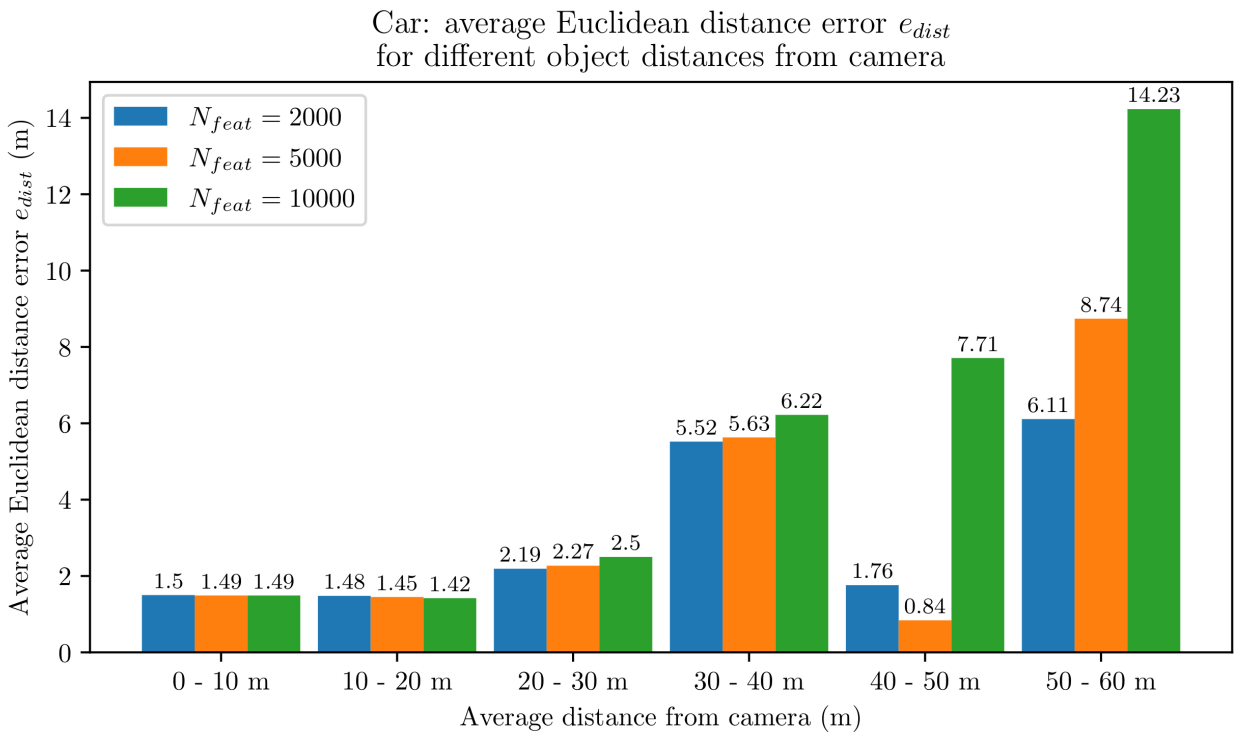
Quantitative Evaluation 3D Object Tracking



Evaluation

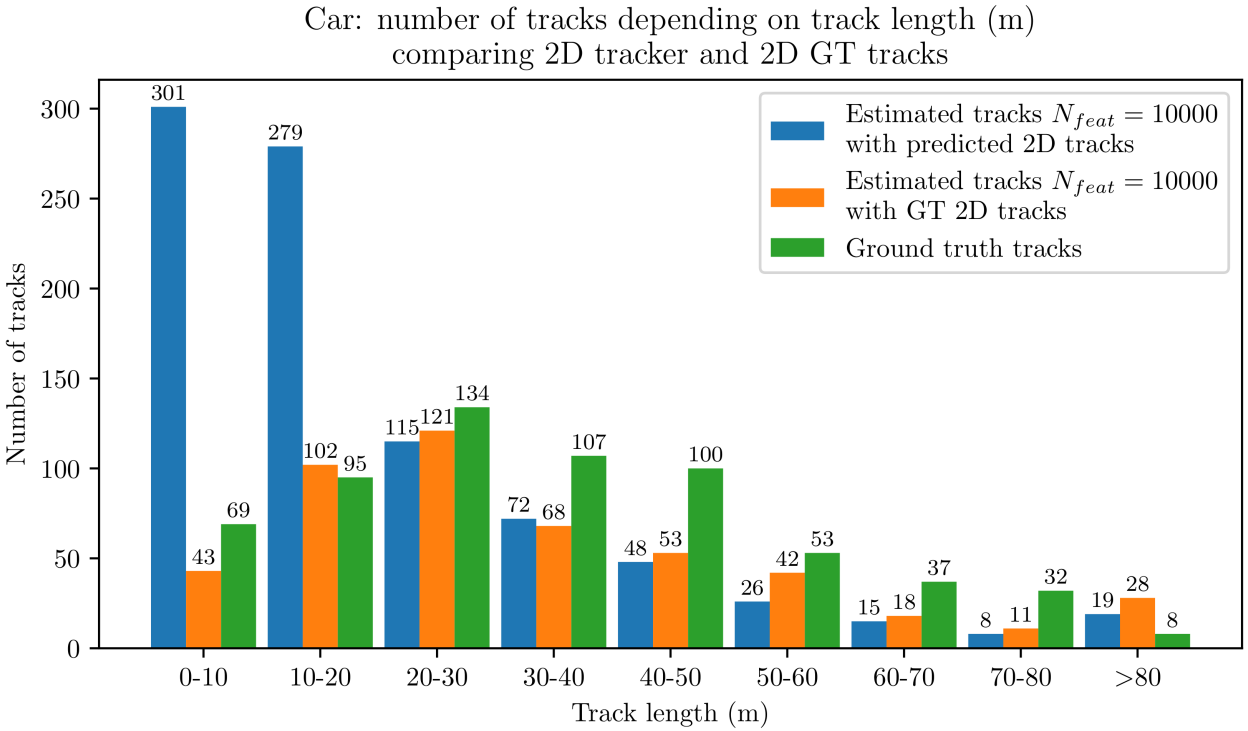
Quantitative Evaluation 3D Object Tracking

Only considered tracks that are already present with $N_{feat}=2000$



Evaluation

Quantitative Evaluation 3D Object Tracking



Evaluation

Comparison to DynaSLAM II [11]

- Builds up on ORB-SLAM2 [7]
- Pixel-wise-semantic segmentation for each 2D frame
- Smooth-motion prior for optimization
- Estimates 3D bounding boxes

Evaluation

Comparison to DynaSLAM II [11]

Sequence	0003	0005	0010	0011	0011	0018	0018	0019	0019	0020	0020	0020
Object ID GT	1	31	0	0	35	2	3	63	72	0	12	122
DynaSLAM II ATE	**0.69**	**0.51**	0.95	1.05	1.25	0.86	**0.99**	0.86	0.99	**0.56**	**1.18**	**0.87**
Ours ATE	0.72	0.78	**0.53**	**0.38**	**0.66**	**0.16**	1.05	**0.28**	**0.17**	3.98	2.95	6.10
DynaSLAM II 2D TP (%)	50.00	28.96	**81.63**	72.65	53.17	**86.36**	**53.33**	35.26	**29.11**	63.68	**42.77**	**34.90**
Ours 2D TP (%)	**55.74**	**78.79**	31.29	**98.39**	**64.75**	40.15	2.11	**52.02**	9.18	**88.06**	13.53	5.22

Conclusion

- Estimation of trajectory of camera and surrounding objects
- Big variety of classes
- High number of features is crucial to enable successful 3D tracking
- Comparison to DynaSLAM II [11]
  → competitive trajectory localization accuracy

Conclusion

Future Work

- Finetuning of the 2D tracker network
- Integration of semantic segmentation network
- 3D bounding box estimation module
- Meaningful confidence scores for the 3D detections
- Standard benchmark for comparison of dynamic SLAM algorithms

Literature

[1] X. Zhou, D. Wang, and P. Krähenbühl. “Objects as Points”. In: arXiv:1904.07850 (2019).
[2] X. Zhou, V. Koltun, and P. Krähenbühl. “Tracking Objects as Points”. In: ECCV. 2020.
[3] A. Gupta, P. Dollár, and R. Girshick. “LVIS: A Dataset for Large Vocabulary Instance Segmentation”. In: CVPR. 2019.
[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick. “Microsoft COCO: Common Objects in Context”. In: ECCV. 2014.
[5] A. Geiger, P. Lenz, and R. Urtasun. “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite”. In: CVPR. 2012.
[6] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós. “ORB-SLAM: A Versatile and Accurate Monocular SLAM System”. In: IEEE Transactions on Robotics 31.5 (2015), pp. 1147–1163.
[7] R. Mur-Artal and J. D. Tardós. “ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras”. In: IEEE Transactions on Robotics 33.5 (2017), pp. 1255–1262.
[8] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. “ORB: An efficient alternative to SIFT or SURF”. In: ICCV. 2011.

Literature

[9] K. Bernardin, A. Elbs, and R. Stiefelhagen. “Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment”. In: Proceedings of IEEE International Workshop on Visual Surveillance (2006).
[10] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. “A Benchmark for the Evaluation of RGB-D SLAM Systems”. In: IROS. 2012.
[11] B. Bescos, C. Campos, J. D. Tardós, and J. Neira. DynaSLAM II: Tightly-Coupled Multi-Object Tracking and SLAM. 2020. arXiv: 2010.07820 [cs.RO].