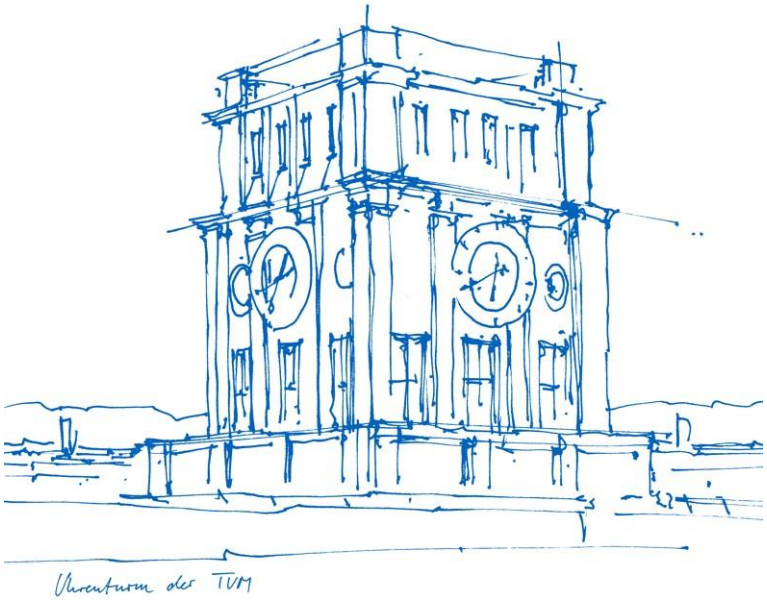
Computer Vision Group
Department of Informatics
Technical University of Munich

Direct Object Tracking

Guided Research Presentation
28.04.2021

Korobov Nikita

Supervisors:
Nikolaus Demmel, Aljosa Osep



©2021 Technical University of Munich

Motivation

- Tracking of objects is “must have” for an autonomous system;
- Static SLAM can benefit from tracking of dynamic objects [1].

Goal

- Build 3D-2D MOT system using direct methods.

[1] Yang, Shichao, and Sebastian Scherer. "Cubeslam: Monocular 3-d object slam.

Related work

Dynamic SLAM

- Dynamic scenes are difficult for standard SLAM systems;
- Excluding dynamic parts may help in SLAM [1];
- SLAM and tracking of dynamic objects may be coupled [2].

3D MOT

- Use 3D info to improve 2D tracking [3, 4];
- Use motion model based tracker given detections [5];
- Use 2D-3D appearance based association [6].

[1] Yang, Shichao, and Sebastian Scherer. "Cubeslam: Monocular 3-d object slam."

[2] Ballester, Irene, et al. "DOT: Dynamic Object Tracking for Visual SLAM."

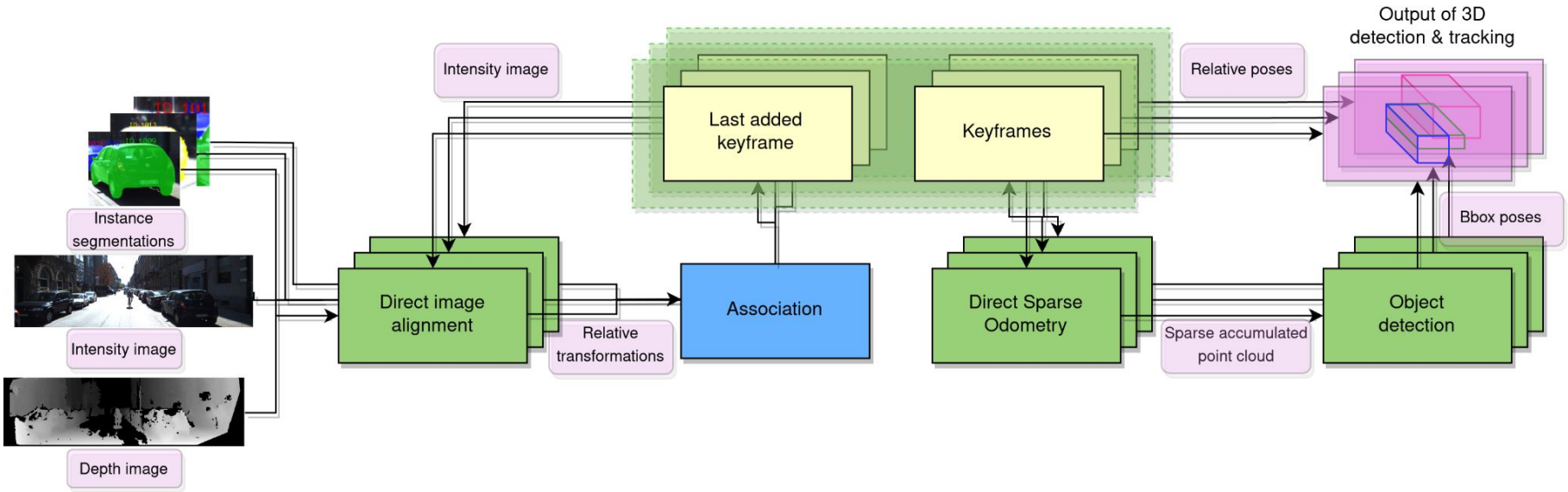
[3] Osep, Aljoša, et al. "Combined image-and world-space tracking in traffic scenes."

[4] Luiten, Jonathon, Tobias Fischer, and Bastian Leibe. "Track to reconstruct and reconstruct to track."

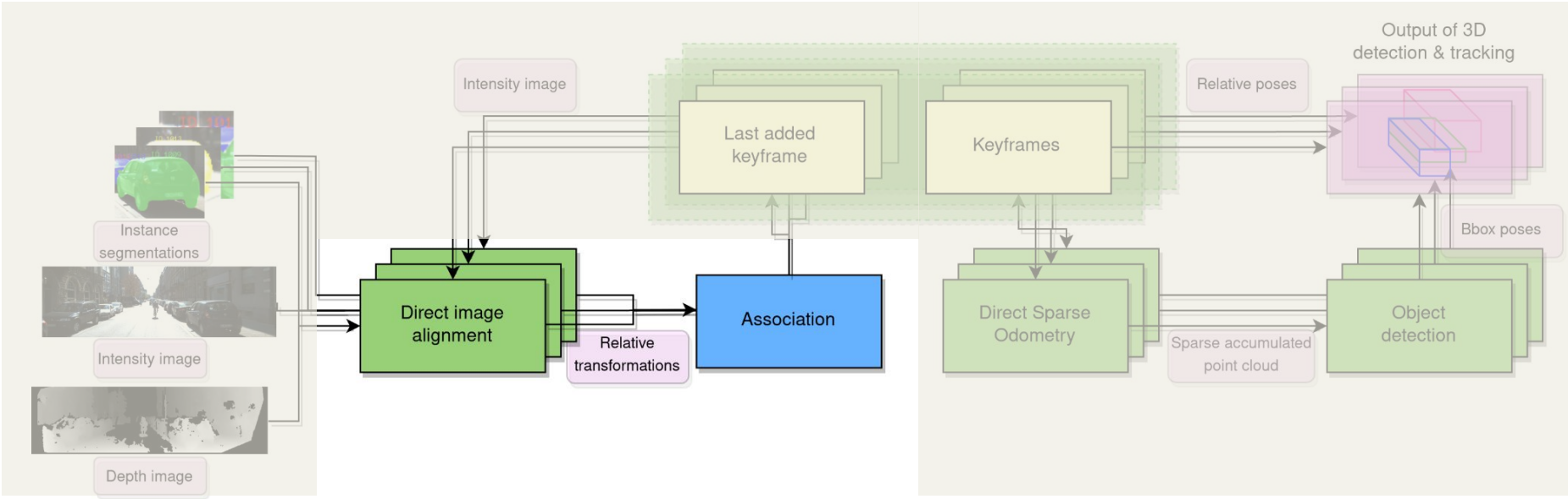
[5] Weng, Xinshuo, et al. "AB3DMOT: A Baseline for 3D Multi-Object Tracking and New Evaluation Metrics."

[6] Baser, Erkan, et al. "Fantrack: 3d multi-object tracking with feature association network."

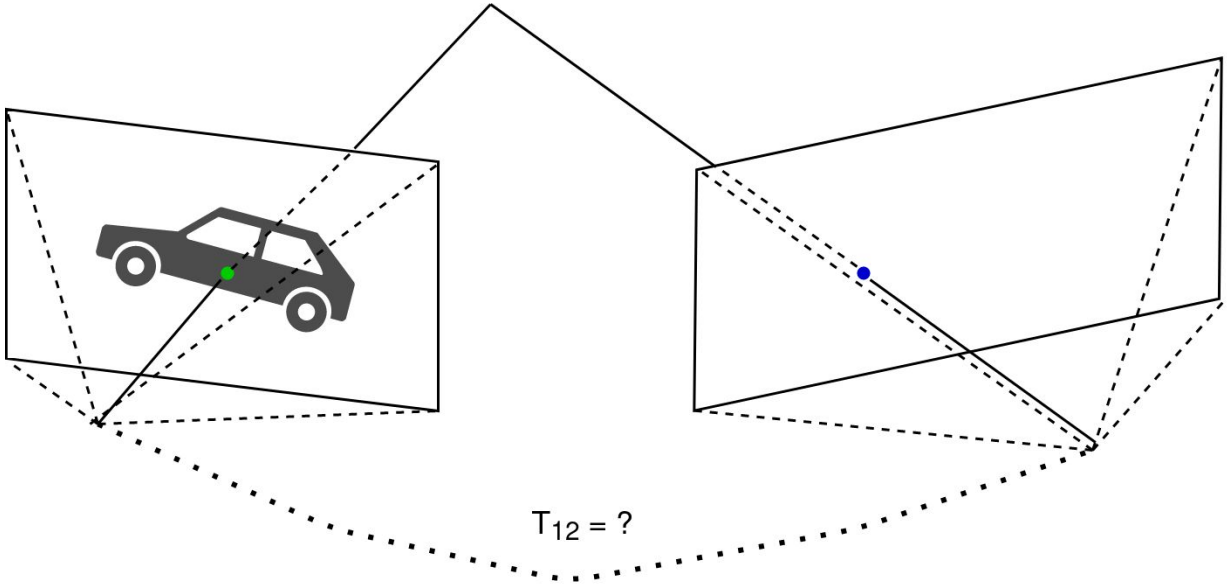
Proposed method



Proposed method



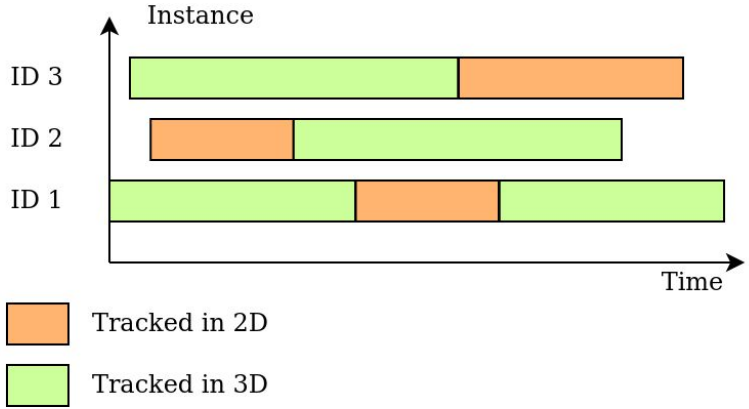
Direct Image Alignment

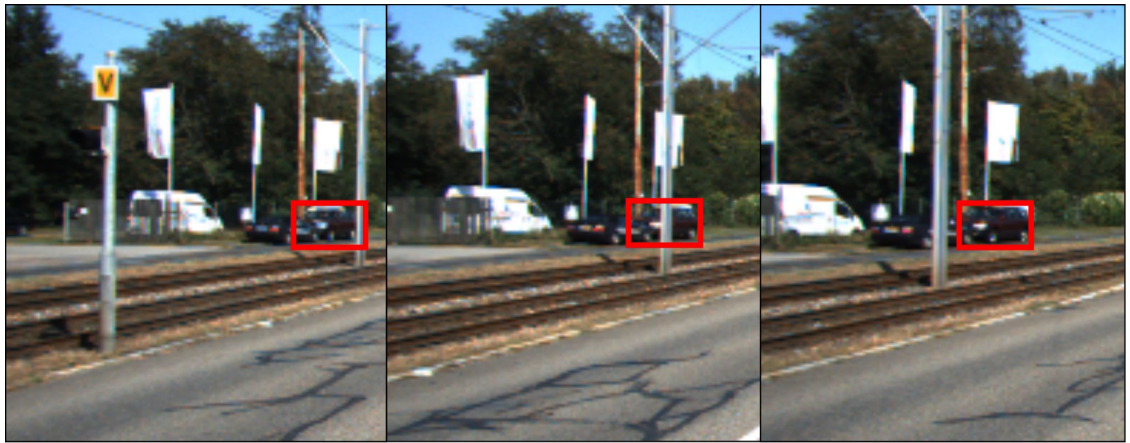


- For each pixel in mask;
- Image pyramid;
- Depth is known for one frame;
- Photometric error.

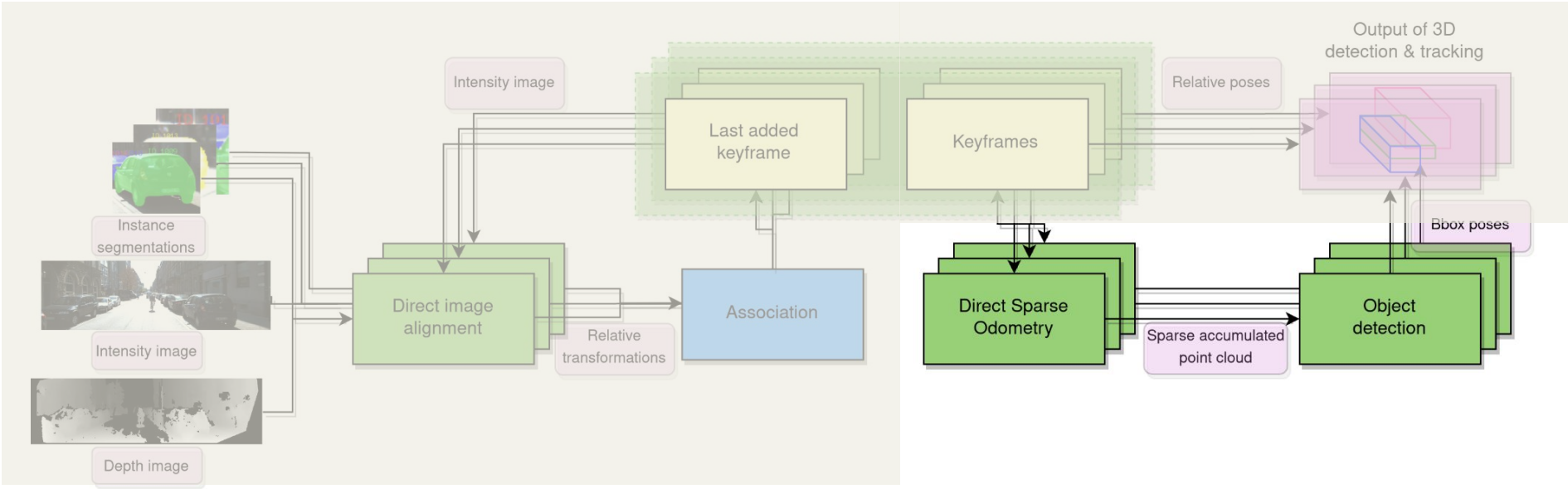
2D tracking

- Sparse optical flow;
- Opportunistic tracking: try 3D, if fails, then 2D;
- Helps to keep the tracklet if 3D tracking fails.

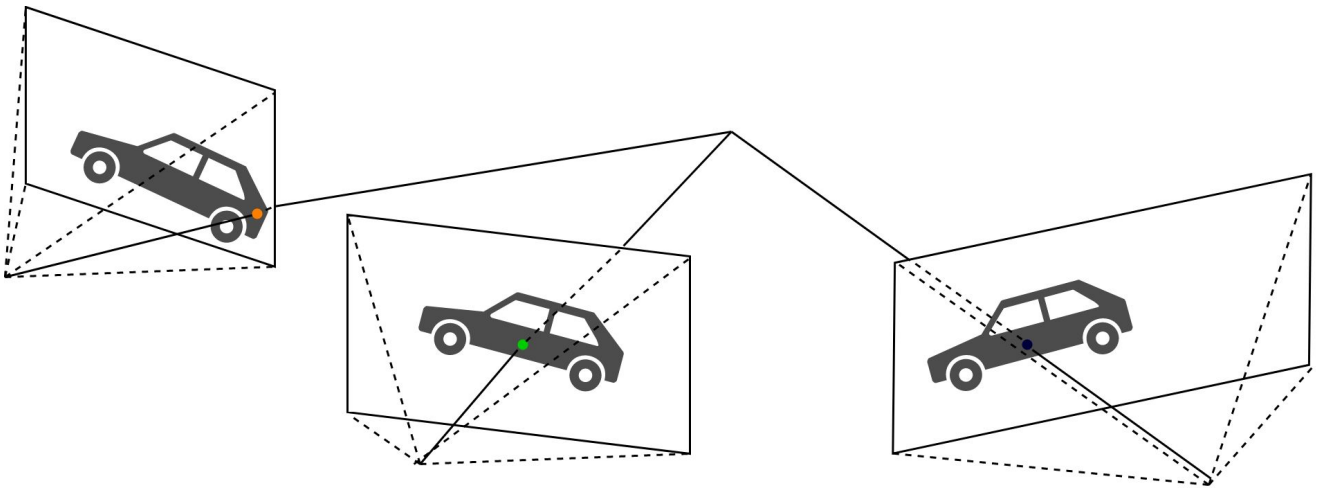




Proposed method

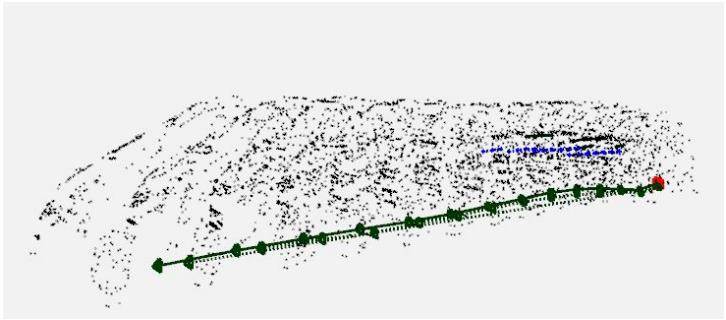


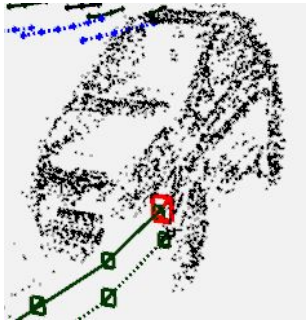
Direct Sparse Odometry

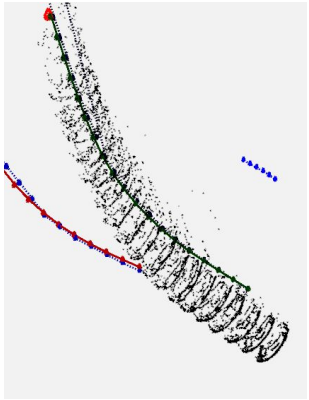


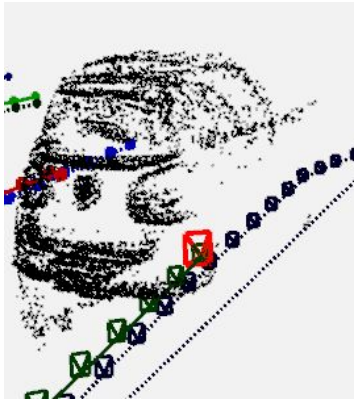
- Sparse points;
- Joint optimization of depth and poses;
- Photometric error.

Accumulated point clouds





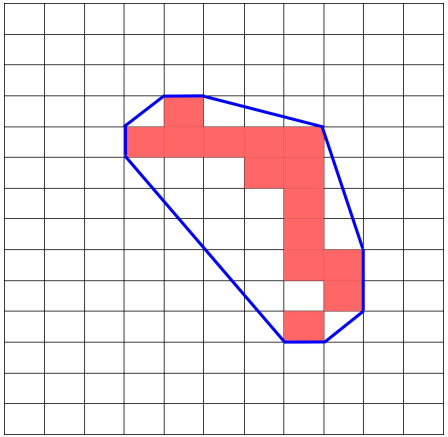




Dashed line - GT trajectory.
Solid line - estimated trajectory.
Red anchor point - reference frame.

Object localization

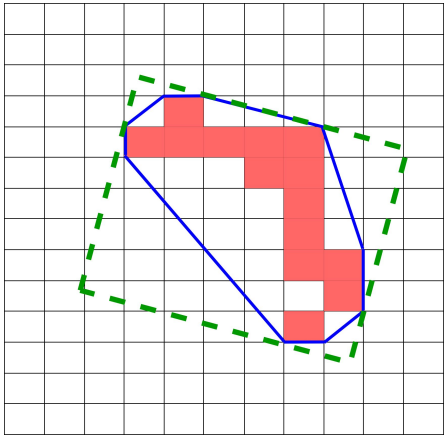
- No 3D supervision is required;
- Dataset specific assumptions;
- Object existence is guaranteed in the accumulated point cloud;
- Finite number of bbox hypothesis;
- Bbox size correction.



Red - occupied cells.
Blue line - convex hull.
Green dashed rectangle - bounding box hypothesis.

Object localization

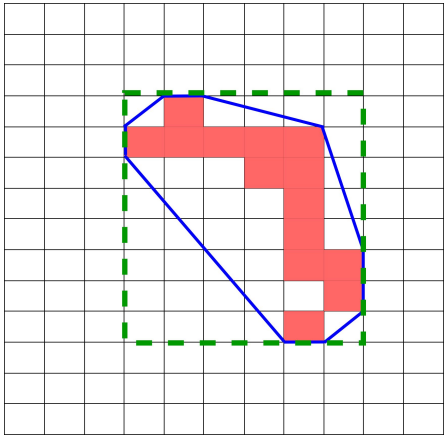
- No 3D supervision is required;
- Dataset specific assumptions;
- Object existence is guaranteed in the accumulated point cloud;
- Finite number of bbox hypothesis;
- Bbox size correction.



Red - occupied cells.
Blue line - convex hull.
Green dashed rectangle - bounding box hypothesis.

Object localization

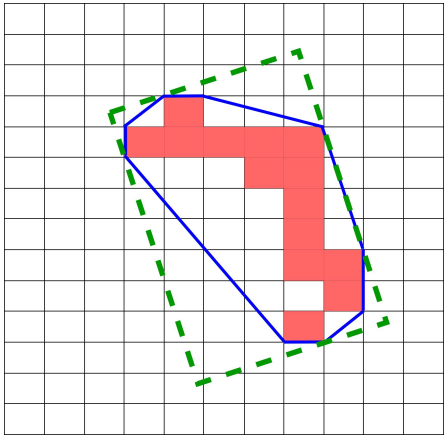
- No 3D supervision is required;
- Dataset specific assumptions;
- Object existence is guaranteed in the accumulated point cloud;
- Finite number of bbox hypothesis;
- Bbox size correction.



Red - occupied cells.
Blue line - convex hull.
Green dashed rectangle - bounding box hypothesis.

Object localization

- No 3D supervision is required;
- Dataset specific assumptions;
- Object existence is guaranteed in the accumulated point cloud;
- Finite number of bbox hypothesis;
- Bbox size correction.



Red - occupied cells.
Blue line - convex hull.
Green dashed rectangle - bounding box hypothesis.

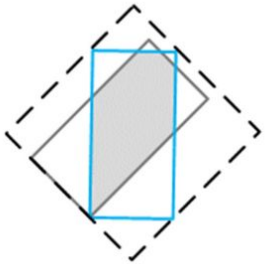
Evaluation 3D MOT metric

- MOTA imbalances DetA and AssA -> HOTA;
- HOTA -> integration over similarity thresholds;
- **3D IoU is 0** for non-overlapping objects -> detection is counted as FP, GT as FN -> **3D GIoU**;

$$GIoU = \frac{A \bigcap B}{A \bigcup B} - \frac{C \backslash (A \bigcap B)}{C}$$

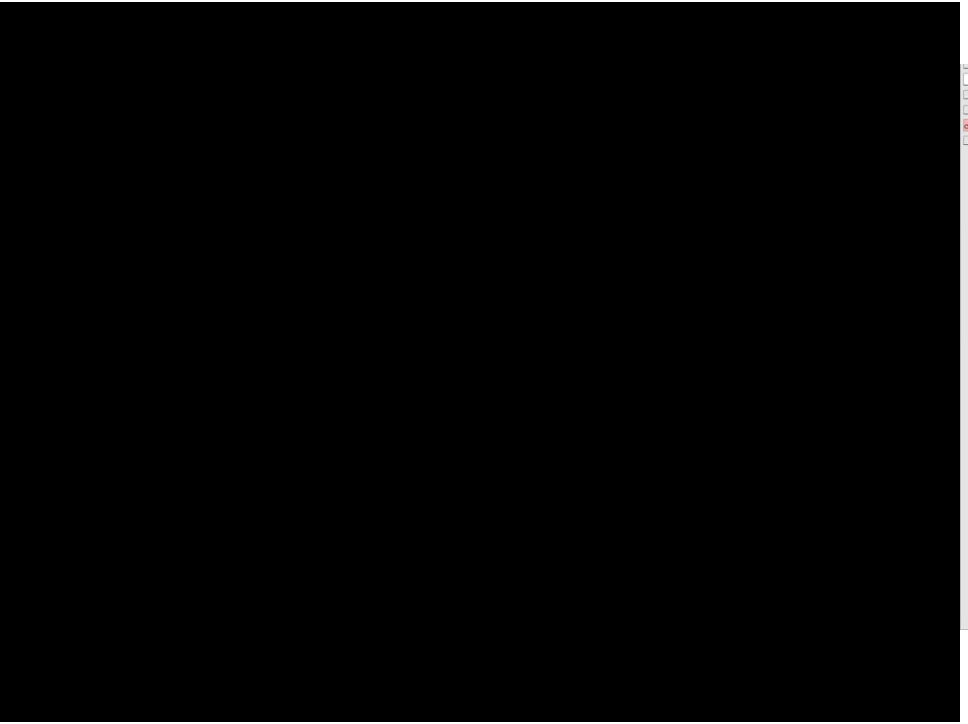
$$HOTA = \int_0^1 \sqrt{AssA_\alpha * DetA_\alpha} d\alpha$$

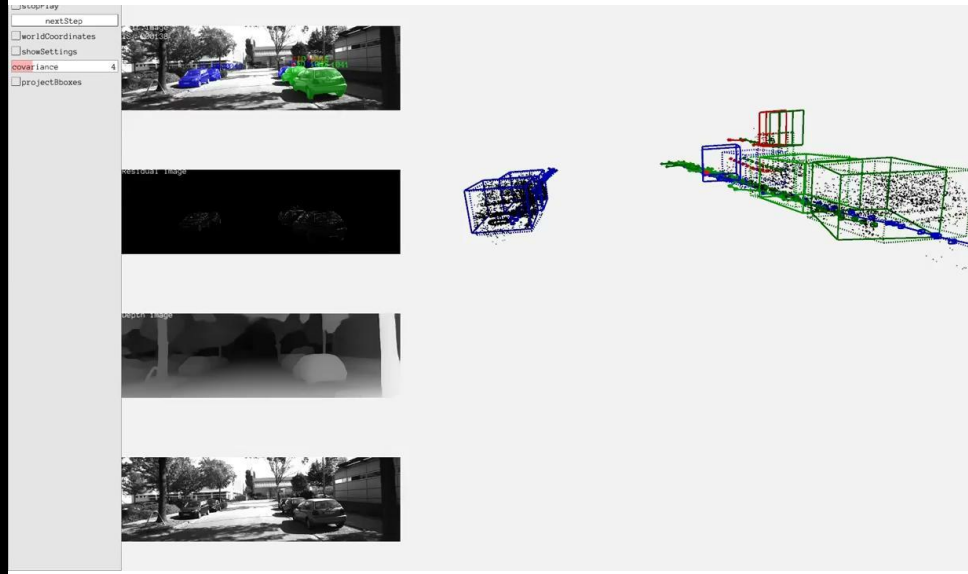
C is the smallest enclosing bbox



$$MOTA = 1 - \frac{\text{Misses} + \text{FP} + \text{Switches}}{\text{GT}}$$

Demo




nextStep
worldCoordinates
showSettings
covariance
4
projectBoxes

Experiments

- **KITTI object tracking dataset [7], validation split (11 sequences);**
- Competitors: PointRCNN [1], DispRCNN [2], DSGN [3], AB3DMOT [4] (as tracker);
- **HOTA [5] + 3D GIoU [6].**

[1] Shi, Shaoshuai, Xiaogang Wang, and Hongsheng Li. "Pointrcnn: 3d object proposal generation and detection from point cloud."

[2] Sun, Jiaming, et al. "Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation."

[3] Chen, Yilun, et al. "Dsgn: Deep stereo geometry network for 3d object detection."

[4] Weng, Xinshuo, et al. "AB3DMOT: A Baseline for 3D Multi-Object Tracking and New Evaluation Metrics."

[5] Luiten, Jonathon, et al. "HOTA: A higher order metric for evaluating multi-object tracking."

[6] Xu, Jun, et al. "3D-GIoU: 3D generalized intersection over union for object detection in point cloud."

[7] Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite."

Results of 3D MOT

Tracker	Detector	HOTA	DetA	AssA	DetRe	DetPr	AssRe	AssPr	LocA
			GIoU						
AB3DMOT	PointRCNN (LiDAR)	**73.85**	70.92	77.38	81.37	79.27	80.93	90.27	88.74
AB3DMOT	DSGN (12 Gb)	48.46	47.08	53.59	50.58	76.23	55.75	91.41	81.70
AB3DMOT	DSGN (full)	**55.78**	52.94	62.41	57.56	76.17	64.84	91.66	82.11
AB3DMOT	DispRCNN (vob)	67.21	66.18	69.59	69.79	84.43	72.00	91.09	85.92
AB3DMOT	DispRCNN (pob)	**67.35**	67.49	68.28	71.71	83.16	70.65	89.07	85.27
Ours	Bbox	**49.16**	46.50	54.06	54.32	56.00	59.51	68.11	70.77
			IoU						
AB3DMOT	PointRCNN (LiDAR)	**65.59**	61.65	70.67	72.38	70.51	75.14	82.98	81.80
AB3DMOT	DSGN (12 Gb)	40.60	35.32	51.93	40.08	60.40	55.19	84.29	75.10
AB3DMOT	DSGN (full)	**46.51**	39.82	59.71	46.00	60.87	63.46	85.44	75.89
AB3DMOT	DispRCNN (vob)	58.37	55.57	63.40	61.59	71.31	67.06	82.45	78.59
AB3DMOT	DispRCNN (pob)	**57.62**	55.519	61.24	61.346	71.145	64.71	80.342	78.243
Ours	Bbox	**31.60**	26.90	39.05	35.48	36.58	45.93	51.27	65.07

3D MOT task; HOTA with 3D GIoU and HOTA with 3D IoU.

Results of 2D MOT

Tracker	Detector	MOTA	MOTP	IDs	FP	FN
AB3DMOT	PointRCNN (LiDAR)	75.62	86.97	28	701	1110
AB3DMOT	DSGN (12 Gb)	57.24	87.03	204	224	2797
AB3DMOT	DSGN (full)	64.64	88.93	130	314	2223
AB3DMOT	DispRCNN (vob)	84.05	89.97	64	91	1048
AB3DMOT	DispRCNN (pob)	87.76	89.99	65	93	1067
MOTSFusion	RRC + BB2SegNet	94.0	-	9	45	400
CIWT	Regionlets	74.38	82.85	26	-	-
Ours	BB2SegNet for 2D + Convex hull for 3D	89.13	82.18	78	192	550

2D bounding box MOT evaluation on validation set. For 3D methods the 3D bounding box is projected on the image plane.

Ablation study

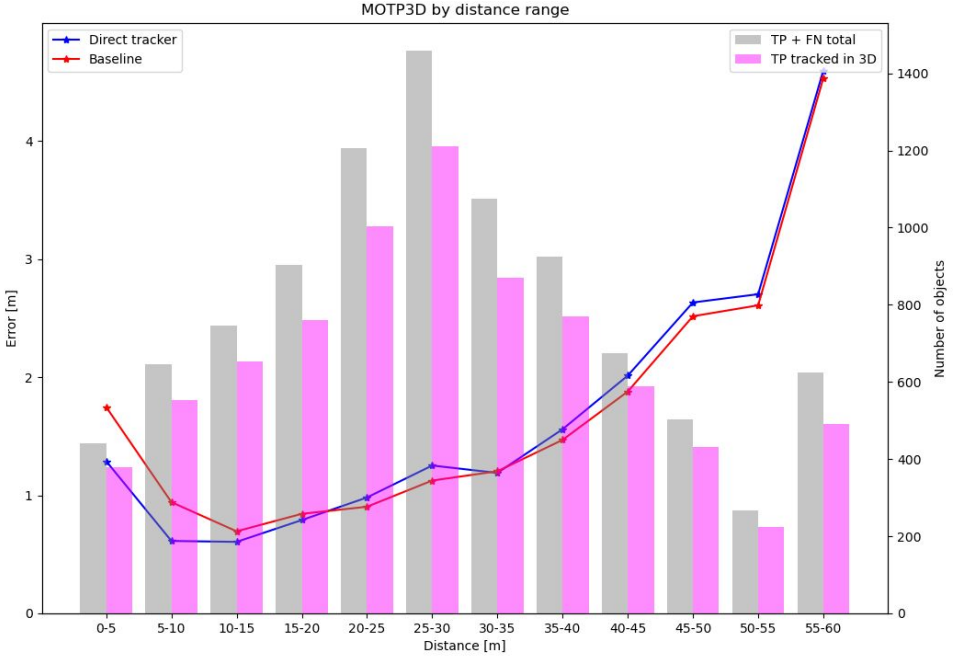
Setup	HOTA	DetA	AssA	DetRe	DetPr	AssRe	AssPr	LocA
LibELAS depth [29]	44.35	40.30	51.53	48.65	50.43	57.57	65.78	69.19
SGBM Depth [28], OpenCV	38.87	36.08	43.62	44.77	46.37	50.89	58.85	68.06
W/o RANSAC pose initialization	48.00	45.11	52.81	52.17	56.78	58.16	68.08	71.04
W/o DIA failure filtering	48.04	45.88	52.47	53.43	55.69	57.61	68.09	70.56
W/o DSO optimization	48.67	46.05	53.29	53.43	56.37	58.67	68.00	70.82
W/o 2D tracking	46.69	45.97	49.26	53.24	56.41	53.79	68.82	70.73
With GT masks	51.17	45.98	58.90	54.18	54.75	66.02	66.40	70.348
Baseline	50.88	48.67	54.66	56.40	58.21	60.09	68.57	71.65
Full system	49.16	46.50	54.06	54.32	56.00	59.51	68.11	70.77

HOTA + 3D GIoU

- **Baseline**: tracking as in full system, but object detection is from single depth map (rather than accumulated point cloud);
- Method significantly relies on the accurate depth map;
- 2D tracking significantly helps to associate the detections, when 3D tracking fails.

Full version vs Baseline

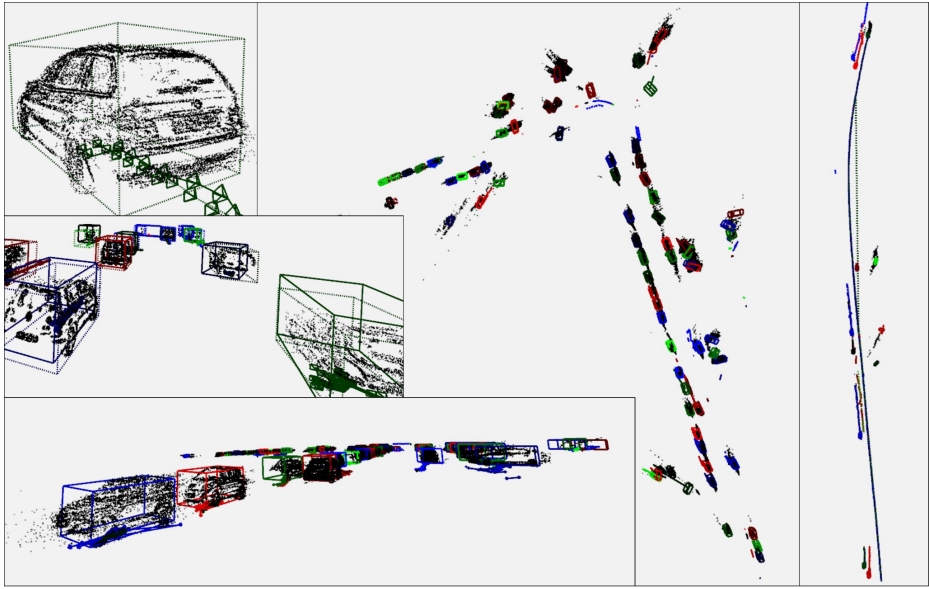
$$MOTP3D = \frac{\sum_{i,t} \left\| \mathbf{p}^i_{gt}(t) - \mathbf{p}^i_{det}(t) \right\|}{\sum_t n_{tp}(t)}$$



- Baseline is worse for objects located close to the camera.

Conclusion

- Sparse and direct method to track objects in 2D and 3D;
- No 3D supervision required;
- Promising results for 3D and 2D MOT;
- HOTA + GIoU as the 3D MOT metric.



Future work

- Better 3D object detector in point cloud;
- Need for fair evaluation of the competing methods.