BAD SLAM: Bundle Adjusted Direct RGB-D SLAM

Thomas Schöps[1], Torsten Sattler[2], Marc Pollefeys[1,3]

[1]Department of Computer Science, ETH Zürich [2]Chalmers University of Technology [3]Microsoft

Benchmark Dataset & Open Source Code: www.eth3d.net

Contributions & Conclusions

1. A novel RGB-D SLAM approach
- Using alternating **direct Bundle Adjustment**, demonstrating that this is **real-time capable** on a GPU for short videos
- Released as **open source** (BSD licensed)

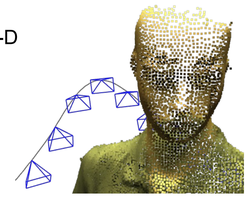
2. A well-calibrated SLAM benchmark
- For **visual-inertial mono, stereo, and RGB-D SLAM**
- Using **well-calibrated synchronized global-shutter cameras** with active infrared stereo for depth estimation
- With ground truth poses by a motion capturing system
- Consists of a **training set (61 datasets),** and a **test set (35 datasets) without public ground truth**
- With an **online evaluation service**

Conclusions
- To some extent, dense & direct BA is possible in real-time
- Direct RGB-D SLAM methods seem to perform better on our benchmark than on existing ones due to its good calibration

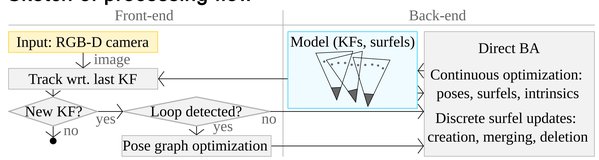
SLAM approach overview

Data representation

- **Keyframes** to store sparse RGB-D frames from input video
- **Surfels** to model the observed surfaces (dense but slightly sparsified)



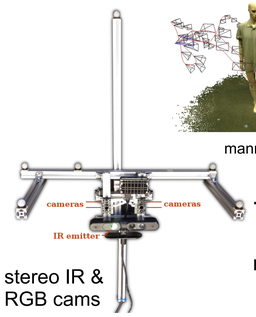
Sketch of processing flow



- Odometry-style camera pose tracking in front-end
- Simultaneous **alternating optimization of map and poses** (and optionally intrinsics) in back-end

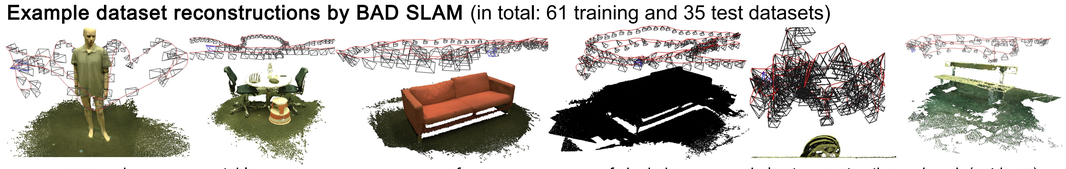
Benchmark dataset overview

Camera rig



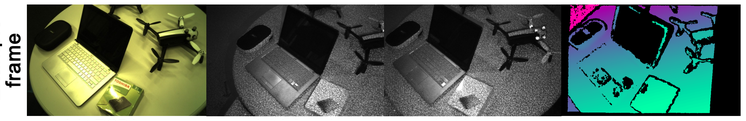
stereo IR & RGB cams

Example dataset reconstructions by BAD SLAM (in total: 61 training and 35 test datasets)



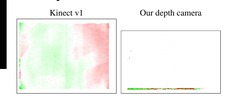
mannequin | table scene | sofa | sofa in darkness | helmet reconstruction | bench (outdoors)

Example frame



right RGB image (no white balancing!) | left infrared image | right infrared image | depth image in right RGB frame

Self-calibrated depth deformation



Benchmark vs. related datasets

Benchmark	Real data	RGB-D	Stereo	Global shutter	Sync'ed cameras	IMU	Accurate GT	Geometry GT	Benchmark
TUM RGB-D	✓	✓				(1)	✓		(2)
TUM VI	✓		✓	✓	✓	✓	✓		
TUM Mono	✓			✓	-				
CoRBS	✓	✓					✓	✓	
ICL-NUIM		✓		✓	✓		✓	(3)	
VIPER Odometry				✓	-		✓		✓
InteriorNet		✓	✓	✓	✓	✓	✓	✓	
KITTI Odometry	✓	(4)	✓	✓	✓	✓			✓
EuRoC MAV	✓		✓	✓	✓	✓	✓	(5)	
ETH3D SLAM	✓	✓	✓	✓	✓	✓	(6)		✓

☐ Highlighted: In contrast to other datasets, we provide a benchmark dataset with real-world RGB-D data, recorded with synchronized global-shutter cameras.

Notes: (1) Accelerometer but not gyroscope measurements are available. (2) While this dataset has a test set, it is not well suited for benchmarking since it shows the same scenes as the training set, and there is no online leaderboard. (3) Available in an extended version of the dataset. (4) Sparse measurements of a spinning laser scanner are available. (5) Structure ground truth is available for some of the sequences. (6) A motion capturing system is used for all but a few training datasets, for which ground truth is obtained using Structure-from-Motion.

Cost function & fast direct Bundle Adjustment scheme

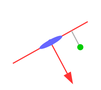
Total cost: $C(K,S) = \sum_{k \in K} \sum_{s \in S_k} \left(\rho_{\text{Tukey}} \left(\sigma_D^{-1} r_{\text{geom}}(s,k) \right) + w_{\text{photo}} \rho_{\text{Huber}} \left(\sigma_p^{-1} r_{\text{photo}}(s,k) \right) \right)$

Geometric residual: point-plane distance between surfel and measurement $r_{\text{geom}}(s,k) = \left(\mathbf{T}_G^k \mathbf{n}_s\right)^T \left(\pi_{D,k}^{-1}(\hat{\pi}_{D,k}(\mathbf{T}_G^k \mathbf{p}_s)) - \mathbf{T}_G^k \mathbf{p}_s\right)$

Photometric residual: descriptor difference between surfel and measurement

Gradient magnitude - based descriptor:

$$r_{\text{photo}}(s,k) = \left\| \begin{pmatrix} I(\pi_{I,k}(\mathbf{s}_1)) - I(\pi_{I,k}(\mathbf{p}_s)) \\ I(\pi_{I,k}(\mathbf{s}_2)) - I(\pi_{I,k}(\mathbf{p}_s)) \end{pmatrix} \right\|_2 - d_s$$



surfel position: $\mathbf{T}_G^k \mathbf{p}_s$
surfel normal: $\mathbf{T}_G^k \mathbf{n}_s$
measurement: $\pi_{D,k}^{-1}(\hat{\pi}_{D,k}(\mathbf{T}_G^k \mathbf{p}_s))$
result: $r_{\text{geom}}(s,k)$

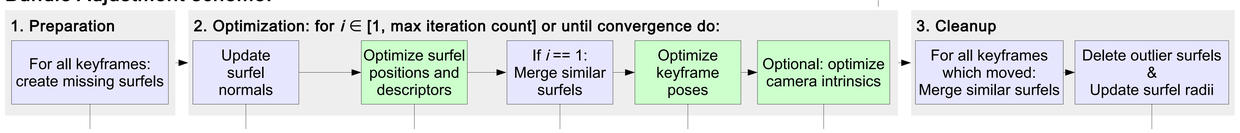
geometric residual sketch

surfel projection
center sample: $\pi_{I,k}(\mathbf{p}_s)$
border samples: $\pi_{I,k}(\mathbf{s}_1)$ & $\pi_{I,k}(\mathbf{s}_2)$

image

photometric residual sketch

Bundle Adjustment scheme:



Sparse surfel creation
Default: one surfel per 4x4 block of pixels

Treated as auxiliary variables
→ Fast update: Just average all observations

Optimized with Gauss-Newton
Position optimization restricted to surfel normal direction

In order to remove surfels created from outliers

Optimized with Gauss-Newton

Optimized with Gauss-Newton
Optimize for intrinsics and depth distortion

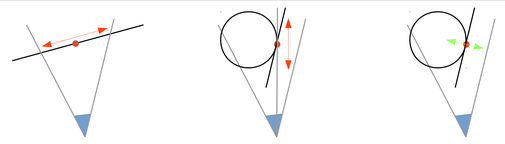
Update surfel sizes to the highest image resolution they are observed with

Legend: **Surfel updates** | **Continuous optimization**

Details: Surfel position parametrization

- Geometric & photometric residual do not always sufficiently constrain surfel positions (e.g., in planar white walls)
- Optimizing only a surfel's depth in its source keyframe seems intuitive, but yields a potentially ill-posed ray-plane intersection problem together with the geometric residual

→ Restrict surfel position to move along surfel normal directions



free optimization
→ may be unconstrained (e.g., for a planar white wall)

optimization along viewing direction
→ point-plane intersection may be ill-conditioned

optimization along surfel normal

Results

TUM RGB-D benchmark

	fr1/desk	fr2/xyz	fr3/office	avg. rank
BundleFusion	**1.6 (1)**	1.1 (3)	2.2 (4)	2.7 (2)
DVO SLAM	2.1 (5)	1.8 (6)	3.5 (8)	6.3 (6)
ElasticFusion	2.0 (4)	1.1 (3)	1.7 (2)	3.0 (4)
Kintinuous	3.7 (8)	2.9 (9)	3.0 (6)	7.7 (8)
MRSMap	4.3 (9)	2.0 (7)	4.2 (9)	8.3 (9)
ORB-SLAM2	**1.6 (1)**	**0.4 (1)**	**1.0 (1)**	**1.0 (1)**
PSM SLAM	**1.6**	-	3.1	-
RGB-D SLAM	2.3 (6)	0.8 (2)	3.2 (7)	5.0 (5)
VoxelHashing	2.3 (6)	2.2 (8)	2.3 (5)	6.3 (6)
Ours (fixed intr.)	3.6	1.2	2.5	-
Ours	1.7 (3)	1.1 (3)	1.7 (2)	2.7 (2)

Among the selected datasets, BAD SLAM shares the second average rank with BundleFusion. ORB-SLAM2 takes the first place clearly. We evaluate SE(3) ATE RMSE here.

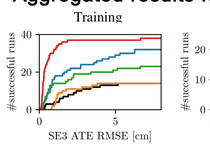
Impact of distortions

	clean		async		rs		async & rs	
	avg.	med.	avg.	med.	avg.	med.	avg.	med.
BundleFusion	0.34	0.22	1.10	1.14	1.10	1.02	1.48	1.40
DVO SLAM	0.32	0.23	2.33	0.72	5.10	1.37	4.94	1.39
ElasticFusion	1.11	0.90	1.98	1.17	2.70	1.77	3.19	2.52
ORB-SLAM2	0.47	0.30	0.60	0.40	3.25	1.57	3.49	1.55
Ours	**0.15**	**0.02**	**0.40**	**0.21**	**0.99**	**0.87**	**1.01**	**0.98**

We simulate asynchronous RGB-D frames (async) and rolling shutter (rs) in synthetic datasets. Both effects significantly affect the methods. We evaluate SE(3) ATE RMSE here.

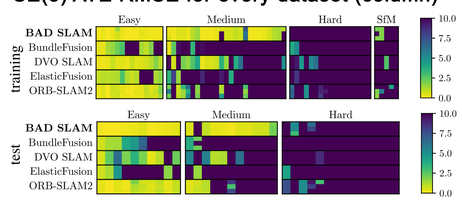
Evaluation on our benchmark

Aggregated results for the whole dataset

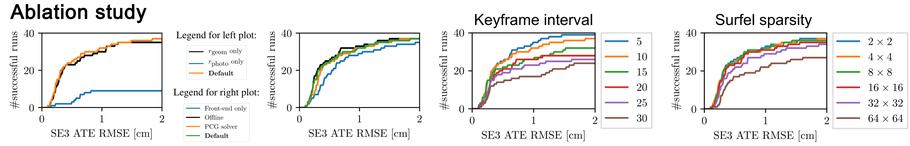


→ BAD SLAM performs best, while DVO SLAM performs on a similar level as ORB-SLAM2

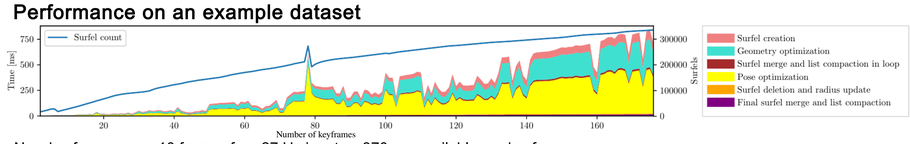
SE(3) ATE RMSE for every dataset (column)



Ablation study



Performance on an example dataset



New keyframe every 10 frames for ~27 Hz input → 370 ms available per keyframe

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

Computer Vision and Geometry Lab

Acknowledgments:
Thomas Schöps was supported by a Google PhD Fellowship.