A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos

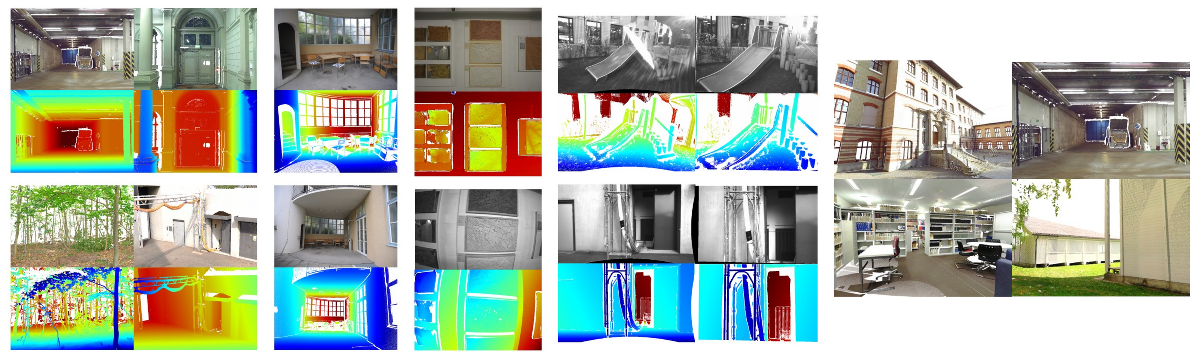
www.eth3d.net

Thomas Schöps[1], Johannes L. Schönberger[1], Silvano Galliani[2], Torsten Sattler[1], Konrad Schindler[2], Marc Pollefeys[1,4], Andreas Geiger[1,3]

[1]Department of Computer Science, ETH Zürich [2]Institute of Geodesy and Photogrammetry, ETH Zürich [3]Autonomous Vision Group, MPI for Intelligent Systems, Tübingen [4]Microsoft, Redmond

Key features

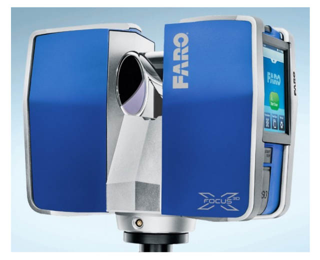
- **High-resolution DSLR images** (24 Mpx), available as RAW
- **Hand-held multi-camera videos** cover the use-case of reconstruction on mobile devices
- **Diverse set of scenes,** indoors and outdoors, 6 DOF camera motion
- **Online evaluation** on training & test set. **Visualizations** of 3D point clouds and depth maps facilitate easy comparison (may not work with mobile browsers)



Hardware

Laser scanner — **Faro Focus X 330**

- Colored 360° scans from static position
- Configured for up to 28M pts / scan
- ~9 minutes / scan

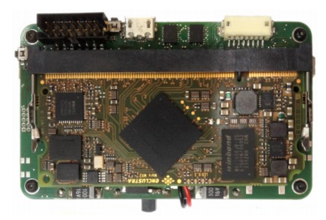


DLSR camera — **Nikon D3X**

- 6048 x 4032 resolution
- Lens with ~85° FOV
- RAW images, mostly fixed intrinsics

Multi-camera rig — **[3]**

- 752 x 480 resolution
- 2 stereo pairs, with ~54° resp. ~83° FOV
- ~13.6 Hz synchronized recording
- Global shutter, automatic exposure

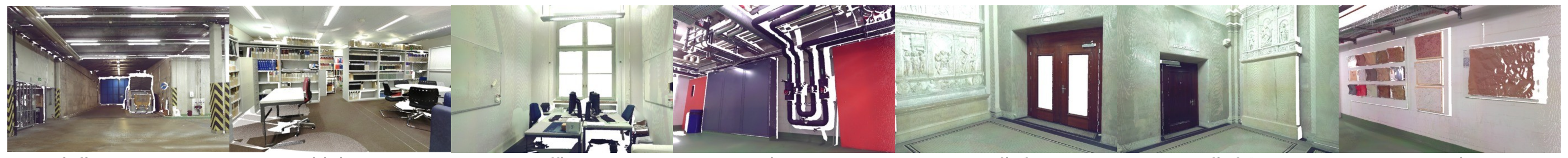


Benchmark tasks & Evaluation

Outdoor scenes

courtyard, electro, facade, forest, meadow, playground, terrace

Indoor scenes

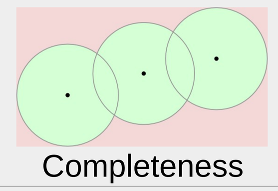


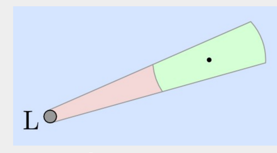
delivery area, kicker, office, pipes, relief, relief 2, terrains

High-resolution multi-view stereo on DSLR images	13 datasets, 454 images in total
Low-resolution many-view stereo on camera rig videos	5 datasets, 4796 images in total
Low-resolution two-view stereo on rig camera pairs	16 frames, 64 images in total

Evaluation of reconstructed point clouds (for a given distance threshold):

- Compare to laser scan points seen in ≥ 2 images
- Compute accuracy and completeness
- F-score (harmonic mean) as single ranking metric





Evaluation: equal to the popular Middlebury two-view stereo benchmark [5]

Comparison to related works

Benchmark	Setting	Resolution	Online Eval.	6DoF Motion	MVS	Stereo	Video	Varying FOV	Website
Middlebury MVS	Laboratory	0.3 Mpx	✓		✓				http://vision.middlebury.edu/mview/
Middlebury	Laboratory	6 Mpx	✓			✓			http://vision.middlebury.edu/stereo/
DTU	Laboratory	2 Mpx			✓				http://roboimagedata.compute.dtu.dk/?page_id=36
MPI Sintel	Synthetic	0.4 Mpx	✓	✓		✓	✓		http://sintel.is.tue.mpg.de/
KITTI	Street scenes	0.5 Mpx	✓		✓	✓	✓		www.cvlibs.net/datasets/kitti/
Strecha	Buildings	6 Mpx		✓	✓				http://cvlabwww.epfl.ch/data/multiview
Tanks and Temples (*)	Varied	8 Mpx	✓	✓	✓		✓		www.tanksandtemples.org
ETH3D	Varied	0.4 / 24 Mpx	✓	✓	✓	✓	✓	✓	www.eth3d.net

(*) Created simultaneously to ours. Evaluates complete reconstruction pipelines and thus does not include ground truth for image poses.

Relative rankings (scores) on MVS (left) and two-view stereo (right) benchmarks. Default parameters were used on ours. Metrics: bad 2 non-occ (stereo), F-score (ours MVS), see [6] (Strecha).

Method	Strecha	Ours
PMVS	3 (68.9)	3 (41.2)
Gipuma	4 (48.8)	4 (33.2)
COLMAP	2 (75.9)	**1 (64.7)**
CMPMVS	**1 (78.2)**	2 (48.9)

Method	Middle.	KITTI	Ours
SPS-Stereo	5 (29.3)	2 (5.3)	**1 (3.4)**
MeshStereo	2 (14.9)	4 (8.4)	3 (7.1)
SGM+D.	4 (29.2)	3 (6.3)	2 (5.5)

Method	Middle.	KITTI	Ours
MC-CNN	**1 (10.1)**	**1 (3.9)**	4 (8.9)
ELAS	3 (25.7)	5 (9.7)	5 (10.5)

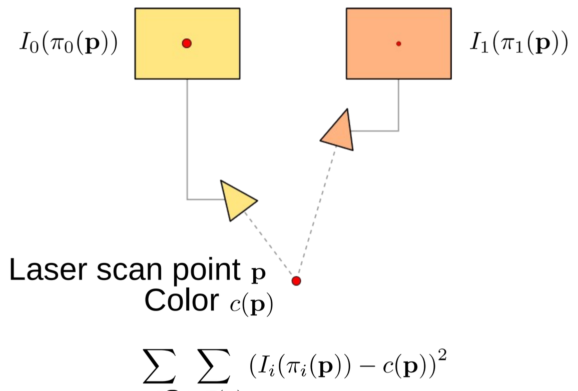
Registration pipeline

1. Preprocessing: Laser scan cleanup (auto. + manual) → Image masking

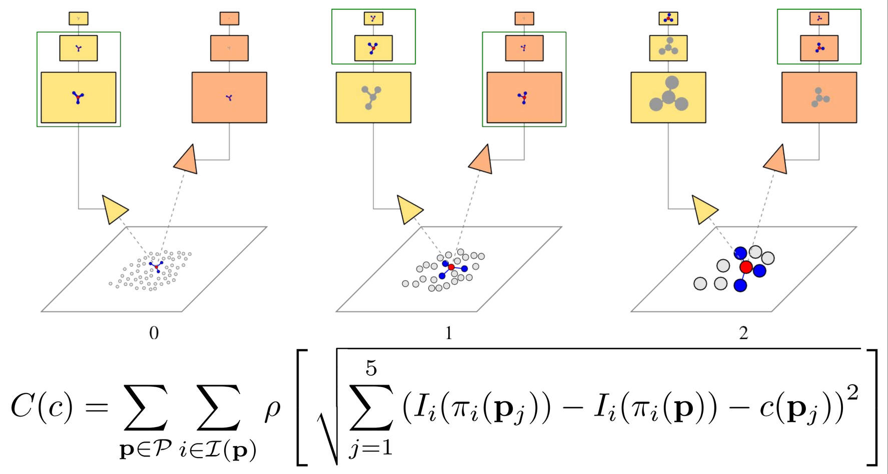
2. Initial registration: Laser scan cube map rendering → Joint SfM reconstruction with scale

3. Registration refinement: Laser scan refinement (ICP) → Image and intrinsics refinement

- Extension of [7]:



$$\sum_{\mathbf{p}\in\mathcal{P}}\sum_{i\in\mathcal{I}(\mathbf{p})}(I_i(\pi_i(\mathbf{p}))-c(\mathbf{p}))^2$$

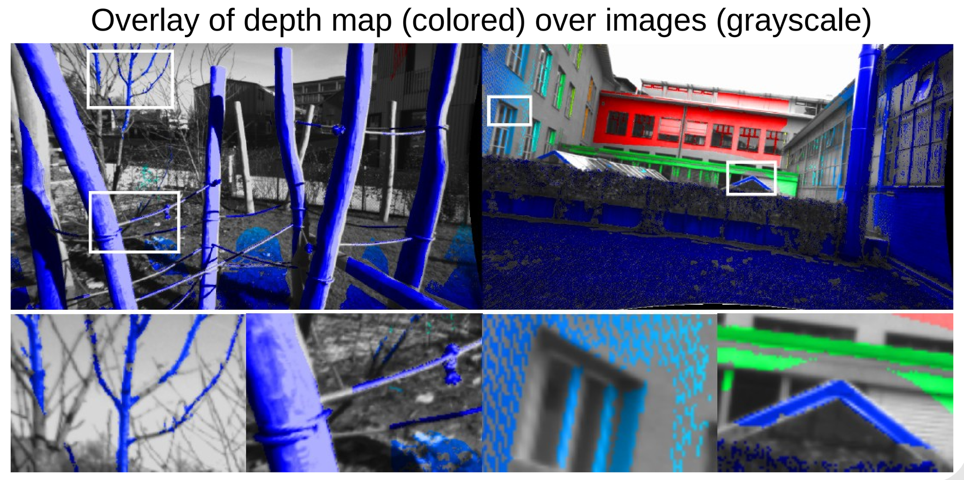


$$C(c)=\sum_{\mathbf{p}\in\mathcal{P}}\sum_{i\in\mathcal{I}(\mathbf{p})}\rho\left[\sqrt{\sum_{j=1}^{5}\left(I_i(\pi_i(\mathbf{p}_j))-I_i(\pi_i(\mathbf{p}))-c(\mathbf{p}_j)\right)^2}\right]$$

- Occlusion handling using reconstructed mesh
- Use of image and scan colors
- Use gradients instead of raw pixel intensities
- Multi-resolution cost term handles scale differences

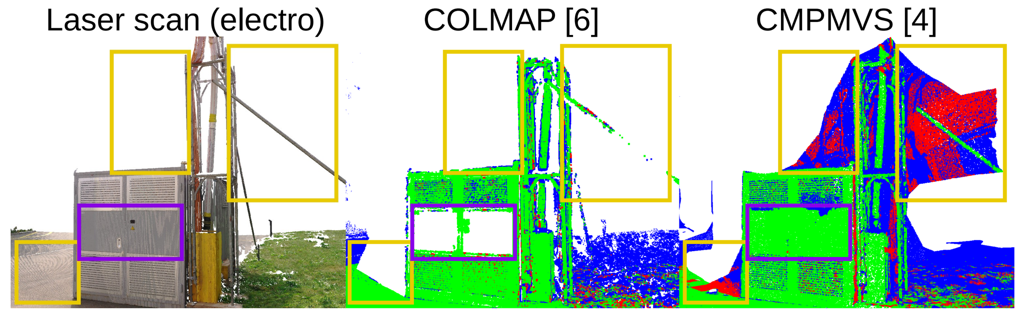
Qualitative evaluation

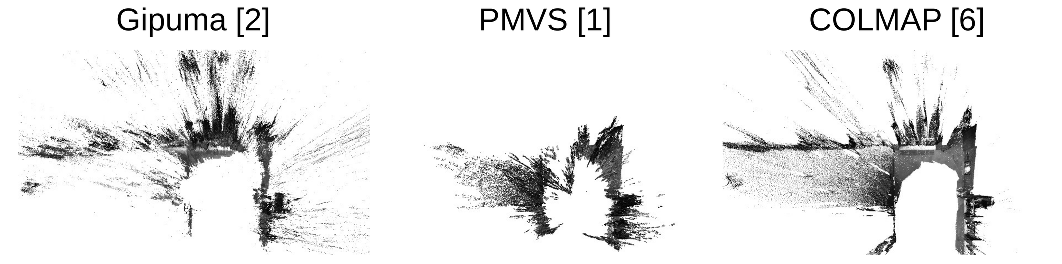
Overlay of depth map (colored) over images (grayscale)



Open challenges

- **Weakly textured surfaces**
- **Runtime**
- **Properly exploiting high view redundancy**

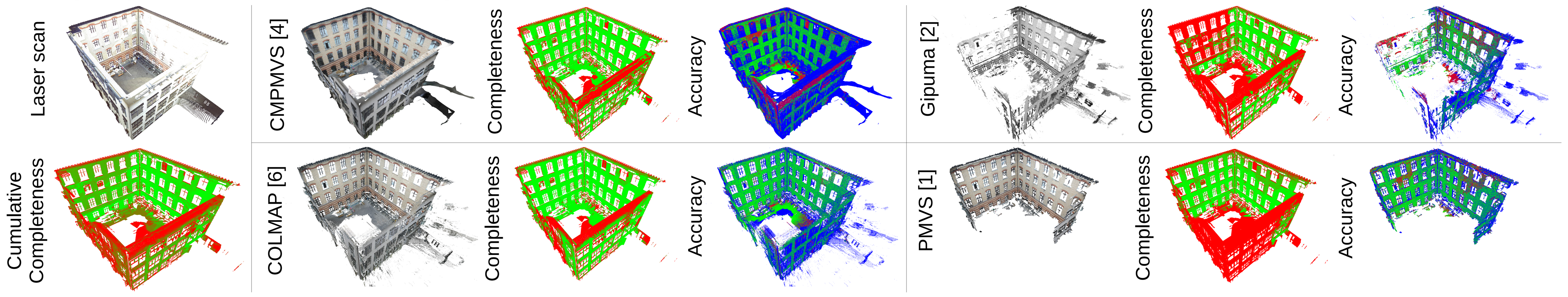




Results

See www.eth3d.net for the current leaderboards!

Example qualitative results on courtyard DSLR dataset



Initial leaderboards. Left: multi-view (accuracy/completeness/F-score), right: two-view. All methods use default parameters.

Method	Indoor	Outdoor	Mobile	DSLR
CMPMVS	67.2 / 47.3 / 55.5	44.2 / 40.0 / 42.0	14.4 / 7.4 / 9.8	71.6 / **57.6** / 63.8
COLMAP	**90.2 / 51.1 / 65.2**	**80.9 / 53.1 / 64.1**	**69.5 / 41.2 / 51.8**	**91.7** / 56.2 / **69.7**
Gipuma	74.9 / 24.0 / 36.3	52.8 / 20.8 / 29.9	31.1 / 13.4 / 18.7	76.5 / 25.9 / 38.7
PMVS	85.1 / 28.0 / 42.1	72.2 / 27.8 / 40.1	48.7 / 18.8 / 27.2	90.1 / 31.3 / 46.5

Method	bad 0.5	bad 1	bad 2	bad 4	avgerr	rms	A50	A90	A95	A99
SPS-Stereo	56.91	21.29	**3.43**	**1.43**	**0.83**	**1.61**	2.22	**1.36**	**2.11**	**6.52**
SGM+D.	57.79	22.43	5.48	2.65	1.03	2.43	**1.14**	7.05	6.46	10.69
MeshStereo	**28.99**	**13.23**	7.09	4.38	0.87	2.18	1.61	2.55	4.46	10.65
MC-CNN	32.51	13.85	8.92	8.59	8.48	17.03	49.01	27.63	28.79	45.79
ELAS	41.20	21.56	10.50	4.56	1.06	3.03	4.54	3.98	5.24	9.68

References

[1] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. PAMI, 32(8):1362–1376, 2010.
[2] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In ICCV, 2015.
[3] P. Gohl, D. Honegger, S. Omari, M. Achtelik, M. Pollefeys, and R. Siegwart. Omnidirectional Visual Obstacle Detection using Embedded FPGA. In IROS, 2015.
[4] M. Jancosek, and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In CVPR, 2011.
[5] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In GCPR, 2014.
[6] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In ECCV, 2016.
[7] Q. Zhou and V. Koltun. Color map optimization for 3d reconstruction with consumer depth cameras. In SIGGRAPH, 2014.

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich
CVG Computer Vision and Geometry Lab
PRS Photogrammetry Remote Sensing

Acknowledgments:
Thomas Schöps was supported by a Google PhD Fellowship. This project received funding from the European Union's Horizon 2020 research and innovation programme under grant No. 688007 (Trimbot2020). This project was partially funded by Google Tango.