Dense 3D Reconstruction for Mixed Reality in Medical Training: Classical methods vs Deep Learning

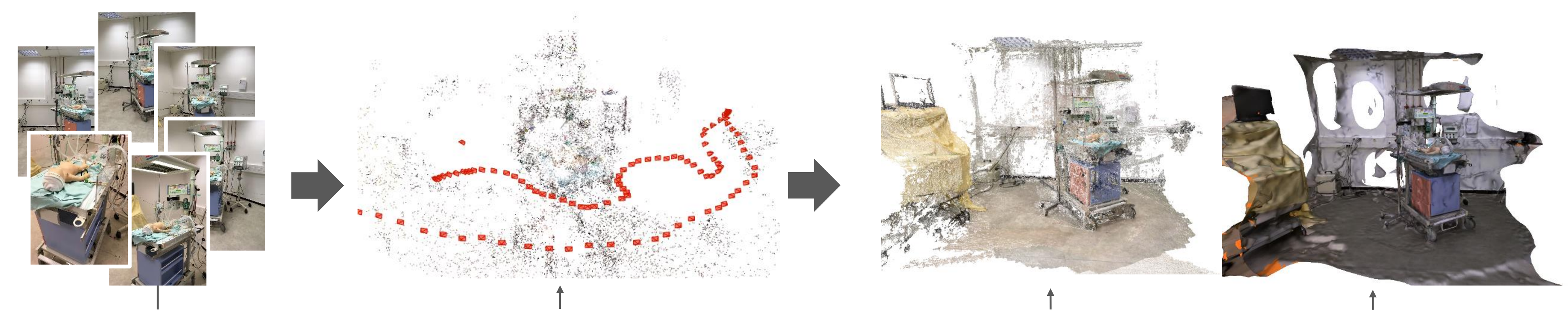
<u>Kristina PROKOPETC</u> and Romain DUPONT

LVML - Laboratory for Vision, Modeling and Localization, CEA List Paris-Saclay, France

1 - MOTIVATION AND OBJECTIVES

- **Learning by simulation** is an educational approach that is widely adopted in medical practice including neo-natal reanimation training.
- **Recent solutions employ Virtual Reality (VR)** alongside with traditional techniques to enrich the experience but it has **two major shortcomings**: 1) the non-tangible side of the VR often bothers learners and requires a pre-learning phase and 2) the risk of motion sickness.
- We work on a **Mixed Reality (MR)** system that **solves VR-related problems and leverages all simulation types** using modern computer vision techniques where **Dense 3D Reconstruction is a key-component.**
- **Deep learning solutions for Dense 3D have high claims against analytical methods**, **yet the gap in performance is ambiguous.** To this end, **we propose a comparative study on challenging scenario of MR in medical scenes.**

2 - DENSE 3D RECONSTRUCTION PIPELINE

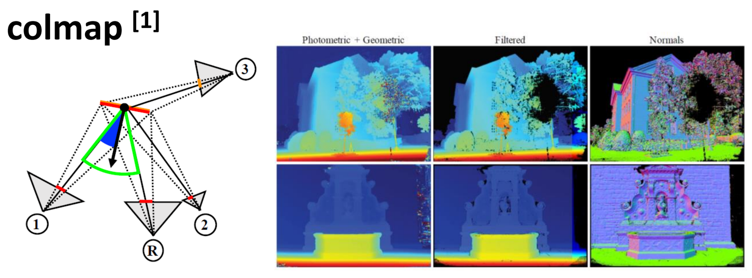


	Initialization	Visibility Model	Dense Correspondence and Depth Estimation	Depth refinement and Fusion	Other
Classical MVS	high resolution images, known cameras, sparse point cloud, depth range	handcrafted stereo-pair selection + geometric priors	patch-based plane estimation with heuristics + handcrafted spatial and temporal propagation (PatchMatch Stereo[6], PlaneSweep Stereo[7]) + handcrafted similarity metrics	handcrafted geometric consistency checks + use of surface normals	requires no training or up-sampling
Learned MVS	similar to classical methods + Ground Truth depth	similar to classical methods	learned (patch-based) stereo and semantic features + fronto-parallel plane-sweeping + depth regression from temporal feature volumes (multi-class classification problem)	residual learning or probability-based filtering, no surface normals	real+ synthetic data, may need up-sampling

3 - THE COMPARATIVE STUDY

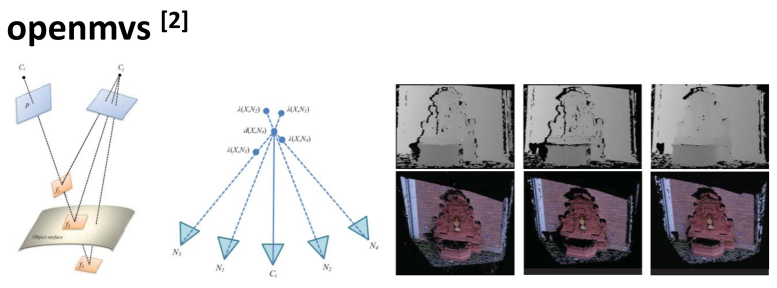
Classical MVS

colmap [1]



Pixelwise view selection with geometric priors (resolution , triangulation angle, incident angle) for reference (R) and three source views (1-3) and temporal smoothness term + examples of depth and normal maps

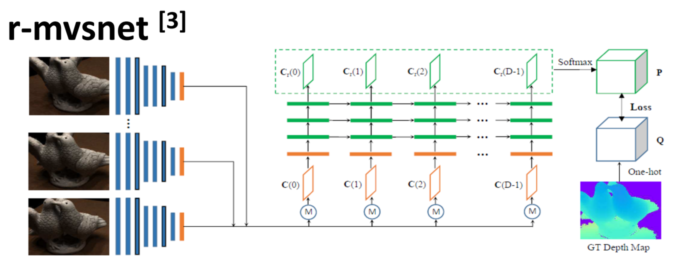
openmvs [2]



Pixelwise patch-based plane estimation with minimal aggregated matching cost where a support plane is represented in 3D + depth filtering with back-projection checks + depthmaps and back-projected 3D points

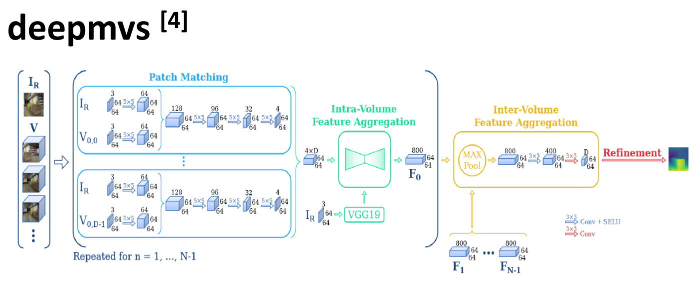
Learned MVS

r-mvsnet [3]



Deep feature extraction + fronto-parallel warping w.r.t reference camera + multi-depth cost map computation and recurrent regularization = classification problem with the cross-entropy loss

deepmvs [4]



PatchMatch network for feature extraction on fronto-parallel planes + semantic features + spatial and temporal feature aggregation + residual filtering = classification with the cross-entropy loss

WHAT IS BETTER?

Classical MVS

- **Generally perform better** but often **more slow** due to more complex optimization schemes
- Success is driven by careful choice of heuristics and several pivotal ideas
- **Rely heavily on Lambertian surface assumption** but **can handle slanted surface** orientation
- **Provide more detailed reconstructions** which are favorable for Mixed Reality
- **Rarely limited by number of images, their size or computational power**
- Safe choice but **difficult to improve more!**

Learned MVS

- **Do not generalize well** (require fine-tuning) and **not necessarily faster**
- Driven by pivotal ideas from classical methods but **not everything is learned (however, it can be!)**
- **Semantic features relax some appearance constraints** but still **assume fronto-parallel scene**
- Often **lack fine details** due to reduced resolution and other factors (e.g. view selection priors)
- **Often restrained by GPU limits and require up-sampling of the output**
- **Have great potential for improvement!**

DATASETS AND EVALUATION PROTOCOL

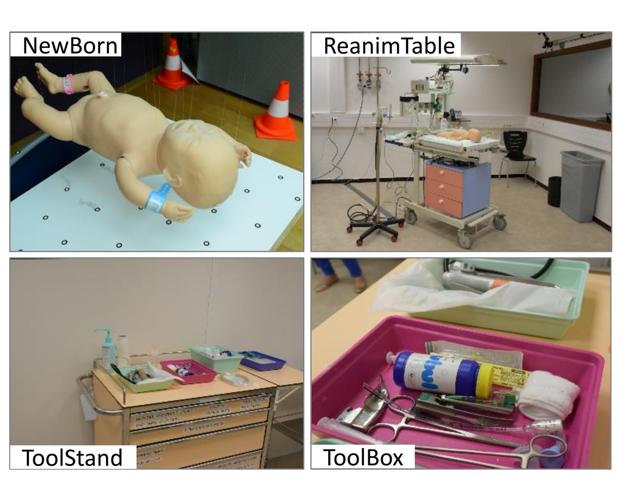
Datasets acquisition

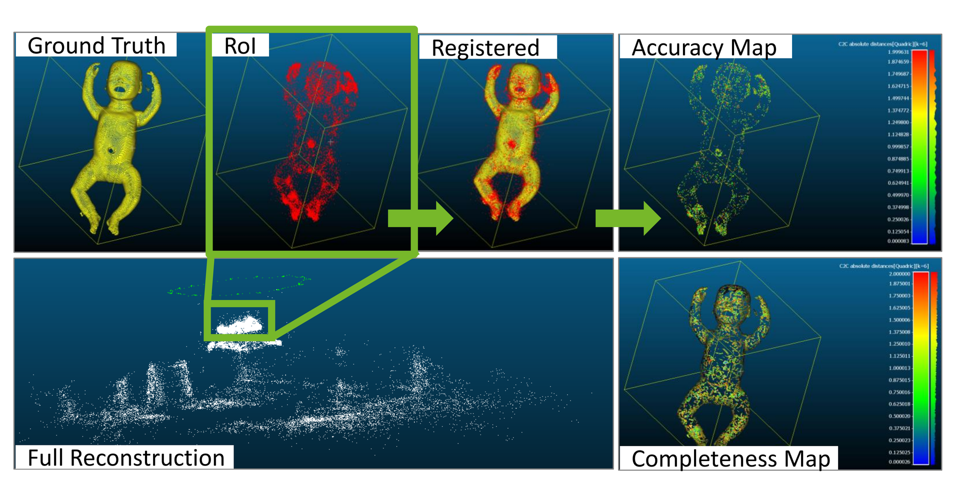
- different sensors (DSLR, iPhone)
- 6DoF motion, different Nº images
- 1152x864 resolution
- sparse initialization via colmap [1]

Ground Truth

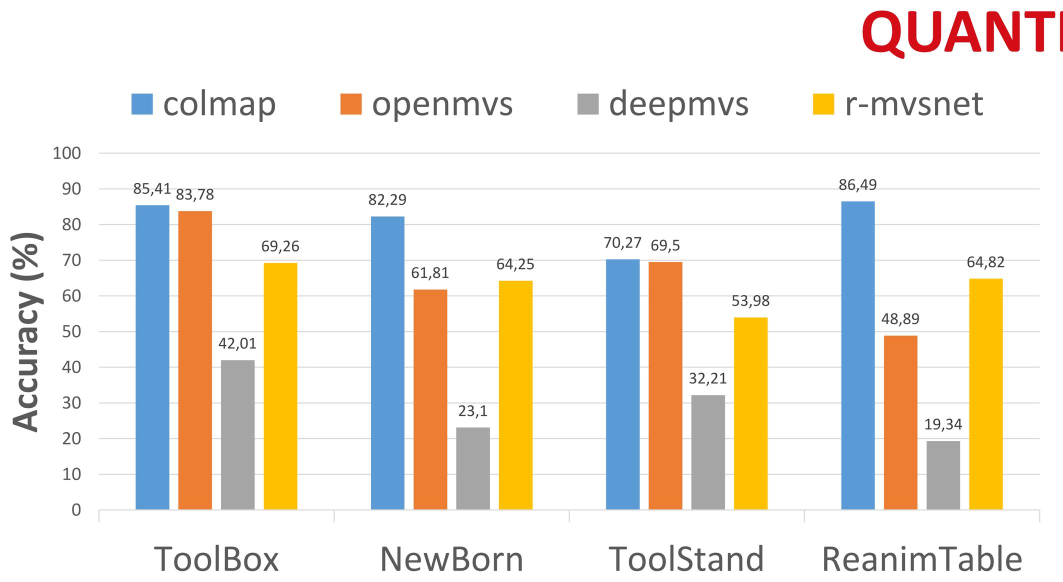
- high-precision laser scanner
- dense point cloud from mesh
- not always complete

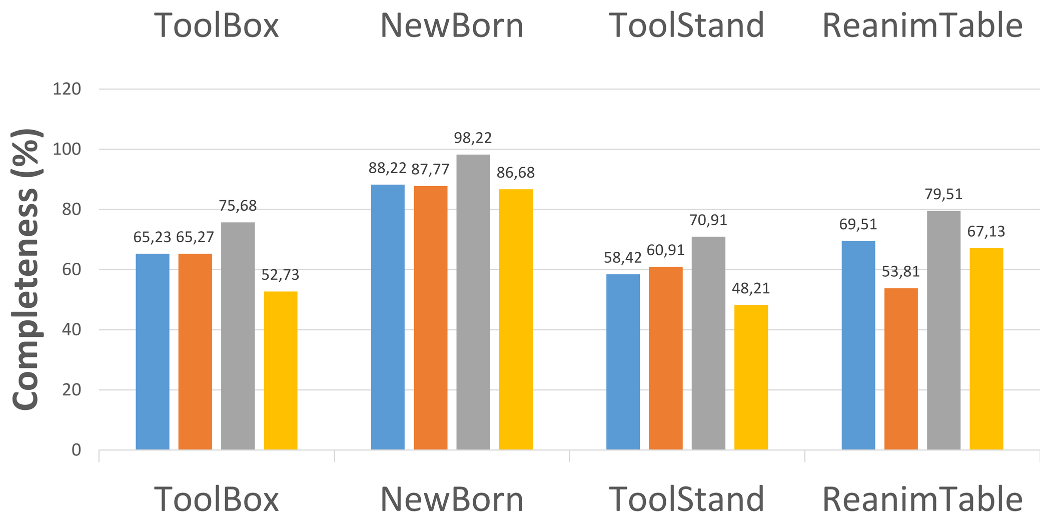
Evaluation as in ETH3D Benchmark [5]





QUANTITATIVE RESULTS



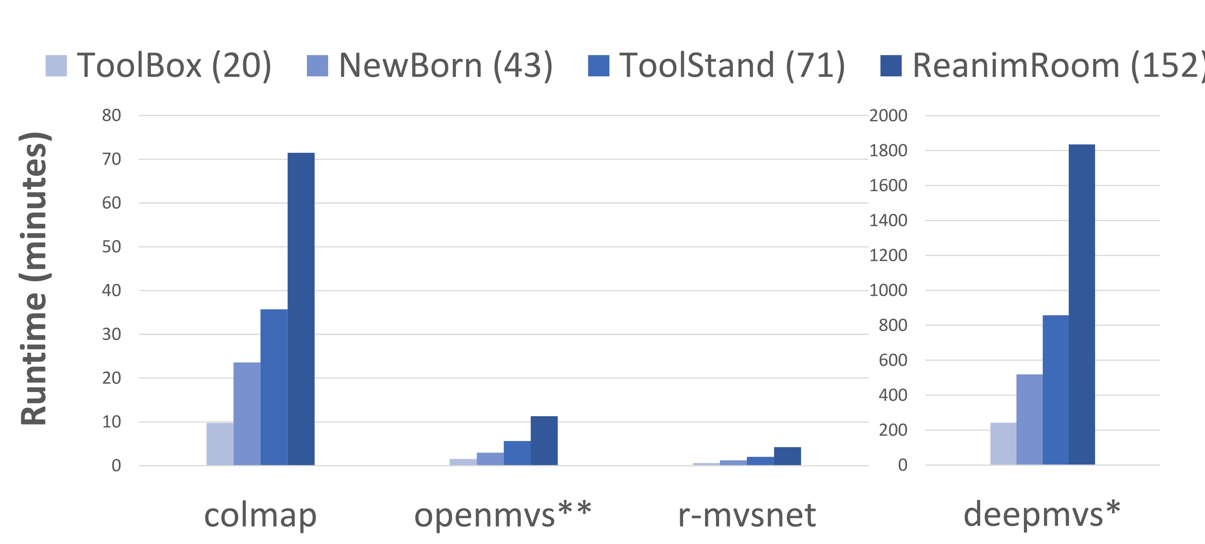


Accuracy : A fraction of the *Reconstruction* which is closer than <u>2mm</u> to the *Ground Truth*.

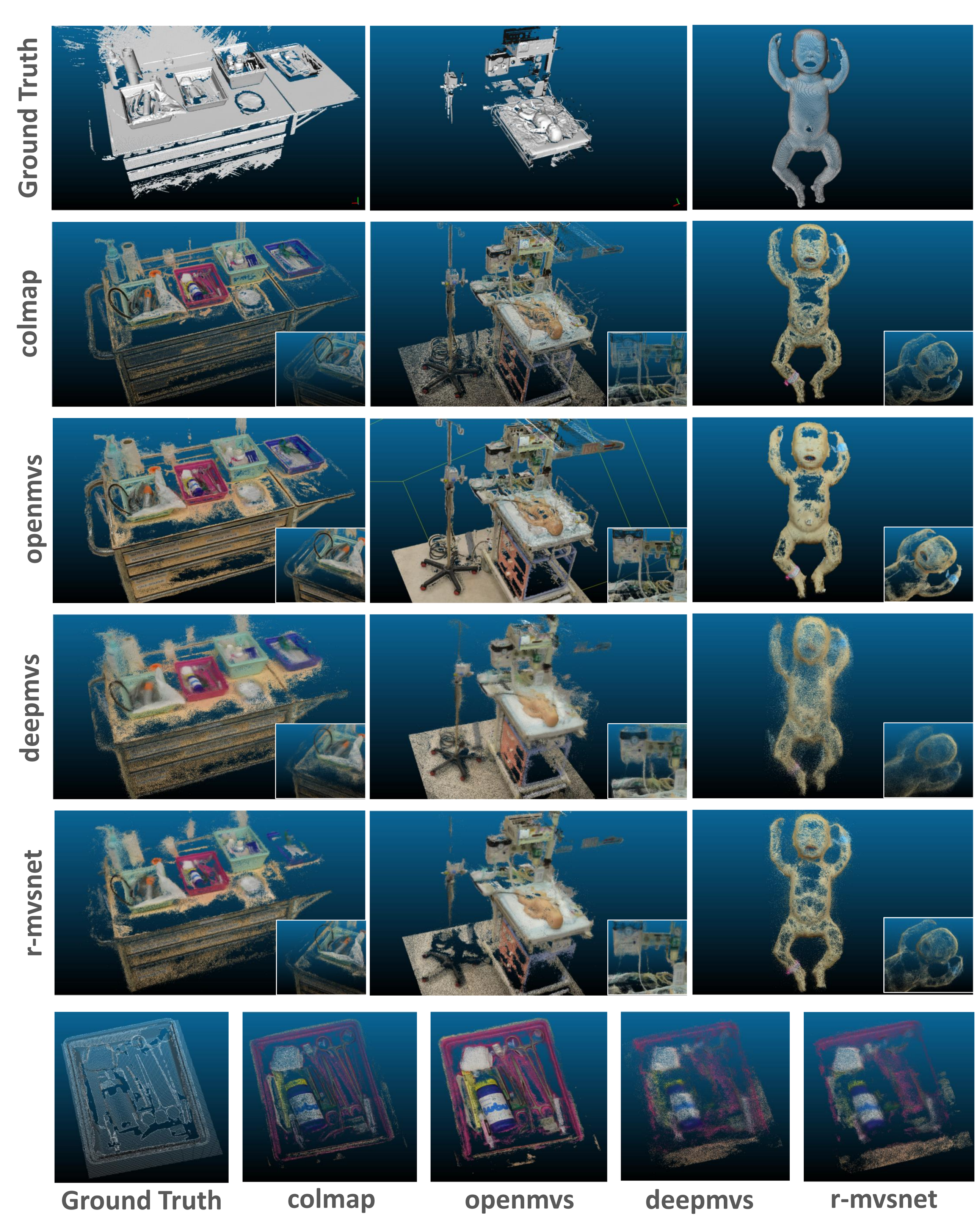
Completeness : A fraction of the *Ground Truth* which is closer than <u>2mm</u> to the *Reconstruction*.

Runtime : All methods run on the computer with AMD CPU with 16 cores x32, 32GB RAM, GeForce GTX 1080 ti 11Gb GPU

* partial execution on CPU and GPU
** multithreaded CPU only (20) - number of images in a dataset



QUALITATIVE RESULTS



[1] Schönberger, J.L., et al. "Pixelwise view selection for unstructured multi-view stereo". In *European Conference on Computer Vision* (2016).
[2] Shen, S. "Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes". *IEEE transactions on image processing* (2013).
[3] Yao, Y., et al. "Recurrent MVSNet for High-resolution Multi-view Stereo Depth Inference". In *Int. Conf. on Computer Vision and Pattern Recognition* (2019).
[4] Huang, P.H., et al. "Deepmvs: Learning multi-view stereopsis". In *Int. Conf. on Computer Vision and Pattern Recognition* (2018).
[5] Schops, T., et al. "A multi-view stereo benchmark with high-resolution images and multi-camera videos". In *Int. Conf. on Computer Vision and Pattern Recognition* (2017).
[6] Bleyer, M., et al. "PatchMatch Stereo - stereo matching with slanted support windows". In *British Machine Vision Conference* (2011).
[7] Gallup, D., et al. "Real-time plane-sweeping stereo with multiple sweeping directions". In *Int. Conf. on Computer Vision and Pattern Recognition* (2007).