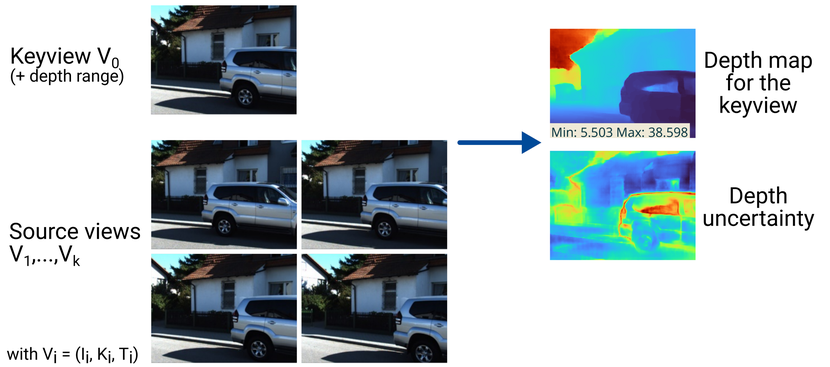
A Benchmark and a Baseline for Robust Multi-view Depth Estimation

Philipp Schröppel Jan Bechtold Artemij Amiranashvili Thomas Brox

schroepp@cs.uni-freiburg.de

3DV 2022

Task: Multi-view Depth Estimation

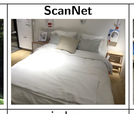


Given multiple views of a scene, estimate depth map for the keyview.

Robust Multi-view Depth Benchmark

Benchmark for multi-view depth estimation with a **focus on robust application to arbitrary real-world data**:

- **test sets based on multiple diverse existing datasets**:

Test set	KITTI	ScanNet	ETH3D	DTU	T&T
					
domain	driving	indoor	in- & outdoor	tabletop	in- & outdoor
Structure	Video	Video	None	None	None
scene scale	2 – 85 m	0.2 – 9 m	0.3 – 60 m	0.4 – 1.2 m	1.1 – 42 m
# samples	93	200	104	110	69

- training set intentionally left undefined
  → **evaluation in a zero-shot cross-dataset fashion**

Benchmark features **multiple evaluation settings**:

- **input modalities**: images, intrinsics, ground truth poses, ground truth depth range
- optional **alignment** between predicted and ground truth depths

---

Benchmark **evaluates**:

- **depth estimation performance**:
  - Absolute Relative Error (rel)
  - Inlier Ratio with a threshold of 1.03 (τ)
- **uncertainty estimation performance**:
  - Sparsification error curves
  - Area Under Sparsification Error (AUSE)
- performance depending on the number of source view

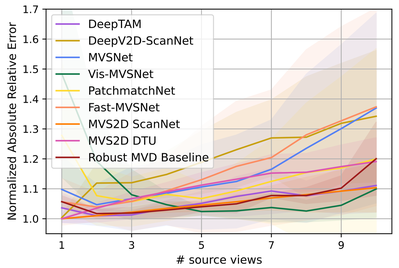
Results on the Robust MVD benchmark

We evaluate state-of-the-art models on the benchmark with different settings:
a) classical approaches
b) no poses, no depth range, with alignment (depth-from-video)
c) with poses, with depth range, no alignment (multi-view stereo)
d) with poses, no depth range, no alignment (absolute scale)

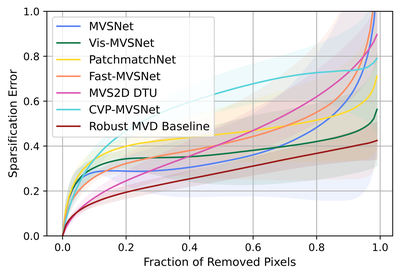
Approach	GT Poses	GT Range	Align	KITTI rel ↓	KITTI τ ↑	ScanNet rel ↓	ScanNet τ ↑	ETH3D rel ↓	ETH3D τ ↑	DTU rel ↓	DTU τ ↑	T&T rel ↓	T&T τ ↑	Average rel ↓	Average τ ↑	Average time [s] ↓
a)																
Colmap	✓	✗	✗	**12.0**	**58.2**	**14.6**	**34.2**	**16.4**	**55.1**	**0.7**	**96.5**	**2.7**	**95.0**	**9.3**	**67.8**	≈ 3 min
Colmap Dense	✓	✗	✗	26.9	52.7	38.0	22.5	89.8	23.2	20.8	69.3	25.7	76.4	40.2	48.8	≈ 3 min
b)																
DeMoN	✗	✗	$\|\mathbf{t}\|$	15.5	15.2	**12.0**	**21.0**	17.4	15.4	21.8	16.6	13.0	23.2	16.0	18.3	**0.08**
DeepV2D KITTI	✗	✗	med	(3.1)	(74.9)	23.7	11.1	27.1	10.1	24.8	8.1	34.1	9.1	22.6	22.7	2.07
DeepV2D ScanNet	✗	✗	med	**10.0**	**36.2**	(4.4)	(54.8)	**11.8**	**29.3**	**7.7**	**33.0**	**8.9**	**46.4**	**8.6**	**39.9**	3.57
c)																
MVSNet	✓	✓	✗	22.7	36.1	24.6	20.4	35.4	31.4	(1.8)	(86.0)	8.3	73.0	18.6	49.4	0.07
MVSNet Inv. Depth	✓	✓	✗	18.6	30.7	22.7	20.9	21.6	35.6	(1.8)	(86.7)	6.5	74.6	14.2	49.7	0.32
CVP-MVSNet	✓	✓	✗	156.7	2.2	137.1	15.9	156.4	13.6	(4.0)	(68.4)	24.7	52.9	95.8	30.6	0.49
Vis-MVSNet	✓	✓	✗	**9.5**	**55.4**	8.9	33.5	**10.8**	**43.3**	(1.8)	(87.4)	**4.1**	**87.2**	**7.0**	**61.4**	0.70
PatchmatchNet	✓	✓	✗	10.8	45.8	**8.5**	**35.3**	19.1	34.8	(2.1)	(82.8)	4.8	82.9	9.1	56.3	0.28
Fast-MVSNet	✓	✓	✗	14.4	37.1	17.0	24.6	25.2	32.0	(2.5)	(81.8)	8.3	68.6	13.5	48.8	0.30
MVS2D ScanNet	✓	✓	✗	21.2	8.7	(27.2)	(5.3)	27.4	4.8	17.2	9.8	29.2	4.4	24.4	6.6	**0.04**
MVS2D DTU	✓	✓	✗	226.6	0.7	32.3	11.1	99.0	11.6	(3.6)	(64.2)	25.8	28.0	77.5	23.1	0.05
d)																
DeMoN	✓	✗	✗	16.7	13.4	75.0	0.0	19.0	16.2	23.7	11.5	17.6	18.3	30.4	11.9	0.08
DeepTAM	✓	✗	✗	68.7	0.4	(6.7)	(39.7)	20.4	19.8	58.0	9.1	40.0	12.9	38.8	16.4	0.85
DeepV2D KITTI	✓	✗	✗	(20.4)	(16.3)	25.8	8.1	30.1	9.4	24.6	8.2	38.5	9.6	27.9	10.3	1.43
DeepV2D ScanNet	✓	✗	✗	61.9	5.2	(**3.8**)	(**60.2**)	18.7	28.7	9.2	27.4	33.5	38.0	25.4	31.9	2.15
MVSNet	✓	✗	✗	14.0	35.8	1568.0	5.7	507.7	8.3	(4429.1)	(0.1)	118.2	50.7	1327.4	20.1	0.15
MVSNet Inv. Depth	✓	✗	✗	29.6	8.1	65.2	28.5	60.3	5.8	(28.7)	(48.9)	51.4	14.6	47.0	21.2	0.28
CVP-MVSNet	✓	✗	✗	158.2	1.2	2289.0	0.1	1735.3	1.2	(8314.0)	(0.0)	415.9	9.5	2582.5	2.4	0.50
Vis-MVSNet	✓	✗	✗	10.3	**54.4**	84.9	15.6	51.5	17.4	(374.2)	(1.7)	21.1	65.6	108.4	31.0	0.82
PatchmatchNet	✓	✗	✗	29.0	16.3	70.1	16.7	99.4	3.5	(82.6)	(5.6)	39.4	19.3	64.1	12.3	0.18
Fast-MVSNet	✓	✗	✗	12.1	37.4	287.1	9.4	131.2	9.6	(540.4)	(1.9)	33.9	47.2	200.9	21.1	0.35
MVS2D ScanNet	✓	✗	✗	73.4	0.0	(4.5)	(54.1)	30.7	14.4	5.0	57.9	56.4	11.1	34.0	27.5	**0.05**
MVS2D DTU	✓	✗	✗	93.3	0.0	51.5	1.6	78.0	0.0	(**1.6**)	(**92.3**)	87.5	0.0	62.4	18.8	0.06
Robust MVD Baseline	✓	✗	✗	**7.1**	41.9	**7.4**	**38.4**	**9.0**	**42.6**	**2.7**	**82.0**	**5.0**	**75.1**	**6.3**	**56.0**	0.06

bold: best results (parentheses): trained on data from the same domain

Effect of the number of source views:



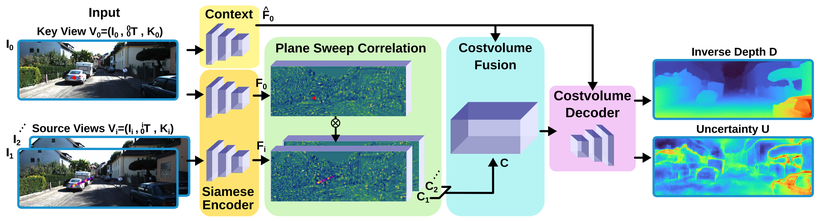
Uncertainty evaluation:



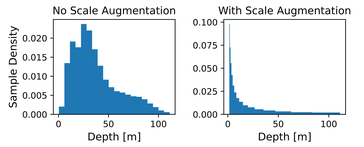
Findings

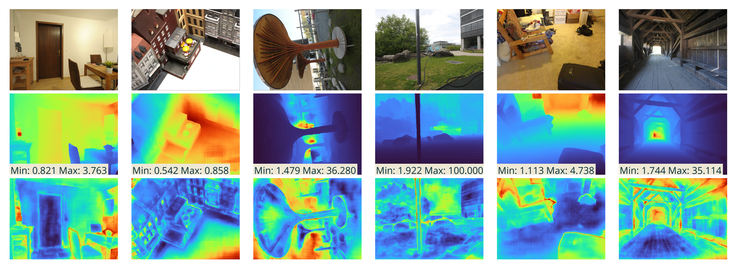
- learned models perform significantly better on training domain (see table)
- models perform significantly worse in the absolute scale setting (see table)
- most multi-view fusion strategies are suboptimal (left plot)
- suboptimal alignment between estimated uncertainties and actual errors (right plot)
- Robust MVD Baseline model shows more consistent performance across test sets and works in the absolute scale setting (see table)

Robust MVD Baseline Model



- **Network architecture**: based on DispNet
- **Training data**: StaticThings3D + BlendedMVS
- Novel data augmentation: **Scale augmentation**





Usage of the Robust MVD Benchmark

We provide code to use the benchmark at:
https://github.com/lmb-freiburg/robustmvd
→ dataloaders for all test sets
→ evaluation code
→ code to run all evaluated models
→ leaderboard coming soon

Link to code

Summary

→ We **show problems of current multi-view depth models**: cross-domain generalization, multi-view fusion, uncertainty estimation
→ We introduce a **benchmark to improve upon these problems**
→ The benchmark is complimentary to existing benchmarks and we **encourage future work on depth-from-video or multi-view stereo to additionally evaluate on the Robust MVD benchmark**
→ The **Robust MVD Baseline model can be used as baseline** and for robust multi-view depth estimation in applications where camera poses are known

Acknowledgements

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Climate Action within the project "KI Delta Learning" (Förderkennzeichen 19A19013N). The authors would like to thank the consortium for the successful cooperation. Funded by the Deutsche Forschungsgemeinschaft (DFG) - 417962828.