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

Philipp Schröppel
University of Freiburg

Jan Bechtold
Bosch

Artemij Amiranashvili
University of Freiburg

Thomas Brox
University of Freiburg

Abstract

Recent deep learning approaches for multi-view depth estimation are employed either in a depth-from-video or a multi-view stereo setting. Despite different settings, these approaches are technically similar: they correlate multiple source views with a keyview to estimate a depth map for the keyview. In this work, we introduce the Robust Multi-view Depth Benchmark that is built upon a set of public datasets and allows evaluation in both settings on data from different domains. We evaluate recent approaches and find imbalanced performances across domains. Further, we consider a third setting where camera poses are available and the objective is to estimate the corresponding depth maps with their correct scale. We show that recent approaches do not generalize across datasets in this setting. This is because their cost volume output runs out of distribution. To resolve this, we present the Robust MVD Baseline model for multi-view depth estimation, which is built upon existing components but employs a novel scale augmentation procedure. It can be applied for robust multi-view depth estimation, independent of the target data. We provide code for the proposed benchmark and baseline model at https://github.com/lmb-freiburg/robustmvd.

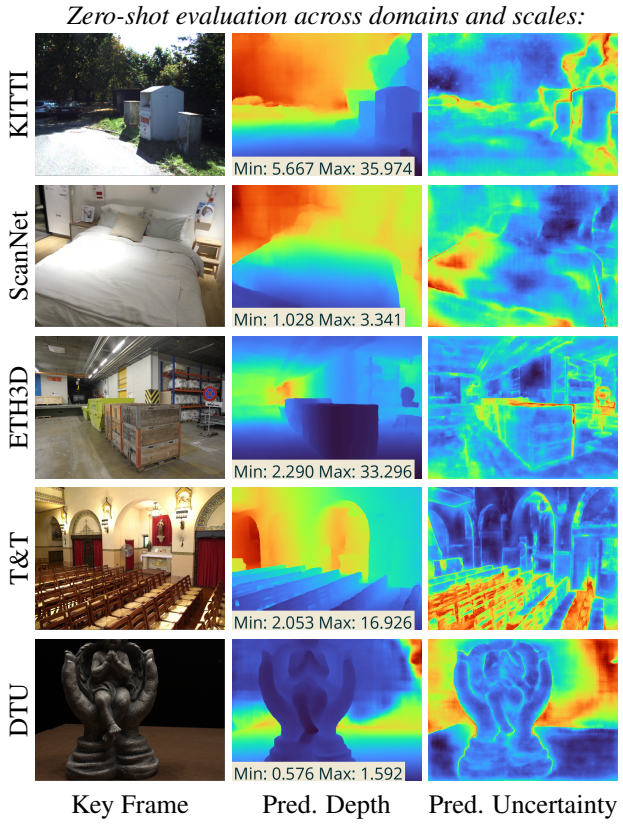


Figure 1. We introduce the Robust Multi-View Depth (MVD) Benchmark to evaluate multi-view depth estimation models regarding robust application on arbitrary real-world data. The benchmark comprises evaluation of depth and uncertainty estimates on multiple existing datasets from different domains and includes a setting with given camera poses and evaluation against absolute depth maps at real-world scale. We provide a baseline model that is built on existing components and generalizes well across domains and can be applied and extended for robust multi-view depth estimation. Predictions shown above are from the baseline model on benchmark data. Purple colors indicate small values, red colors large values, and the ranges are given in meters.

1. Introduction

Since the early days of computer vision, depth is reconstructed using the motion parallax between multiple views [13, 11, 15, 16]. The principle of motion parallax is generic. It works the same in all domains, just like physics is the same everywhere in the world. Consequently, classical geometry-based approaches are not bound to a training distribution, but are agnostic to data from different domains.

In recent years, approaches based on deep learning have emerged for multi-view depth estimation. They are employed either in a depth-from-video setting with images from a video with small and incremental but unknown camera motion [22, 31, 19], or a multi-view stereo setting with unstructured but calibrated image collections [6, 27, 12]. Usually, at the core of these approaches are deep networks that correlate learned features from multiple images and learn to decode the obtained cost volume to an estimated depth map. This design, in principle, allows the network to base its estimates on the motion parallax, which should enable good generalization across domains and consistent

predictions for different scene scales. However, approaches are often evaluated only on data similar to their training domain. Furthermore, evaluation is predominantly done only up to a relative scene scale: in depth-from-video, predictions are aligned to ground truth depths based on the median values; in multi-view stereo, models are supplied with minimum and maximum depth values and predict relative values within this range.

In this work, we introduce a benchmark based on existing datasets to evaluate multi-view depth models regarding generalization across domains. Moreover, as specific cases like small camera motion, occlusions, or texture-less regions are potentially problematic, it is beneficial if a model comes with a measure of its uncertainty, which should be aligned with the depth prediction error. In particular, the Robust Multi-view Depth Benchmark (1) evaluates estimated depth maps on data from different domains in a zero-shot fashion and (2) evaluates uncertainties with the Area Under Sparsification Error Curve metric. Further, it (3) includes evaluation in an absolute setting where ground truth camera poses are given to the model and evaluation is done against ground truth depths at their correct scale. As the scale is provided through the poses, evaluation is done without a given depth range and without alignment. This setting is relevant in practice, *e.g*. in robotics or multi-camera setups where camera poses are known.

We evaluate the depth and uncertainty estimates of recent models in their original relative depth-from-video or multi-view-stereo settings, as well as in the absolute depth estimation setting described above. We find that (1) almost all models have imbalanced performances across domains, (2) uncertainties show only limited alignment with the prediction error, and (3) models mostly perform well on a relative scale, but cannot be directly applied to estimate depths with their correct scale across datasets. We attribute the problems at an absolute scale to out-of-distribution statistics in the correlation cost volume: depth-from-video models learn to use only the cost volume scores corresponding to absolute depth values seen during training; multi-view stereo models overfit to the cost volume distributions within the given minimum and maximum depth values and hence require a sufficiently accurate depth range to be known.

The problems in depth estimation at absolute real-world scale limit practical application. To resolve this, we build a simple baseline model for robust cross-domain, scale-agnostic multi-view depth estimation. The model is mostly based on existing components, such as the DispNet architecture [14] and trained on the BlendedMVS dataset [28] and a static version of the FlyingThings3D dataset [14]. We only add scale augmentation as a new component to randomize across scales during training. This plain baseline achieves what the use of motion parallax promises: it generalizes across domains and scales.

2. Previous methods and benchmarks

Depth-from-video In depth-from-video, depth maps are estimated from consecutive images of a video. Typically, it is assumed that the camera intrinsics are known, but not the camera motion. Hence, the task usually also comprises estimating the camera motion between images. DeMoN [22] was the first deep learning based approach for this task. DeMoN consists of a single network, which estimates depth and camera motion jointly from a pair of consecutive images. Later approaches are DeepTAM [31] and DeepV2D [19], which both process more than two frames, and estimate depth and camera motion with separate mapping and tracking modules, that are applied alternatingly. In such approaches, the mapping and tracking module typically overfit to the scene scale seen during training. Applying the models to scenes at a different scale requires aligning predictions to the scene scale based on additional information. Furthermore, our studies show that the mapping modules of such approaches do not generalize across scale, *i.e*. it is not generally possible to input ground truth camera poses at real-world scale and obtain absolute depth. We argue that this is a shortcoming, as the concept of mapping motion parallaxes to depths given camera motion is independent of the scale.

Multi-view stereo In multi-view stereo, the task is to estimate the 3D geometry of an observed scene from an unstructured set of multiple images with known intrinsics and camera poses. Here, we focus on depth maps as a 3D geometry representation. DeepMVS [6] was the first deep network based approach for this task. DeepMVS brings the keyview in correspondence with source views with a correlation layer that samples patches from source images based on candidate depth values and compares them to patches from the key image. The resulting view-wise matching features are fused by max-pooling. MVSNet [27] takes a similar approach, but compares source views and the keyview in a learned feature space, and fuses multi-view information based on the variance across source views. Many follow up works build upon this concept. R-MVSNet [27] reduces memory consumption by recurrent application. CVP-MVSNet [25] and CAS-MVSNet [5] correlate in a coarse-to-fine fashion to reduce computational constraints and enable higher output resolutions. Vis-MVSNet [30] improves fusion of multi-view information with a late-fusion strategy based on predicted uncertainties. Regarding different scene scales, all these approaches require the minimum and maximum depth value of the observed scene as input and predict depths relative to this range. Our studies show that these approaches have problems in a more general setting where ground truth poses are given, but the depth range of the observed scene is unknown.

Benchmarks and datasets Learned depth-from-video approaches are mostly evaluated on KITTI [4, 21] and

ScanNet [2]. KITTI is a benchmark suite for key tasks in vision-based autonomous driving, including depth estimation. ScanNet is a dataset for 3D scene understanding with annotated RGB-D videos of indoor scenes that were acquired at scale with an elaborate capturing framework. Learned multi-view stereo approaches are mostly evaluated on DTU [9, 1], ETH3D [17], and Tanks and Temples [10]. DTU consists of 80 scenes, each showing a tabletop object that was captured with a camera and a structured light scanner mounted on a robot arm. Tanks and Temples consists of real-world scenes that were captured indoors and outdoors with a high resolution video camera and an industrial laser scanner. Likewise, the ETH3D high-resolution multi-view stereo benchmark consists of images of diverse indoor and outdoor scenes, captured with a high resolution DSLR camera and an industrial laser scanner. Training is often done on the same datasets, namely on KITTI, ScanNet, and DTU. Additionally, some approaches train on BlendedMVS [28], which is designed specifically for large diversity to improve generalization. In this work, we additionally train on the FlyingThings3D dataset [14], which has been shown to enable good generalization in other matching-based tasks like disparity [14] and optical flow estimation [8, 20].

3. Robust Multi-view Depth Benchmark

Key considerations In this work, we aim to evaluate multi-view depth models regarding robust depth estimation on arbitrary real-world data. To reflect this, we propose the Robust Multi-view Depth (MVD) Benchmark based on the following four key considerations:

(1) Depth estimation performance should be independent of the target domain. As a proxy, the benchmark defines test sets from diverse existing datasets. The training set is not defined, but must differ from test datasets. Evaluation is done in a zero-shot cross-dataset setting without fine-tuning. This simulates robustness to arbitrary, potentially unseen real-world data.

(2) The benchmark should be applicable to different multi-view depth estimation settings. To this end, the benchmark allows different input modalities and optional alignment between predicted and ground truth depths.

(3) Estimated uncertainty measures should be aligned with the depth estimation error. This is evaluated with the Area Under Sparsification Error Curve metric [7].

(4) The evaluation should not be affected by the selection of source views. For this, a procedure to find and evaluate with a quasi-optimal set of source views is used.

Relation to existing benchmarks For multi-view stereo, multiple established benchmarks exist, *e.g.* DTU [9, 1], ETH3D [17], and Tanks and Temples [10]. We consider the proposed benchmark as complementary to these benchmarks. Existing multi-view stereo benchmarks evaluate 3D reconstruction performance on the basis of fused point-clouds. Complementarily, the proposed benchmark evaluates the generalization capabilities of learned models based on their typical raw outputs, namely depth maps and uncertainties. We encourage future works to evaluate 3D reconstruction performance on existing benchmarks, but additionally evaluate generalization capabilities on the Robust MVD Benchmark. Depth-from-video models are usually trained and evaluated on existing datasets, *e.g.* KITTI or ScanNet. Usually the test sets are less diverse than in the proposed benchmark. We hence encourage future works on depth-from-video to evaluate generalization capabilities on the Robust MVD Benchmark. Depth estimation at absolute scale is usually not evaluated. However, we consider this setting as relevant in practice and encourage future work to evaluate in this setting on the Robust MVD Benchmark.

In the following, we first describe the setup of the Robust MVD Benchmark in Sec. 3.1. We then present results of recent multi-view depth models, as well as the proposed Robust MVD Baseline model on the benchmark in Sec. 3.2. We provide details on the baseline model in Sec. 4.

3.1. Setup

Test sets The test sets of the Robust MVD Benchmark are defined based on the KITTI [21], ScanNet [2], ETH3D [17], DTU [9, 1] and Tanks and Temples [10] datasets, as they are common for multi-view stereo and depth-from-video evaluation and cover diverse domains and scene scales.

Test set	KITTI [4, 21]	ScanNet [2]	ETH3D [17]	DTU [9, 1]	T&T [10]
domain	driving	indoor	in- & outdoor	tabletop	in- & outdoor
setting	DFV	DFV	MVS	MVS	MVS
cam motion	small	small	large	small	small
scene scale	2 – 85 m	0.2 – 9 m	0.3 – 60 m	0.4 – 1.2 m	1.1 – 42 m
split	test split from	test split from	orig. train	val. split from	orig. train
based on	Eigen *et al.* [3]	Tang and Tan [18]	split	Yao *et al.* [27]	split
full res.	1226x370	640x480	6048x4032	1600x1200	1962x1092
# samples	93	200	104	110	69

Table 1. **Test sets of the Robust MVD Benchmark** are based on KITTI, ScanNet, ETH3D, DTU, and Tanks and Temples (T&T). These datasets are commen for depth-from-video (DFV) or multi-view stereo (MVS) and cover different domains and scene scales.

Each test set is a set of samples from the respective dataset. Each sample has input views $V = (V_0, \cdots, V_k)$, consisting of a keyview V_0 and source views $V_{1,..,k}$, and (potentially sparse) ground truth depth values $\mathbf{z}^*$ for the keyview. Each view $V_i = (\mathbf{I}_i, {}^i_0\mathbf{T}, \mathbf{K}_i)$ consists of an image $\mathbf{I}_i$, a pose ${}^i_0\mathbf{T}$ relative to the keyview, and intrinsics $\mathbf{K}_i$. The task is to estimate a dense depth map $\mathbf{Z}$ for the keyview V_0 from the input data. The test sets are chosen such that they are as comparable to existing data splits as possible. The test sets are deliberately rather small to speed up evaluation, but samples have been selected to cover a large diversity. An overview of the test sets is given in Tab. 1 and further details are provided in the Appendix.

Training set The benchmark does not define a training set as the objective is robustness to arbitrary real-world data,

independent of a specific training setup. However, it must be specified in case training data is used that overlaps with test datasets of the benchmark.

Evaluation settings The benchmark allows evaluation with different input modalities that are provided to the model and with an optional alignment between predicted and ground truth depths. The provided input modalities always include the images $\mathbf{I}_i$ and intrinsics $\mathbf{K}_i$ for each view and can optionally include the poses ${}^{i}_{0}\mathbf{T}$ and the ground truth depth range (z^*_{min}, z^*_{max}) with minimum and maximum ground truth depth values. To account for the scale-ambiguity of some models, predicted depth maps can optionally be aligned to the ground truth depth maps before computing the metrics, *e.g.* based on the ratio of the median ground truth depth and the median predicted depth.

In literature, depth-from-video models are typically applied without poses and ground truth depth range and evaluated with alignment. Multi-view stereo models are typically applied with poses and ground truth depth range and evaluated without alignment. Both settings evaluate depth estimations on a relative scale, *i.e.* up to an unknown scale factor or within a given depth range. In contrast, the benchmark additionally evaluates depth estimation on an absolute scale. Here, the models are provided with poses but without depth range and the task is to estimate depth maps at absolute real-world scale. Evaluation is done without alignment.

Depth estimation metrics Results are reported per test set for the Absolute Relative Error (rel) and the Inlier Ratio (τ) with a threshold of 1.03 [3, 21]:

$$\text{rel} = 100 \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} \frac{\left| z_{i,j} - z^*_{i,j} \right|}{z^*_{i,j}} \tag{1}$$

$$\tau = 100 \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} [\max\left(\frac{z_{i,j}}{z^*_{i,j}}, \frac{z^*_{i,j}}{z_{i,j}} \right) < 1.03] \tag{2}$$

where j indexes the m pixels with valid ground truth depth, i indexes the n samples in a test set, and $[\cdot]$ denotes the Iverson bracket. The Absolute Relative Error indicates the average relative deviation of predicted depth values from ground truth depth values in percent. The Inlier Ratio indicates the percentage of pixels with correct predictions, where a prediction is considered correct if it has an error below $3\,\%$. In addition to results on individual test sets, average metrics and model runtimes are reported over all test sets.

Estimated depth maps are upsampled to the full resolution before computing the metrics. Additionally, to remove the effect of implausible outliers, depth estimates are clipped to a range of $0.1\,\text{m}$ to $100\,\text{m}$. We conjecture that this is a reasonable range for real-world application.

Uncertainty estimation metrics Results are reported with commonly used Sparsification Error Curves and the Area Under Sparsification Error Curve (AUSE) metric [7]. For the Sparsification Error Curves, the most erroneous pixels are gradually excluded from the error metric based on actual pixel errors (oracle uncertainty) versus estimated pixel uncertainties. The Sparsification Error Curve then is the difference of the oracle-based and uncertainty-based error reduction. The AUSE is the area under the Sparsification Error Curve. An AUSE of 0 is optimal and indicates perfect alignment between estimated uncertainties and actual errors. More details are provided in the Appendix.

Source view selection To factor out effects from the selection of source views on the model performance, the benchmark finds and evaluates with a quasi-optimal set of source views for each model. For a given sample, the model is run for all pairs (V_0, V_i) of the keyview and a single source view and the resulting Absolute Relative Errors are stored. The set of source views is then grown incrementally by adding the source views in the order of the stored Absolute Relative Errors. Results are reported for the set of source views with the overall lowest Absolute Relative Error. Additionally, the Absolute Relative Error is plotted over the size of the source view set.

3.2. Robust MVD Benchmark results

Evaluated models In this work, we evaluate the COLMAP [16, 15], DeMoN [22], DeepTAM [31], DeepV2D [19], MVSNet [27], CVP-MVSNet [25], Vis-MVSNet [30], PatchmatchNet [23], Fast-MVSNet [29], and MVS2D [26] models on the proposed benchmark. This choice reflects seminal works that lay ground for later improvements, as well as works that represent the current state of the art. For all models, we use the original provided code and weights, except for MVSNet where we use the PyTorch implementation from Xiaoyang Guo, as it gave better performance than the original Tensorflow version. We additionally evaluate a MVSNet that we re-trained with plane sweep sampling in inverse depth space. For DeepV2D, we evaluate the KITTI and ScanNet models. For MVS2D, we evaluate the ScanNet and DTU models. Note that we intentionally not re-train models on a specific uniform dataset, as the objective of the benchmark is generalization across diverse test sets, independent of the training data.

Results In Tab. 2, we report results of evaluated models on the proposed Robust MVD Benchmark. We report results up to a relative scale in the typical depth-from-video and multi-view stereo settings, as well as on an absolute scale. In the following, we discuss the results.

Classical approaches For a comparison to classical approaches, in Tab. 2a we report results of COLMAP [16, 15] on the benchmark. The results of applying COLMAP with default parameters cannot be directly compared to those of learned models, as COLMAP estimates depth maps at a lower density ($54\,\%$ in average) and we compute the metrics only for pixels with a valid prediction. We additionally

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 [16, 15]	✓	✗	✗	**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 [16, 15]	✓	✗	✗	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 [22]	✗	✗	$\|\mathbf{t}\|$	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 [19]	✗	✗	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 [19]	✗	✗	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 [27]	✓	✓	✗	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 [27]	✓	✓	✗	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 [25]	✓	✓	✗	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 [30]	✓	✓	✗	**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 [23]	✓	✓	✗	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 [29]	✓	✓	✗	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 [26]	✓	✓	✗	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 [26]	✓	✓	✗	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 [22]	✓	✗	✗	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 [31]	✓	✗	✗	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 [19]	✓	✗	✗	(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 [19]	✓	✗	✗	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 [27]	✓	✗	✗	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 [27]	✓	✗	✗	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 [25]	✓	✗	✗	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 [30]	✓	✗	✗	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 [23]	✓	✗	✗	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 [29]	✓	✗	✗	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 [26]	✓	✗	✗	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 [26]	✓	✗	✗	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

Table 2. **Quantitative results for the evaluated multi-view depth models on the Robust MVD Benchmark with different evaluation settings: a)** Classical approaches. **b)** Evaluation without poses, without depth range, with alignment. This is the common setting in depth-from-video literature. **c)** Evaluation with poses, with depth range, without alignment. This is the common setting in multi-view stereo literature. **d)** Absolute scale evaluation with poses, without depth range, without alignment. Results are reported for the Absolute Relative Error (rel) and Inlier Ratio (τ) with a threshold of 1.03 on each test set and as averages across all test sets. Additionally, the average runtime in seconds of each model across all test sets is reported. All results are for the quasi-optimal selection of source views of each model. (Parentheses) denote training on data from the same domain. **Bold** denotes best results.

report results for COLMAP without filtering, which results in dense predictions but lower accuracy.

Evaluation up to a relative scale In Tab. 2b and 2c, we report results in the typical depth-from-video and multi-view stereo settings up to a relative scale. It shows that all models perform significantly better on the training domain.

Evaluation on an absolute scale In Tab. 2d, we provide results of the evaluated models in an absolute scale depth estimation setting. For DeepV2D and DeepTAM, we only use the mapping module and input ground truth poses. For models that require a given depth range, we assume an unknown depth range and provide a default range of $0.2\,\mathrm{m}$ to $100\,\mathrm{m}$. This covers the range of all test sets and simulates real-world applications with no information except poses.

In this setting, all evaluated models perform significantly worse. Depth-from-video models perform worse on datasets with a different depth range than the training data (*e.g.* DeepV2D-ScanNet on KITTI). Multi-view stereo models perform worse on datasets where the depth differs from the given default depth range (*e.g.* MVSNet on DTU). Most evaluated models internally build and decode a cost volume, which is computed in a plane sweep stereo fashion by correlating source views with the keyview for specific (inverse) depth values. We attribute the performance decrease to out-of-distribution cost volume statistics. Depth-from-video models learn to use only the cost volume scores corresponding to absolute depth values seen during training. Multi-view stereo models overfit to the cost volume distribution within the provided depth range.

In practice, this means that existing depth-from-video models cannot be generally used with known ground truth camera poses. Multi-view stereo models in turn require a sufficiently accurate depth range of the observed scene to be known. Even though this depth range can be obtained by running structure-from-motion, this comes at the cost of increased runtime and complexity.

The proposed Robust MVD Baseline model shows consistent performance across all test sets. We conjecture that

the model really learned to exploit multi-view cues that generalize across domains. Furthermore, the proposed scale augmentation enables absolute scale depth estimation independent of the scene scale.

Performances depending on source views In Fig 2, we plot performances for different numbers of source views. Additionally, model runtimes for different numbers of

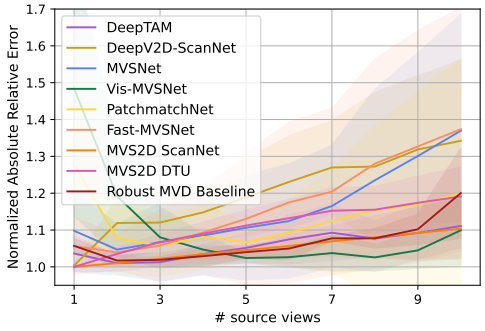


Figure 2. **Effect of the number of source views on the performance of evaluated models.** Each plot shows the average Absolute Relative Error across all test sets relative to the quasi-optimal performance of each model (Tab. 2). The shaded area indicates the standard deviation across test sets.

source views are provided in the Appendix. For all models, we plot results in the respective setting that gives best average results according to Tab. 2. In an ideal curve, the error would decrease with additional source views and converge to a minimal value when more views do not contain additional information. The evaluation shows that multi-view fusion strategies of most models are suboptimal.

Uncertainty evaluation In Fig. 3, we plot sparsification error curves for evaluated models that predict a measure of their depth prediction uncertainty. In Tab. 3, we report the

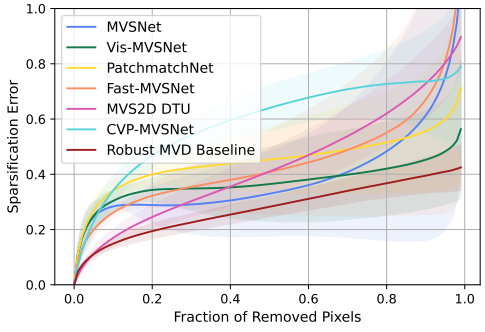


Figure 3. **Evaluation of estimated uncertainty measures.** Lower curves indicate better alignment between estimated uncertainties and actual errors. The area under the curves is the Area Under Sparsification Error Curve, which is reported in Tab. 3.

corresponding Area Under Sparsification Error Curve metric. Again, we report results for each model in the respective setting that gives best average performance. The results of previous models show a suboptimal alignment between estimated uncertainties and errors, whereas the Robust MVD Baseline model gives better uncertainties.

Approach	KITTI	ScanNet	ETH3D	DTU	T&T	Average rel ↓	Average AUSE ↓
MVSNet [27]	**0.18**	0.69	0.35	0.39	0.32	18.6	0.39
Vis-MVSNet [30]	0.28	0.53	0.37	**0.27**	0.39	7.0	0.37
PatchmatchNet [23]	0.47	0.55	0.52	0.28	0.40	9.1	0.45
Fast-MVSNet [29]	0.28	0.73	0.42	0.29	0.48	13.5	0.44
MVS2D DTU [26]	0.41	0.50	0.43	0.31	0.47	77.5	0.43
CVP-MVSNet [25]	0.56	0.68	0.57	0.56	0.55	95.9	0.58
Robust MVD Baseline	0.24	**0.33**	**0.28**	0.28	**0.24**	**6.3**	**0.27**

Table 3. **Evaluation of estimated uncertainties** with the Area Under Sparsification Error Curve (AUSE). An AUSE of 0 means optimal alignment of uncertainties and errors.

4. Robust MVD Baseline

In the following, we describe the Robust MVD Baseline, which is designed specifically as a baseline for robust depth estimation across domains and scene scales and can serve as baseline for evaluation on the proposed benchmark. The model is mostly based on existing components and we provide ablation studies for individual components in Tab. 4.

4.1. Model architecture

The Robust MVD Baseline model builds on the simple DispNet [14] network architecture, but is adapted to the given multi-view setting with non-rectified images. More specifically, as illustrated in Fig. 4, and using the notation defined in Sec. 3.1, the model architecture is structured as follows: (1) a siamese encoder network f_θ that maps input images $\mathbf{I}_i$ to feature maps, $\mathbf{F}_i = f_\theta(\mathbf{I}_i)$, (2) a correlation layer that correlates keyview features $\mathbf{f}_0$ with source view features $\mathbf{f}_i$ in a plane sweep fashion, resulting in view–wise cost volumes $\mathbf{C}_{1,..,k}$, (3) a context encoder network h_σ that maps the key image to features $\hat{\mathbf{F}}_0 = h_\sigma(I_0)$ that are used to decode cost volumes, (4) a fusion module g_ρ that fuses the cost volumes from multiple source views to a fused representation $\mathbf{C} = g_\rho(\mathbf{C}_{1,..,k}, \hat{\mathbf{F}}_0)$ via weighted averaging with learned weights, and (5) a 2D convolutional cost volume decoder network $(\mathbf{D}, \mathbf{U}) = k_\phi(\mathbf{C}, \hat{\mathbf{F}}_0)$ that maps the fused cost volume to an output inverse depth map $\mathbf{D}$, and an uncertainty map $\mathbf{U}$. The inverse depth map $\mathbf{D}$ holds predicted inverted depth values $d = 1/z$ for every keyview pixel. The plane sweep correlation has been shown to work well in other multi-view depth architectures [19, 31, 27, 12, 6] and is explained in the Appendix.

In the first experiments, we apply the base model in dual-view mode, using only a single source view. This factors

Approach	KITTI		ScanNet		ETH3D		DTU		T&T		Average		
	rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑	time [ms] ↓
a) Scale augmentation													
No scale augmentation	18.5	19.6	81.9	8.9	42.0	15.8	804.2	0.0	17.9	43.2	192.9	17.5	32.1
With scale augmentation	**15.2**	**21.7**	**8.5**	**31.5**	**19.4**	**22.7**	**5.7**	**49.5**	**14.5**	**50.5**	**12.7**	**35.2**	32.7
b) Training data													
ST3D	15.2	21.7	**8.5**	31.5	19.4	22.7	5.7	49.5	14.5	50.5	12.7	35.2	32.7
BMVS	11.1	27.3	9.3	29.5	**12.4**	**31.6**	**4.6**	**62.9**	**9.4**	52.9	**9.4**	40.8	34.7
ST3D+BMVS	**10.2**	**27.7**	8.7	**31.7**	14.6	30.7	4.7	61.4	11.3	**57.3**	9.9	**41.8**	33.6
c) Model architecture													
MVSNet architecture	11.3	**29.7**	15.2	23.4	36.8	25.9	123.8	48.8	**10.8**	**60.4**	39.6	37.6	193.2
DispNet architecture	**10.2**	27.7	**8.7**	**31.7**	**14.6**	**30.7**	**4.7**	**61.4**	11.3	57.3	**9.9**	**41.8**	33.6
d) Uncertainty estimation													
Deterministic	10.2	27.7	8.7	31.7	14.6	30.7	4.7	61.4	11.3	57.3	9.9	41.8	33.6
Laplace distribution	**9.3**	**31.9**	**8.2**	**35.0**	**11.7**	**38.1**	**3.4**	**76.6**	**9.1**	**63.7**	**8.4**	**49.1**	35.6
e) Multi-view fusion													
1 source view	9.3	31.9	8.2	35.0	11.7	38.1	3.4	76.6	9.1	63.7	8.4	49.1	35.6
Averaging	6.7	40.1	7.5	38.5	9.7	39.9	3.0	79.6	6.0	74.2	6.6	54.5	58.6
Learned view weights	**6.6**	**42.0**	**7.4**	**38.7**	9.2	**42.9**	2.9	80.6	7.6	**76.0**	6.8	**56.0**	61.7
Learned view weights + Eraser	7.1	41.9	**7.4**	38.4	**9.0**	42.6	**2.7**	**82.0**	**5.0**	75.1	**6.3**	**56.0**	59.2

Table 4. **Ablation studies for the Robust MVD Baseline model**. All results are for the absolute scale depth estimation setting (Tab. 2d). **a)** Scale augmentation is essential for generalization across scene scales. **b)** Joint training on StaticThings3D and BlendedMVS gives best performance. **c)** A DispNet architecture performs better than a MVSNet architecture. **d)** Predicting parameters of a Laplace distribution instead of point estimates improves performance. **e)** Multi-view fusion via weighted averaging with learned weights work slightly better than simple averaging. The last model is the Robust MVD Baseline model.

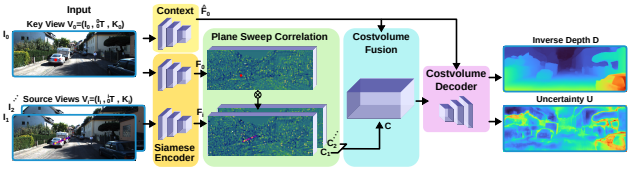


Figure 4. **Robust MVD Baseline model architecture**. The model consists of a siamese encoder network, a plane sweep correlation that correlates source views with the keyview, a fusion module, and a decoder that predicts inverse depths and uncertainties.

out effects from the multi-view cost volume fusion and allows for an isolated evaluation of effects from data augmentation, training dataset, model architecture, and uncertainty estimation. Following this, we evaluate different strategies for fusing multi-view information. In Tab. 4c, we compare the DispNet architecture with a MVSNet architecture.

4.2. Data augmentation

Standard photometric and spatial augmentations are applied uniformly to all views. Additionally, to prevent the model from overfitting to the depth distribution of the training data, we introduce a novel data augmentation strategy that we term scale augmentation. Scale augmentation re-scales ground truth translations ${}^{i}_{0}\mathbf{t}$ during training before feeding them to the model. Likewise, the ground truth inverse depth map $\mathbf{D}^{*}$ is scaled with the inverse scaling factor. Inverse depth values outside the range $[0.009\,\mathrm{m}^{-1}, 2.75\,\mathrm{m}^{-1}]$ are masked. To choose the scaling factor, a histogram of the depth values that were seen during previous training iterations, is maintained. The size of the histogram bins increases logarithmically, as consistent performance across the full depth range requires training for smaller depth values at a finer resolution. This is illustrated by Fig. 5. Scaling factors are then computed as the ratio of

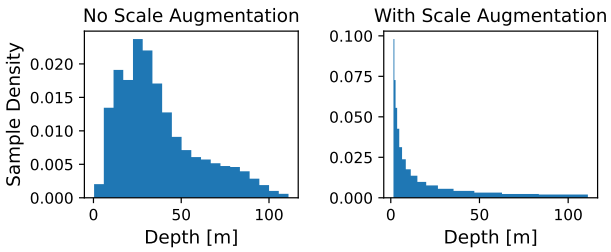
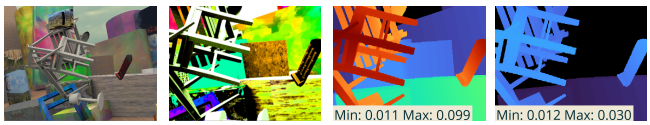


Figure 5. Effect of scale augmentation on the distribution of depth values seen during training on StaticThings3D. With scale augmentation, smaller depth values are sampled at a higher density. All bins in the "With Scale Augmentation" histogram cover the same area. This gives consistent performance across scene scales.

the depth label of the histogram bin with the lowest count and the median ground truth depth value of the current sample. Fig. 6 shows effects of the data augmentation on an exemplary sample. As shown by the results in Tab. 4a, scale augmentation is a key component for enabling the model to generalize across different scene scales.

4.3. Training data

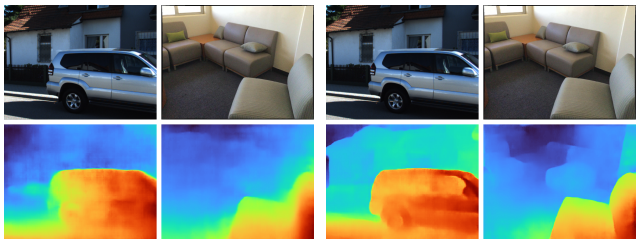
The Robust MVD Baseline model is jointly trained on a static version of the existing FlyingThings3D dataset [14],



(a) (b) (c) (d)

Figure 6. **Training data and augmentation**: (a) keyview image $\mathbf{I}_0$ of a StaticThings3D training sample, (b) augmented keyview image, (c) ground truth inverse depth $\mathbf{D}^*$, and (d) $\mathbf{D}^*$ after scale augmentation with a randomly sampled scaling factor of 3.27. Translations ${}^{i}_{0}\mathbf{t}$ to source views are scaled with the same factor.

that we term StaticThings3D (see Fig. 6), and on the existing BlendedMVS [28] dataset. StaticThings3D is similar to FlyingThings3D: it contains 2250 train and 600 test sequences with 10 frames per sequence, showing randomly placed ShapeNet objects in front of random Flickr backgrounds. However, in StaticThings3D, all objects are static, and only the camera moves. The advantage of using this randomized synthetic dataset is that it reduces the possibility of a model to overfit to domain-specific priors. In Tab. 4b, we compare joint training on StaticThings3D and BlendedMVS against training on a single dataset. Joint training performs quantitatively on par with training solely on BlendedMVS, but results in more accurate object boundaries, as shown in Fig. 7. Further training details are provided in the Appendix.

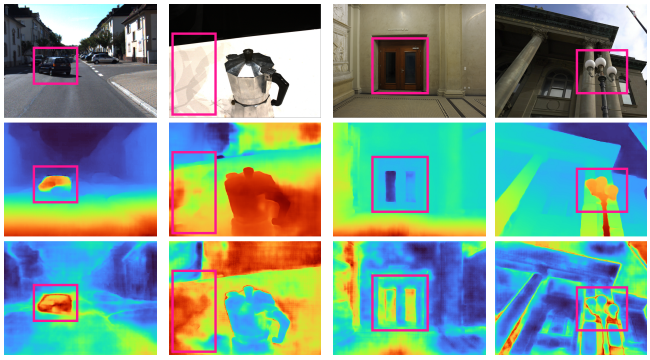


(a) (b)

Figure 7. **Effect of the training dataset**: the first row shows keyview images (KITTI and ScanNet), and the second predicted inverse depth maps. (a) Model trained on BlendedMVS. (b) Model trained jointly on BlendedMVS+StaticThings3D.

4.4. Uncertainty estimation

Instead of predicting a point estimate of the inverse depth map, the Robust MVD Baseline model predicts parameters of a Laplace distribution, as in [7] and [30]. For this, an additional output channel is added to the network such that one channel encodes the predicted location parameter and the other the predicted scale parameter. Training is then done by minimizing the negative log likelihood. Effects on the depth prediction performance are evaluated in Tab. 4d. Predicted uncertainties are evaluated in Tab. 3 and shown qualitatively in Fig. 8.



(a) (b) (c) (d)

Figure 8. **Uncertainty estimation**: the first row shows keyview images, the second predicted inverse depth maps, and the third predicted uncertainty maps (red is uncertain). The model outputs high uncertainties for problematic cases *e.g.* (a) moving objects, (b) textureless regions, (c) windows, or (d) fine structures.

4.5. Multi-view fusion

We evaluate two strategies for multi-view fusion, namely averaging of cost volumes from multiple source views, and weighted averaging with learned weights, *e.g.* as in [24]. For the weighted averaging, a small 2D convolutional network with two layers is applied with shared weights to all view-wise cost volumes and outputs pixel-wise weights for each view. We conduct multi-view training with an eraser data augmentation, where regions in source views are randomly replaced with the mean color. Results for both multi-view fusion strategies are given in Tab. 4e. The model with learned weights is the Robust MVD Baseline model.

5. Conclusion

We presented the Robust MVD Benchmark to evaluate the robustness of multi-view depth estimation models on different data domains. The benchmark supports different evaluation settings, *i.e.* different input modalities and optional alignment between predicted and ground truth depths. We found that existing methods have imbalanced performance across domains and cannot be directly applied to arbitrary real-world scenes for estimating depths with their correct scale from given camera poses. We also demonstrated that this can be resolved mostly with existing technology. Together with the benchmark, we provide a robust baseline method that can serve as a basis for future work.

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.

Further, we thank Özgün Cicek and Christian Zimmermann for their comments on the text and Stefan Teister for keeping our systems running.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 2016. 3

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, July 2017. 3

[3] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, January 2014. 3, 4

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 2013. 2, 3

[5] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, 2020. 2

[6] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. DeepMVS: Learning multi-view stereopsis. In *CVPR*, June 2018. 1, 2, 6

[7] Eddy Ilg, Özgün Çiçek, Silvio Galesso, Aaron Klein, Osama Makansi, F. Hutter, and T. Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *ECCV*, 2018. 3, 4, 8

[8] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 3

[9] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, 2014. 3

[10] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and Temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 3

[11] Vladimir Kolmogorov and Ramin Zabih. Multi-camera scene reconstruction via graph cuts. In *ECCV*, 2002. 1

[12] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa Narasimhan, and Jan Kautz. Neural RGB-D sensing: Depth and uncertainty from a video camera. In *CVPR*, 2019. 1, 6

[13] Hugh Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828):133–135, 1981. 1

[14] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, June 2016. 2, 3, 6, 7

[15] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, 2016. 1, 4, 5

[16] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *ECCV*, 2016. 1, 4, 5

[17] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, July 2017. 3

[18] Chengzhou Tang and Ping Tan. BA-Net: Dense bundle adjustment networks. In *ICLR*, May 2019. 3

[19] Zachary Teed and Jia Deng. DeepV2D: Video to depth with differentiable structure from motion. *ICLR*, 2020. 1, 2, 4, 5, 6

[20] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*. Springer, 2020. 3

[21] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant CNNs. In *3DV*, October 2017. 2, 3, 4

[22] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. DeMoN: Depth and motion network for learning monocular stereo. In *CVPR*, 2017. 1, 2, 4, 5

[23] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. PatchmatchNet: Learned multi-view patchmatch stereo, 2021. 4, 5, 6

[24] Qingshan Xu and Wenbing Tao. PVSNet: Pixelwise visibility-aware multi-view stereo network. *arXiv preprint arXiv:2007.07714*, 2020. 8

[25] Jiayu Yang, Wei Mao, Jose M. Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *CVPR*, June 2020. 2, 4, 5, 6

[26] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. MVS2D: Efficient multi-view stereo via attention-driven 2d convolutions. In *CVPR*, 2022. 4, 5, 6

[27] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. *ECCV*, September 2018. 1, 2, 3, 4, 5, 6

[28] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. *CVPR*, 2020. 2, 3, 8

[29] Zehao Yu and Shenghua Gao. Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *CVPR*, 2020. 4, 5, 6

[30] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *BMVC*, 2020. 2, 4, 5, 6, 8

[31] Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. DeepTAM: Deep tracking and mapping. In *ECCV*, September 2018. 1, 2, 4, 5, 6