

Depth-Guided Sparse Structure-from-Motion for Movies and TV Shows

Sheng Liu*
 University at Buffalo
 sliu66@buffalo.edu

Xiaohan Nie, Raffay Hamid
 Amazon Prime Video
 {nxiaohan, raffay}@amazon.com

Abstract

Existing approaches for Structure from Motion (SfM) produce impressive 3-D reconstruction results especially when using imagery captured with large parallax. However, to create engaging video-content in movies and TV shows, the amount by which a camera can be moved while filming a particular shot is often limited. The resulting small-motion parallax between video frames makes standard geometry-based SfM approaches not as effective for movies and TV shows. To address this challenge, we propose a simple yet effective approach that uses single-frame depth-prior obtained from a pretrained network to significantly improve geometry-based SfM for our small-parallax setting. To this end, we first use the depth-estimates of the detected keypoints to reconstruct the point cloud and camera-pose for initial two-view reconstruction. We then perform depth-regularized optimization to register new images and triangulate the new points during incremental reconstruction. To comprehensively evaluate our approach, we introduce a new dataset (**StudioSfM**) consisting of 130 shots with 21K frames from 15 studio-produced videos that are manually annotated by a professional CG studio. We demonstrate that our approach: (a) significantly improves the quality of 3-D reconstruction for our small-parallax setting, (b) does not cause any degradation for data with large-parallax, and (c) maintains the generalizability and scalability of geometry-based sparse SfM. Our dataset can be obtained at <https://github.com/amazon-research/small-baseline-camera-tracking>.

1. Introduction

Estimating camera motion and 3-D scene geometry in movies and TV shows is a standard task in video production. Existing Structure from Motion (SfM) approaches for 3-D scene reconstruction produce high-quality results especially for images with large parallax [16, 6, 36, 34]. However, creating engaging viewing experience in movies and TV shows often constrains the amount of camera movement while filming a shot. This often leads to insufficient paral-

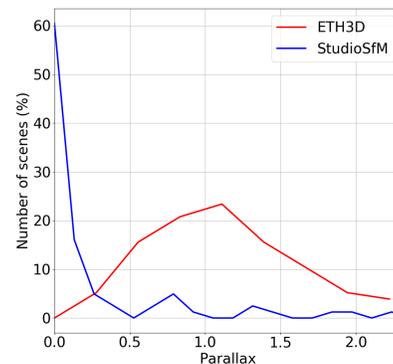


Figure 1: **Comparison of Parallax-Distribution:** Parallax-distribution of ETH3D [34] is plotted with StudioSfM – a new dataset with 21K frames in 130 manually annotated shots from 15 TV episodes (see § 4.1 for details of data and computation of parallax). The long-tail distribution of StudioSfM shows that small-motion parallax is significantly more common in studio-produced content than in standard SfM datasets.

lax compared to standard SfM datasets captured specifically for 3-D reconstruction (see Figure 1 for more details).

This insufficient parallax is one of the key challenges [9] that limits the effectiveness of well-developed geometry-based SfM approaches [10, 2, 33, 1, 43, 28] that recover camera motion and geometry based on the principle of motion-parallax. Shots with small motion-parallax are ill-conditioned for 3-D reconstruction as algebraic methods for two-view reconstruction are numerically unstable in such situations [26]. Conventional SfM pipelines (*e.g.*, COLMAP [33]) use various heuristics to handle small-parallax data, *e.g.*, by using inlier ratio to decide the two-view motion type to prevent two-view reconstruction from using panoramic image pairs, and filtering out points with small triangulation angles. These heuristics however require careful tuning and can fail completely when using data which has no image pairs with sufficient parallax.

In contrast, learning-based approaches [17, 47, 40, 42] are able to handle data with small parallax more effectively as they can learn to predict depth and pose from large-scale labelled datasets. However, as these methods do not incorporate geometric-consistency constraints between images,

*This work was done when the author was an intern at Amazon.

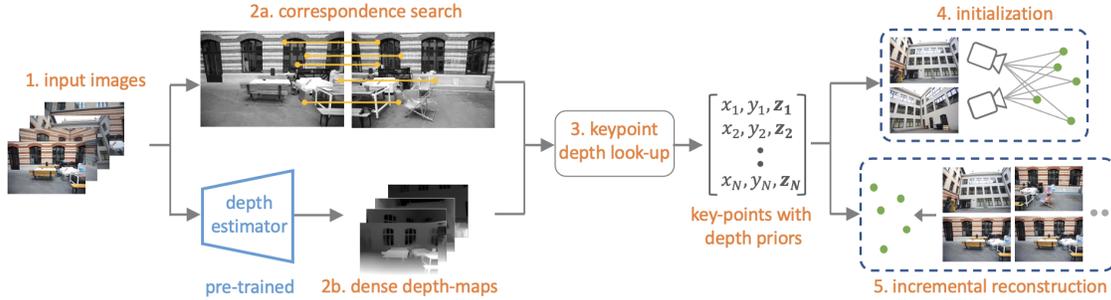


Figure 2: **Proposed Pipeline** – Given a set of input images (Step 1), we detect 2-D keypoints and match them across frames, *i.e.*, correspondence search (Step 2-a), as well as use a pretrained network to obtain their single-frame based dense depth-maps (Step 2-b). We apply bi-linear interpolation to look-up the depths of the detected key-points from these dense depth-maps (Step 3). We use the detected 2-D keypoints along with their depth-priors to improve the initialization (Step 4) and incremental reconstruction steps (Step 5).

their pose and depth estimates are not as accurate [19]. Furthermore, the generalizability of these approaches heavily depends on the scale of labeled data used for their training, which can be laborious and expensive to collect.

Recently proposed hybrid approaches [37, 38, 19] have achieved more accurate results than learning-based approaches by employing learned depth priors as implicit constraints for geometric consistency. However, these approaches do not use robust estimators thus making them heavily dependent on the quality of used optical flow which can adversely affect their robustness. Moreover, these approaches require heavy compute and memory resources. This prevents them from scaling to larger problems.

Key Contributions: To address these challenges, we propose a novel hybrid approach that combines the strengths of: (a) geometry-based SfM to achieve high-accuracy without requiring additional labelled data, and (b) learning-based SfM to effectively handle data with insufficient parallax. As illustrated in Figure 2, our approach builds on the standard geometry-based SfM pipeline and particularly improves its initialization and incremental reconstruction steps by leveraging single-frame depth-priors obtained from a pretrained deep network. Specifically:

- Instead of using epipolar geometry for initial two-view reconstruction, we directly utilize monocular depth obtained from a pretrained model to accurately recover the initial camera pose and point cloud.
- During the incremental reconstruction step, we propose a depth-prior regularized objective function to be able to accurately register and triangulate new images and points.

We demonstrate that our approach is robust to a variety of pretrained networks used to obtain the depth-prior, and maintains the generalizability and scalability of geometry-based SfM pipeline by maximally relying on its well-engineered implementations (*e.g.* COLMAP [33]).

To comprehensively evaluate our approach, we collect a new dataset (**StudioSfM**) containing 130 shots with 21K frames from 15 TV-episodes. The ground truth camera pose

and point clouds were created manually by professionals using commercial CG software (see § 4.1 for details). We use StudioSfM to demonstrate that our approach offers significantly more accurate camera poses and scene geometry over existing state-of-the-art approaches under small-parallax setting in studio-produced content, while does not cause any degradation on standard SfM datasets [34] with large parallax, and maintains the generalizability and scalability of standard SfM pipelines.

2. Related Work

a. Geometry-Based SfM: Geometry-based SfM [2, 33, 1, 43, 28] approaches have undergone tremendous improvements over the past few decades in terms of their robustness, accuracy, completeness and scalability. Most of these approaches first detect and match local image features [24, 13, 32, 7], followed by estimating the two-view motion using epipolar geometry [14] and then reconstructing the 3-D scene either globally or incrementally using bundle adjustment [39]. One of the most widely used open-source geometry-based SfM pipeline is COLMAP [33] which is often used as a preliminary step for state-of-the-art dense reconstruction approaches [25, 23, 27]. Like most geometry-based approaches however, it requires images with sufficiently large baselines. Our approach improves COLMAP [33] to make it work robustly for small parallax setting often found in movies and TV shows.

Previous geometry-based SfM approaches geared for videos with small motion [46] simplify the rotation matrix and parameterize bundle adjustment using inverse depth of reference image. Work in [12] makes the same simplification and parameterization as [46] but assumes that camera intrinsics are unknown and optimizes them in bundle adjustment. These works show improved results only for videos with very small accidental motion and do not generalize to data with relatively larger motion as is the case in movies and TV shows. Unlike [46, 12] that use priors for camera motion, our approach uses priors for scene geometry which

is robust to both narrow as well as wide baseline data.

b. Learning-Based SfM: To jointly estimate motion and depth in an end-to-end fashion, work in [40] stack multiple encoder-decoder networks for their iterative estimation. To improve the robustness of pose estimation, work in [42] construct pose-cost volume similar to depth-cost volume [45] used in stereo matching to predict camera pose iteratively. Unlike [40, 42] which rely on ground truth labels for training, our approach utilizes off-the-shelf pre-trained depth estimators without the need of labels from target data.

c. Hybrid SfM: Hybrid approaches attempt to optimize camera pose and depth by using geometry-consistency constraints. Work in [37] represents depth as a linear combination of depth basis maps, and computes the camera motion and depth by aligning deep features using differentiable gradient descent. Work in [38] uses dense optical flow to build dense correspondences, and iterates between learning based depth estimation and optimization based motion estimation. Work in [19] optimizes the re-projection loss by allowing depth to deform as splines for low-frequency alignment. Depth filters are used for high-frequency alignment to recover the details. In our approach, we do not rely on optical flow which enables our approach to work on both videos as well as un-ordered image-sets.

d. Monocular Depth Estimation: Recent improvements in deep networks and the availability of large-scale depth-data have contributed to the remarkable progress in monocular depth-estimation [18]. Work in [11] learns a depth estimation network in a self-supervised manner using monocular videos. Work in [31] focuses on mixing multiple datasets for training using multiple objectives which are invariant to depth scale and range. Work in [4] divides depth into bins whose centers are estimated adaptively per-image, and are linearly combined to predict the final depth value. We use off-the-shelf pre-trained monocular depth estimators to generate depth-priors for sparse keypoints. Although monocular depth estimates are inconsistent across frames, we show that using them as priors in SfM pipeline helps the reconstruction process to converge to a better solution.

3. Approach

3.1. Review of Incremental SfM

As our approach builds on incremental SfM, for completeness we first review the standard incremental SfM pipeline [33] which can be roughly divided into three key components: (i) correspondence search, (ii) initialization, and (iii) incremental reconstruction. We provide details of these component in the following.

a. Correspondence Search: For each image \mathbf{I} in a given set of N images \mathcal{I} , their 2-D keypoints $\mathbf{p} \in \mathbb{R}^2$ and respective appearance-based descriptors are extracted and used to

match all image-pairs $(\mathbf{I}_a, \mathbf{I}_b) \in \mathcal{I}$ using a similarity-metric based on their keypoint-descriptors. A robust estimator such as RANSAC [8] is used to perform robust geometric verification of the matched image-pairs in order to estimate the geometric transformation between them.

b. Initialization: Based on epipolar geometry of the corresponding 2-D keypoints in a matched image-pair $(\mathbf{I}_a, \mathbf{I}_b)$, two-view reconstruction is performed to estimate the initial camera pose $(\mathbf{R}_{\text{init}}, \mathbf{t}_{\text{init}}) \in \text{SE}(3)$ and 3-D point cloud $\mathbf{P} \in \mathbb{R}^3$. Recall that good initialization is critical in incremental SfM pipelines as later steps may not be able to recover from a poor initialization. To this end, heuristics such as number of keypoint matches, triangulation angles and geometric-transformation types are used to select a good image-pair likely to result in high-quality initialization [33].

c. Incremental Reconstruction: New images from the remaining image-set are incrementally incorporated into the reconstruction process by iterating between the following three steps. i – Image Registration : this step registers a new image to the current 3-D scene by first solving the Perspective-n-Point (PnP) problem [3] using RANSAC [5] on 2-D to 3-D correspondences, and then refining the pose of the new image by minimizing its re-projection error. ii – Triangulation : scene points of the new image are triangulated and added to the existing scene. iii – Bundle Adjustment(BA) : this step jointly refines the camera pose and 3-D point cloud by minimizing the total re-projection error of the currently registered images.

Under small-parallax settings, initialization struggles to produce good initial two-view reconstruction due to unstable epipolar geometry, while incremental reconstruction tends to coverage to bad solutions due to large triangulation variation. We now show how these two steps can benefit from depth-prior obtained from a pretrained network. Note that we do not modify BA as our improved previous steps already provide a strong starting point where adding depth-prior to BA does not result in any additional gains.

3.2. Finding Keypoint-Depth

Given an image-set, we use standard COLMAP [33] pipeline for 2-D keypoints detection and matching. Moreover, we use a pretrained monocular depth-estimator to predict the dense depth map \mathbf{D}_i for each image \mathbf{I}_i . The depth of keypoint \mathbf{p}_i in \mathbf{I}_i is extracted from \mathbf{D}_i using bilinear interpolation as $\mathbf{D}_i[\mathbf{p}_i]$. We incorporate this keypoint-depth in the initialization step to get a more accurate estimate of the initial camera pose and 3-D point cloud, and regularize the optimization process of image registration and triangulation to guide the incremental reconstruction towards a better solution. Using the sparse keypoints-depth instead of dense depth map is important to maintain computation and memory efficiency for large scale reconstruction. We empirically demonstrate that our method is agnostic to the

choice of depth estimation model (see § 4.6).

3.3. Initialization

Instead of computing the essential matrix from 2-D to 2-D correspondences between the initial image-pair ($\mathbf{I}_a, \mathbf{I}_b$), and decomposing it into rotation and translation matrices as done in COLMAP [33], we incorporate keypoint-depth information to formulate the initialization step as a Perspective-n-Point (PnP) problem. Specifically, we first create an initial point cloud \mathbf{P}_a by projecting the 2-D keypoints in \mathbf{I}_a into 3-D as follows:

$$\mathbf{P}_a = \mathbf{D}_a[\mathbf{p}_a] \mathbf{K}_a^{-1} \mathbf{h}(\mathbf{p}_a) \quad \forall \mathbf{p}_a \in \mathcal{T}_a \quad (1)$$

where $\mathbf{D}_a[\mathbf{p}_a]$ is the depth of \mathbf{p}_a , $\mathbf{K}_a \in \mathbb{R}^{3 \times 3}$ is the intrinsic matrix of the camera that captured \mathbf{I}_a , $\mathbf{h}(\cdot)$ converts euclidean coordinates to homogeneous coordinates, and \mathcal{T}_a is the set of 2-D keypoints in \mathbf{I}_a . This gives us an initial 3-D point cloud created from keypoints in \mathbf{I}_a .

The relative pose between \mathbf{I}_a and \mathbf{I}_b is then estimated using geometric relationship between 2-D keypoints in \mathbf{I}_b and their corresponding 3-D points in the point cloud (2-D to 3-D correspondences), which is exactly the goal of the PnP problem. Instead of estimating the relative pose using 2-D to 2-D correspondences with epipolar geometry which is unstable under small baseline, using 2-D to 3-D correspondences with PnP approach makes our initialization method much more robust to small baseline since PnP naturally prefers small baseline data.

Note that unlike COLMAP [33] which selects the initial image-pair by considering both triangulation angle and the number of matched keypoints, we select the image pair which has the largest number of matched keypoints with valid depth. We consider depth to be valid for all values except 0 or infinity. Due to our large range of acceptable depth, more matched keypoints are used to generate the initial point cloud with larger scene-coverage, making subsequent reconstruction steps more robust and accurate.

3.4. Depth-Regularized Optimization

The initialization step is followed by: (a) image registration, which registers a new image to the existing scene and (b) triangulation, which triangulates the new points. We define a novel depth-regularized objective to improve these two steps. The intuition of our approach is illustrated in Figure 3 and its details are explained below.

a. Image Registration: We follow the procedure used in COLMAP [33] to select our next image \mathbf{I}_i , and estimate its initial camera pose using PnP problem formulation with RANSAC [5]. We further refine this initial camera pose by minimizing the following objective function:

$$\mathbf{R}_i^*, \mathbf{t}_i^*, \gamma_i, \beta_i = \arg \min_{\mathbf{R}, \mathbf{t}, \gamma, \beta} \sum_{\mathbf{p}_i \in \varphi_i} E_{\text{PR}}(\mathbf{p}_i) + \lambda E_{\text{DC}}(\mathbf{p}_i, \gamma_i, \beta_i) \quad (2)$$

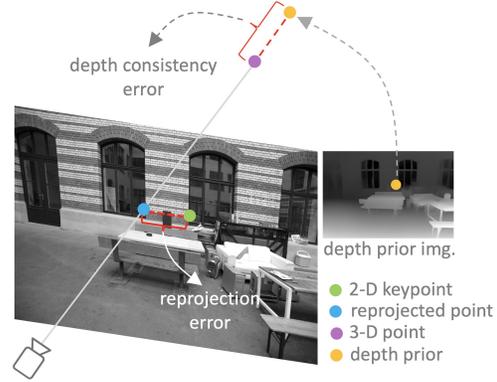


Figure 3: **Depth Regularized Optimization** – In addition to the generally used re-projection error, we use depth-consistency error as a regularizer for optimizing our loss functions for image registration (Equation 2) as well as triangulation (Equation 5).

Here, φ_i is the set of inliers keypoints obtained from RANSAC of initial pose estimation, while E_{PR} is the re-projection loss and E_{DC} is the depth consistency loss. λ is the weight to balance the two losses. E_{PR} is defined as:

$$E_{\text{PR}}(\mathbf{p}_i) = \|\Pi(\mathbf{R}_i \mathbf{P}_i + \mathbf{t}_i) - \mathbf{p}_i\| \quad (3)$$

where Π represents the projection from 3-D points to image plane. Similarly E_{DC} is defined as:

$$E_{\text{DC}}(\mathbf{p}_i, \gamma_i, \beta_i) = \sum_{\mathbf{p}_i \in \varphi(i)} \left| \|\mathbf{R}_i \mathbf{P}_i + \mathbf{t}_i\|_z - \gamma_i \mathbf{D}_i[\mathbf{p}_i] - \beta_i \right| \quad (4)$$

where $\mathbf{D}_i[\mathbf{p}_i]$ is the depth of keypoints \mathbf{p}_i , $[\mathbf{x}]_z \in \mathbb{R}$ ($\mathbf{x} \in \mathbb{R}^3$) is the third element of the 3-D point \mathbf{x} . γ_i and β_i are the scale and shift to align the depth-prior of \mathbf{I}_i with the projected depth from 3-D points..

b. Triangulation: Once image \mathbf{I}_i is registered, the newly observed scene points are added to the existing point cloud via triangulation. We first use DLT [14] and RANSAC [5] to estimate the initial 3-D position and refine it using the following objective function:

$$\mathbf{P}_i^* = \arg \min_{\mathbf{P}_i^{\text{new}}} \sum_{\mathbf{p}_j \in \mathbf{N}(\mathbf{P}_i^{\text{new}})} E_{\text{PR}}(\mathbf{p}_j) + \lambda E_{\text{DC}}(\mathbf{p}_j, \gamma_i, \beta_i) \quad (5)$$

where $\mathbf{P}_i^{\text{new}}$ is the new set of 3-D points observed in \mathbf{I}_i , $\mathbf{N}(\mathbf{P}_i^{\text{new}})$ is the set of 2-D keypoints corresponding to $\mathbf{P}_i^{\text{new}}$, and E_{PR} and E_{DC} are the reprojection and depth consistency errors as defined above. γ_i and β_i are computed from image registration and are kept fixed here. Note that the 3-D point estimated from triangulation based only on reprojection loss has large variance when the triangulation angle is small [21]. Our objective function addresses this challenge by regularizing the position of the 3-D point using the depth consistency error while keeping the reprojection error low.

4. Experiments

4.1. Datasets

We first go over the datasets we used in our experiments.

a. StudioSfM: To undertake a comprehensive comparative evaluation of our approach on studio-produced video-content, we collected a new dataset called StudioSfM which contains 130 shots with 21K frames from 15 TV-episodes. For each full-length TV episode, we first ran shot segmentation [35] to split it into a set of constituent shots and then sparsely sampled these shots in a uniform manner. For each sampled shot, we let professional visual-effects artists generate the ground-truth camera poses and 3-D point clouds through commercial CG software by manually tracking high-quality features, identifying co-planner constraints, and adjusting focal length. We removed the shots which were too challenging to be annotated due to factors such as heavy motion-blur and fully static camera.

To underscore the prevalence of small-baseline in studio-produced video-content, we compare parallax-distribution between StudioSfM dataset with a standard large-scale SfM dataset of ETH3D [34] (shown in Figure 1). We computed parallax as the ratio between the maximum translation of camera motion and the median distance of 3-D point cloud to all cameras. Figure 1 shows that most videos of StudioSfM have small parallax because the shots in movies and TV shows tend to have less camera motion to create an engaging viewing experience. In contrast, ETH3D has a much larger parallax since it is captured specifically for the purposes of 3-D reconstruction using standard approaches.

b. ETH3D: To demonstrate that our approach does not result in any accuracy loss for data with large parallax, we present experiments on ETH3D [34] which is a standard SfM dataset and contains two categories: (a) high-res multi-view with 13 scenes (b) low-res many-view with 5 scenes. The precise camera poses and dense point cloud from a laser scan are provided in the dataset.

4.2. Implementation Details

Our approach builds on the codebase of COLMAP [33]. We use DPT-large [30] as our default depth estimator for producing depth-priors. The influence of using different depth models on our method is analyzed in § 4.6. We resize input image height to 384 while maintaining the original aspect ratio. The dense depth map is resized to the original-image size using nearest neighbor interpolation. The weight λ for depth regularized optimization is always kept fixed at 6. Mask-RCNN [15] is used to create binary masks of humans which are used as input for all compared approaches.

4.3. Baselines

Unlike comparisons provided by previous approaches [19, 20] which only use original COLMAP [33] on videos with

small camera-motion, we fine-tune its hyper-parameters for small-parallax setting to make it much less likely to fail on small-parallax data in order to have a more fair comparison. We call this version of COLMAP [33] as COLMAP++, and compare our approach against the original COLMAP[33], COLMAP++, DeepV2D [38], RCVD [19] and DfUSMC [12].

4.4. Evaluation Metrics

To evaluate camera pose, we compute three commonly used metrics: absolute trajectory error (ATE), relative pose error for translation (T-RPE) and rotation (R-RPE). We refer to the work of [29] for detailed explanation of these metrics. To evaluate 3-D point cloud, we project point cloud to each frame using estimated camera-pose and measure the accuracy of relative depth $\delta = \max(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i})$ and absolute depth $\theta = |y_i - y_i^*|$ under different thresholds where y_i and y_i^* are the estimated and ground truth depth respectively.

4.5. Results

4.5.1 StudioSfM

a. Camera Pose Evaluation: We first evaluate the quality of the estimated camera pose on StudioSfM dataset. The predicted camera poses are aligned with the ground truth camera pose using similarity transformation before computing the metrics. Figure 4 shows the plot of recall against three error metrics and Table 1 shows the area-under-curve (AUC) for each curve. Our approach significantly outperforms other approaches across all three metrics. COLMAP++ performs much better than original COLMAP [33] showing the importance of tuning it to work with small parallax datasets. DfUSMC [12] does not work well on StudioSfM which indicates that their assumptions about camera-motion do not generalize to our data. DeepV2D [38] also shows low performance on StudioSfM likely due to their lack of outliers handling mechanisms.

To further clarify the benefit of our approach for small-parallax settings, we sort videos in StudioSfM data according to their parallax in descending order, and use the top 30% of data as the large-parallax set and bottom 30% as the small-parallax set. We compare our estimated camera pose with COLMAP++ using these two sets. Figure 5 shows the significantly better performance of our approach on small-

Method	ATE AUC		T-RPE AUC		R-RPE AUC	
	0.2 (cm)	2.0 (cm)	0.1 (cm)	0.5 (cm)	0.02 (°)	0.1 (°)
RCVD [19]	4.2	28.0	5.0	17.5	4.4	23.1
DfUSMC [12]	22.1	45.8	23.8	43.0	14.5	25.7
DeepV2D [38]	15.2	43.8	5.6	19.2	4.6	15.8
COLMAP [33]	20.0	49.6	23.1	45.8	25.8	46.7
COLMAP++	24.7	59.1	34.3	60.1	39.7	66.0
Ours	31.8+7.1	65.3+6.2	41.6 +7.3	69.8 +9.7	48.7 +9.0	74.8 +8.8

Table 1: Camera pose evaluation on StudioSfM using AUC –. Recall-curve AUCs for our three considered metrics are shown.

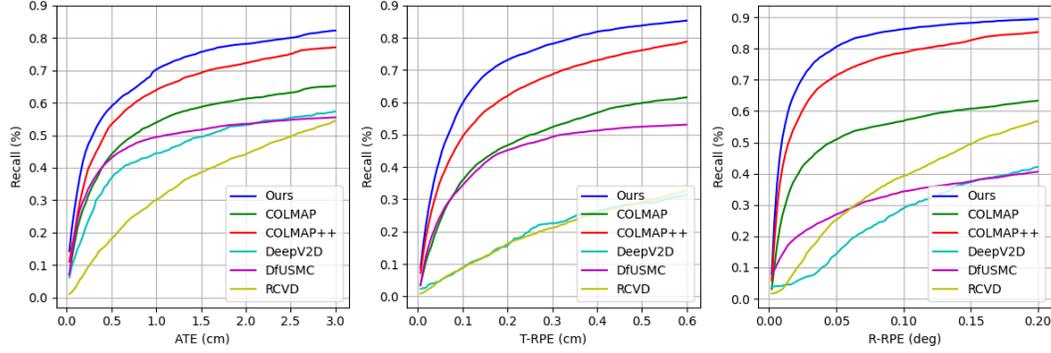


Figure 4: **Evaluation of camera pose on StudioSfM** – The figure shows the recall-curves of multiple comparative methods for ATE (absolute trajectory error), T-RPE (relative pose error for translation) and R-RPE (relative pose error for rotation).

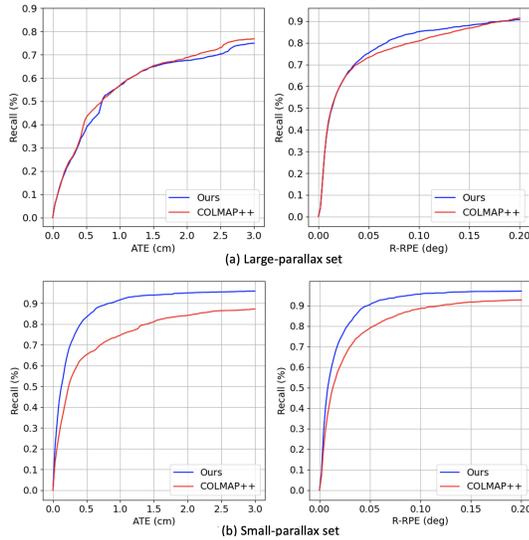


Figure 5: **Camera pose estimation using StudioSfM under:** (a) large-parallax set and (b) small-parallax set. Our approach offers significant improvement over COLMAP++ on small-parallax data.

parallax set, highlighting the importance of using depth-priors in geometry-based SfM for small-parallax settings.

b. Point Cloud Evaluation: To evaluate the quality of estimated point clouds, we first project point clouds into each frame using the estimated camera-pose and then compare depths of the projected points with ground truth depths. Besides computing the accuracy using relative depth error as done in DeepV2D [38], we also compare the accuracy using absolute depth error as our ground truth point clouds are annotated using real-world scale. Table 2 shows that our approach outperforms all other approaches on both relative and absolute depth error. Directly applying DPT-large [30] does not produce accurate depths even though they can visually look good. In contrast, our method of using the output of DPT-large [30] as depth-priors in geometry-based SfM substantially improves the quality of estimated depth.

Method	Relative depth accuracy (%)			Absolute depth accuracy (%)		
	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	$\theta < 5\text{cm}$	$\theta < 10\text{cm}$	$\theta < 25\text{cm}$
DPT-large* [30]	33.5	53.6	64.9	3.8	7.4	14.5
RCVD [19]	43.7	66.5	79.4	5.2	9.1	18.7
DfUSMC [12]	27.6	39.6	46.4	2.4	4.5	8.9
DeepV2D [38]	63.4	80.1	87.5	8.9	15.5	28.8
COLMAP [33]	50.8	55.1	56.9	20.7	27.3	38.3
COLMAP++	72.9	81.4	85.0	22.8	32.6	50.6
Ours	80.0	86.0	89.3	27.1	39.0	57.3

Table 2: **Evaluation of depth estimation on StudioSfM** – Recall is measured here using relative depth and absolute depth. Results for DPT-large [30] are presented here as a reference.

Method	ATE AUC		T-RPE AUC		R-RPE AUC	
	0.2 (cm)	2.0 (cm)	0.1 (cm)	0.5 (cm)	0.02 (°)	0.1 (°)
high-res multi-view						
COLMAP [33]	95.7	99.4	96.7	98.7	27.2	70.6
Ours	99.5	99.9	97.1	99.1	27.7	69.8
low-res many-view						
COLMAP [33]	18.6	74.3	65.8	92.5	0.5	7.5
Ours	42.1	88.8	86.4	96.9	0.4	14.8

Table 3: **Camera pose evaluation on the two categories of ETH3D using AUC** – We report AUC of each metric. Our approach achieves results comparable with COLMAP on high-res multi-view category for all metrics, and outperforms COLMAP on low-res many-view category for most metrics.

4.5.2 ETH3D

To demonstrate the effectiveness of our approach on standard SfM datasets, we assess it on two categories of ETH3D [34] where motion-parallax is significantly larger than StudioSfM. Our approach is compared with original COLMAP [33] which is already tuned for large-parallax. The camera pose comparison is presented in Table 3. On high-res multi-view category both COLMAP [33] and our method achieve impressive performance while our method is still able to slightly outperform COLMAP [33]. Our clear gains over COLMAP [33] on low-res many-view category show that our approach is more robust to low resolution images than COLMAP [33]. The comparison of estimated depth using high-res multi-view category is shown in Ta-

Method	Relative depth accuracy (%)			Absolute depth accuracy (%)		
	$\delta < 1.05$	$\delta < 1.05^2$	$\delta < 1.05^3$	$\theta < 1\text{cm}$	$\theta < 2\text{cm}$	$\theta < 5\text{cm}$
COLMAP [33]	96.9	98.1	98.5	58.7	72.9	86.0
Ours	96.8	98.0	98.4	61.2	75.7	88.1

Table 4: **Evaluation of depth estimation on ETH3D high-res multi-view category** – Accuracy is measured here using both relative depth and absolute depth.

ble 4 in which we achieve better absolute depth accuracy than COLMAP [33]. Our overall better performance on ETH3D demonstrates that our approach does not show any degradation on large-parallax data while offering significant gains for small-parallax settings.

4.6. Ablation Study

a. Method Variants: We compare several variants of our approach with COLMAP++ on StudioSfM dataset. Figure 6 compares the recall curves for ATE and R-RPE between COLMAP++, our approach with only improved initialization (initialization only), our approach with only depth-regularized optimization (optimization only) and our full approach (ours full). We can see that our proposed initialization using depth-prior of keypoints achieves substantial improvement over COLMAP++ showing the criticality of initialization for SfM pipeline to converge to a good solution. With both improved initialization and depth regularized optimization, our full approach performs the best.

b. Depth Estimators: To assess the robustness of our approach to the choice of depth-estimator, we evaluate camera pose estimation using several off-the-shelf pretrained depth estimation models based on various network architectures and trained with different datasets. Specifically, we compare five monocular depth estimation models, including MiDaS small [31] which is designed for mobile devices, DPT-hybrid [30] and DPT-large [30] which are based on Transformers [41], AdaBins [4] which is the latest approach for monocular depth estimation and MC [22] which focuses on human depth estimation. Figure 7 shows that our approach significantly outperforms COLMAP++ using depth priors provided by any of the five different pretrained depth estimation models. The small performance variation among those depth estimators demonstrates that our approach does not rely on a particular depth estimator and is robust to diverse network architectures and training datasets.

c. Depth Noise: In addition to evaluating the use of various depth estimators we also test the robustness of our approach under different amounts of synthetic noise. For each keypoint depth d , we add random Gaussian noise with 0 mean and $\alpha \cdot d$ standard deviation with different values of α . As shown in Table 5, the performance degradation of our approach is only within 5% under the largest added noise level of 0.4 which demonstrates that our pipeline can tolerate significant amounts of errors in the estimated depth-priors.

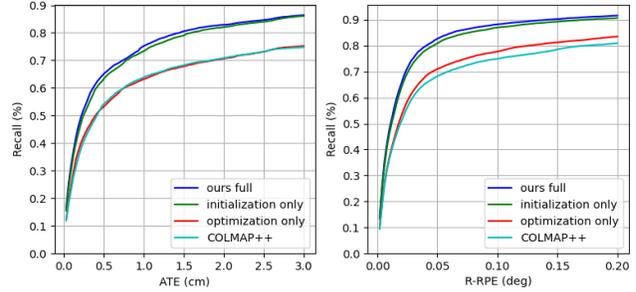


Figure 6: **Ablation study on StudioSfM** – Recall of translation error and relative rotation error are plotted for different variants of our method: "ours full" - our full approach, "initialization only" - our approach with only improved initialization, "optimization only" - our approach with only depth-regularized optimization.

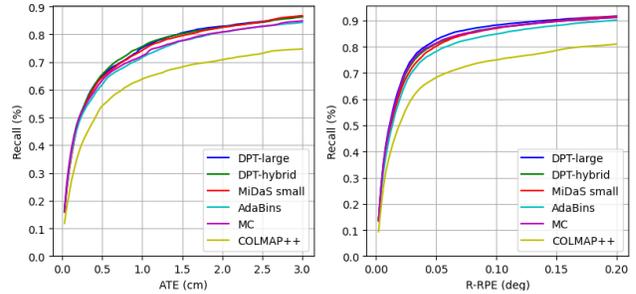


Figure 7: **Ablation study of depth-estimators for StudioSfM** – The recalls for ATE and R-RPE are plotted for our approach with different pretrained monocular depth estimators.

α	ATE AUC		T-RPE AUC		R-RPE AUC	
	0.2 (cm)	2.0 (cm)	0.1 (cm)	0.5 (cm)	0.02 ($^\circ$)	0.1 ($^\circ$)
0.0	31.8	65.3	41.6	69.8	48.7	74.8
0.1	31.1	63.9	39.4	67.6	47.1	73.0
0.2	30.3	62.8	39.2	67.5	47.0	73.5
0.4	28.5	62.7	38.5	67.0	44.6	71.9

Table 5: **Depth noise analysis for camera pose estimation** – Gaussian noise with 0 mean and $\alpha \cdot d$ variance is added to each keypoint with depth d . The performance degradation of our approach is only within 5% for α of 0.4 demonstrating the robustness of our approach to errors in estimated depth-priors.

4.7. Qualitative Evaluation

a. StudioSfM: We compare our approach with other methods qualitatively using five examples from StudioSfM in Figure 8. To compare with RCVD [19], we use their estimated depth image to visualize the depth of the point cloud. Examples 1-4 show a common error observed for COLMAP++ where, unlike our approach, the relative depths between points are incorrect (e.g., the building outside of window in example 2 is estimated closer than the table in the room). Similarly, the camera motion estimated by RCVD [19] tends to have large errors as shown in example 2-5. Both COLMAP++ and our approach achieve accurate reconstruction for example 5 since the motion-parallax is



Figure 8: **Qualitative results on StudioSfM dataset** – Keypoint-depths are visualized in color from red (near) to blue (far) and the camera motion is visualized as a trajectory of red cones. First column shows image and ground truth camera motion, while other columns show the results from different approaches.

sufficient, however, RCVD [19] still produces poor results for this example even though the motion parallax is large.

b. LVU Dataset: We now present qualitative results of our approach and COLMAP++ on a subset of the LVU dataset [44] consisting of video clips from movies. We selected 53 shots with relatively few dynamic objects and small motion blur from the test-set of category "scene". As there is no ground truth provided, we can only evaluate the results by visualizing the camera poses and point clouds. Out of the selected 53 shots, we did not find any shot where results from COLMAP++ were clearly better than ours. Figure 9 shows results of 5 examples demonstrating the higher-quality results produced by our approach. The last row shows an example where our approach produces similar errors as COLMAP++. This is because the estimated depth images [30] of initial image-pair for this example are too erroneous for our approach to effectively guide the subsequent reconstruction process to a better solution.

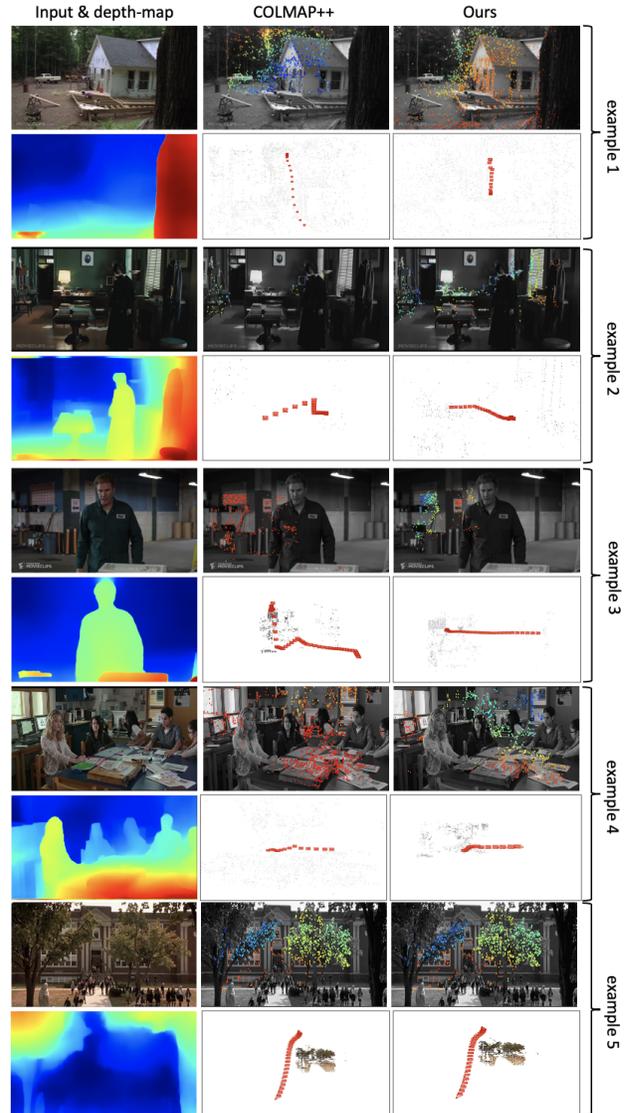


Figure 9: **Qualitative results on LVU dataset [44]** – Depths are visualized using red (near) to blue (far). 1st column shows the input image and its depth, while 2nd and 3rd columns show the results of COLMAP++ and our approach.

5. Conclusions

We presented a simple yet effective SfM approach that uses monocular depth obtained from a pretrained network to improve the incremental SfM pipeline [33]. Experiments using existing and a newly collected dataset show that our approach significantly improves the reconstruction quality for small parallax data while being robust to a variety of pretrained depth networks. Our approach easily integrates with COLMAP [33], and going forward we plan to use it as an initial step for dense reconstruction and novel view synthesis for studio-produced content.

References

- [1] Mapillary AB. Opensfm - open source structure-from-motion pipeline. <https://github.com/mapillary/OpenSfm>, 2019.
- [2] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. Building rome in a day. *International Conference on Computer Vision (ICCV)*, 2009.
- [3] Alex M Andrew. Multiple view geometry in computer vision. *Kybernetes*, 2001.
- [4] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [5] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized ransac. *DAGM-Symposium*, 2003.
- [6] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017.
- [7] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. *CVPR Workshop on Deep Learning for Visual SLAM*, 2018.
- [8] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. 1981.
- [9] Y. Furukawa and C. Hernandez. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 2015.
- [10] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.
- [11] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. *International Conference on Computer Vision (ICCV)*, 2019.
- [12] Hyowon Ha, Sunghoon Im, Jaesik Park, Hae-Gon Jeon, and In So Kweon. High-quality depth from uncalibrated small motion clip. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [13] Chris Harris and Mike Stephens. A combined corner and edge detector. *Proceedings of the 4th Alvey Vision Conference*, 1988.
- [14] Richard Hartley and Andrew Zisserman. Multiple view geometry. *Computer Vision*. Cambridge University Press, 2004.
- [15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.
- [16] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision (IJCV)*, 2020.
- [17] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.
- [18] Faisal Khan, Saqib Salahuddin, and Hossein Javidnia. Deep learning-based monocular depth estimation methods—a state-of-the-art review. *Sensors*, 20(8):2272, 2020.
- [19] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [20] Zihang Lai, Sifei Liu, Alexei A Efros, and Xiaolong Wang. Video autoencoder: self-supervised disentanglement of 3d structure and motion. In *International Conference on Computer Vision (ICCV)*, 2021.
- [21] Seong Hun Lee and Javier Civera. Triangulation: why optimize? *British Machine Vision Conference (BMVC)*, 2019.
- [22] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4521–4530, 2019.
- [23] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [24] David Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, pages 91–110, 2004.
- [25] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM TOG (Proc. SIGGRAPH)*, 2020.
- [26] Yi Ma, Stefano Soatto, Jana Kosecka, and Shankar Sastry. An invitation to 3-d vision: From images to geometric models. *Springer Verlag*, 2004.
- [27] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *European Conference on Computer Vision (ECCV)*, 2020.
- [28] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. Openmvg: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 60–74. Springer, 2016.
- [29] David Prokhorov, Dmitry Zhukov, Olga Barinova, Konushin Anton, and Anna Vorontsova. Measuring robustness of visual slam. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–6. IEEE, 2019.
- [30] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.
- [31] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.
- [32] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. *International Conference on Computer Vision (ICCV)*, 2011.

- [33] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [34] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [35] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, Miguel Bugalho, and Isabel Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 2011.
- [36] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *European Conference on Computer Vision (ECCV)*, 2012.
- [37] Chengzhou Tang and Ping Tan. Ba-net: Dense bundle adjustment networks. *ICLR*, 2019.
- [38] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *ICLR*, 2020.
- [39] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment a modern synthesis. *International workshop on vision algorithms*, 1999.
- [40] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [42] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. Deepsfm: Structure from motion via deep bundle adjustment. In *European Conference on Computer Vision (ECCV)*, 2020.
- [43] Kyle Wilson and Noah Snavely. Robust global translations with 1dsfm. *European Conference on Computer Vision (ECCV)*, 2014.
- [44] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [45] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [46] Fisher Yu and David Gallup. 3d reconstruction from accidental motion. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [47] T. Zhou, M. Brown, N. Snavely, and D.G Lowe. Un-supervised learning of depth and ego-motion from video. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.