Incremental Structure-from-Motion through Affine Transformations of Monocular Depth Priors

Isac Olsteg

Master's thesis
2025:E50

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Centrum Scientiarum Mathematicarum

Incremental Structure-from-Motion through Affine Transformations of Monocular Depth Priors

Isac Olsteg

LUND
UNIVERSITY

Department of Mathematics

Abstract

This thesis presents an incremental Structure-from-Motion pipeline through the use of affine transformations of monocular depth priors. The scene points are triangulated by averaging lifted covisible keypoints utilizing the calculated transformations. Each new image is registered to the reconstruction by first estimating the camera pose, then aligning the depth prior to the observed 3D points. We propose a bundle adjustment method based on reprojection errors and depth deviations through uniform sampling across multi-view shared 2D-2D correspondences. The final reconstruction is found through backprojection of the registered images using the transformed depth priors. We achieve improved completeness and F_1 score compared to the widely adopted COLMAP pipeline, but show that this comes at the cost of accuracy. Furthermore, we observe improved robustness and accuracy in terms of pose estimation for low three-view overlap in images.

Populärvetenskaplig sammanfattning

I denna mastersuppsats undersöks möjligheten att utnyttja transformationer av djupuppskattningar från en mängd bilder med målet att återskapa den fotograferade miljön. Med hjälp av dessa lyckades modellen producera mer fullständiga rekonstruktioner jämfört med tidigare metoder, samt bättre hantera utmanande fall med lite tre-vy överlapp mellan bilder.

Problemet att kunna återskapa miljöer från bilder har länge studerats inom datorseende. Med tiden har tillgången till bildbibliotek och kameror på mobiltelefoner gjort området mer lättillgängligt. Jämfört med att transportera dyr hårdvara, som exempelvis LiDAR skannrar, möjliggör dessa metoder bland annat mer tidseffektiva och billigare alternativ.

Rekonstruktionerna består av en mängd 3D punkter i så kallade punktmoln. Punkterna beräknas från intressanta punkter som återfinns i flera bilder i form av matchningar. I konventionella metoder kan endast 3D punkter beräknas från dessa matchade punkter. På så vis är storleken av punktmolnet direkt kopplad till mängden extraherade punkter. Tätare matchningar löser detta till viss utsträckning, men är generellt mycket tidskrävande.

I detta arbete används monokulära djupkartor för att bemöta problemet. Detta innebär att varje djupuppskattning enbart baseras på information från en bild. Djupkartorna innehåller ett djupvärde för varje pixel i en bild, och kan användas för att återfå 3D strukturen. Vi estimerar affina transformationer för att se till att geometrin från överlappande bilder sammanfaller. På detta vis begränsas ej denna metod av antalet matchningar och kan generera mer fullständiga rekonstruktioner.

Ett ytterligare problem som hanteras i denna metod är kravet för tre-vy överlapp vid registrering av en ny bild till rekonstruktionen. I en iterativ struktur från rörelse modell estimeras placering och rotation för en kamera åt gången. Som följd av hur 3D punkterna beräknas, kräver detta gemensamma observationer i åtminstone tre bilder. De transformerade djupkartorna tillåter att 3D punkter kan tas fram från endast en vy, vilket möjliggör att endast överlapp i två bilder är nödvändigt. Således kan denna metod hantera fall som traditionella metoder ej klarar av. Även om tre bilder har gemensamma observationer visade vår analys att möjligheten att använda två-vys överlapp gynnade metoden till viss utsträckning.

Acknowledgements

I would like to thank my supervisors Viktor Larsson and Gustav Hanning for providing me with valuable insights throughout this project. Furthermore, I would also like to thank the authors of the paper *MP-SfM: Monocular Surface Priors for Robust Structure-from-Motion* for providing me with the triplet data used in the results. Lastly, I would also like to thank the Department of Mathematics at Lund University for granting me access to the necessary hardware.

Contents

1. Introduction **12**
1.1 Problem Statement . . . 12

2. Theory **13**
2.1 Fundamentals of Computer Vision . . . 13
2.2 Feature Extraction and Matching . . . 14
2.3 Structure-from-Motion . . . 14
2.4 Monocular Depth Estimation . . . 16

3. Methodology **21**
3.1 Initialization . . . 21
3.2 Next View Selection . . . 22
3.3 Bundle Adjustment . . . 23
3.4 De-Registering Images . . . 24
3.5 Final Reconstruction . . . 24

4. Results **25**
4.1 Reconstruction Results . . . 25
4.2 Pose Estimation . . . 25
4.3 Low-overlap Pose Estimation . . . 26
4.4 Ablation Studies . . . 28
4.5 Reconstruction Visualizations . . . 29

5. Discussion **31**
5.1 Accuracy and Robustness . . . 31
5.2 Completeness and Reconstruction Quality . . . 32
5.3 Ablation Studies . . . 32

6. Conclusion and Future Work **34**
6.1 Conclusion . . . 34
6.2 Future Work . . . 34

Bibliography **35**

A. Appendix **37**
A.1 Complete Reconstruction Results . . . 37
A.2 Complete Pose Estimation Results . . . 38
A.3 Complete Low-overlap Pose Estimation Results . . . 40

List of abbreviations

SfM: Structure-from-Motion

MDE: Monocular Depth Estimation

GBA: Global Bundle Adjustment

LBA: Local Bundle Adjustment

SIFT: Scale-invariant Feature Transform

1

Introduction

Structure-from-Motion (SfM) is one of the most widely studied computer vision problems. From a collection of unordered images, the goal is to reconstruct the 3D geometry of the scene represented as 3D points known as a *point cloud*. Additionally, the position and orientation of the cameras behind the images are to be determined, referred to as *camera poses*. Optionally, remaining camera parameters, such as the focal length or distortion parameters, may also need to be estimated.

Conventionally, SfM pipelines are divided into two different categories. Global SfM approaches estimate the full point cloud and all cameras simultaneously. Incremental SfM on the other hand, which is the approach adopted in this thesis, starts from an initial image pair and iteratively expands the scene by registering images one by one. The latter is often considered more robust and accurate, but is less scalable compared to global approaches [1]. Both methods commonly start with feature extraction and matching. These 2D-2D correspondences of interesting points, known as *keypoints*, can either be learned or hand-crafted [2]. For smaller collections of images, each pair can be exhaustively matched. Larger datasets may require the need for global descriptors, which allows the matching to be limited to the most similar images [3].

The locations of the 3D points in the scene are found through *triangulation*[4]. To solve this problem, the model needs to be aware of which images observe the same point. The shared observations and their respective images can be stored using graphs known as *tracks*. Traditional SfM pipelines, such as COLMAP[1], allow the model to triangulate one scene point for each track. Sparse feature matchers will therefore generate less complete reconstructions, as the overall number of tracks are few. Dense matches on the other hand, such as RoMa[5], which try to match all pixel-pairs between images, lead to larger reconstructions, but at the cost of an increase in the overall runtime [1].

We aim to address this problem by producing dense reconstructions from sparse matches through the use of monocular depth priors. The monocular depth of an image refers to the predicted depth of each pixel based solely on the image itself [6]. Assuming known (or estimated) camera intrinsics, the depth maps can be used to recover the 3D geometry, or vice versa. As the scale of two monocular depth estimates of the same scene might not coincide, this poses a new set of challenges for the SfM pipeline which we intend to address.

1.1 Problem Statement

The goal of this thesis is to develop an incremental SfM pipeline utilizing affine transformations of monocular depth priors. By aligning the depth estimates along each track, we aim to triangulate a sparse point cloud and register a set of images. From this, based on the camera poses and transformed depth priors we strive to recover a more complete 3D geometry. Ideally, the model is to produce dense reconstructions from a set of sparse 2D-2D correspondences, something not feasible through a typical SfM pipeline. Furthermore, we intend to address failure cases inherent in these models as a result of the requirement of three-view overlap to register new images to the reconstruction.

2

Theory

In this chapter, we present some basic relevant theory required to understand the thesis moving forward. Furthermore, we give detailed explanations of past work. Including arguably the most widely adopted SfM pipeline, and work which form the basis of this thesis.

2.1 Fundamentals of Computer Vision

Any 2D point (x, y) can be extended with a third coordinate 1 creating a triplet $(x, y, 1)$ representing the same point [7]. Triplets differing by a non-zero multiple are said to represent the same point. Homogeneous coordinates refer to these classes of equivalent coordinate triplets. As an example, this means that $(x, y, 1)$ and (sx, sy, s) for any non-zero s represent the same point (x, y), which we recover by dividing by the third coordinate. If the third coordinate is zero we refer to it as a vanishing point. Similarly we can represent three dimensional points (X, Y, Z) in homogeneous coordinates. By mapping 3D homogeneous coordinates with the camera matrix $\mathbf{P} \in \mathbb{R}^{3\times4}$ we obtain the corresponding point in pixel coordinates [4]. We can hereby formulate the camera equations:

$$\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{P} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{2.1}$$

where $\lambda \in \mathbb{R}$. Furthermore, the camera matrix can be deconstructed as $\mathbf{P} = \mathbf{K} \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix}$, where $(\mathbf{R}, \mathbf{t})$ are the extrinsic camera parameters and $\mathbf{K} \in \mathbb{R}^{3\times3}$ the intrinsic. The orthonormal rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times3}$ and translation vector $\mathbf{t} \in \mathbb{R}^3$ give a transformation between the global and camera specific coordinate systems, see fig. 2.1. The calibration matrix $\mathbf{K}$ maps the image plane in the camera coordinate system to the actual image in pixel coordinates. In this thesis we use the following form

$$\mathbf{K} = \begin{bmatrix} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.2}$$

where f is the focal length and (p_x, p_y) is the principal point. By convention the latter is often chosen as the center of the image in pixel coordinates.

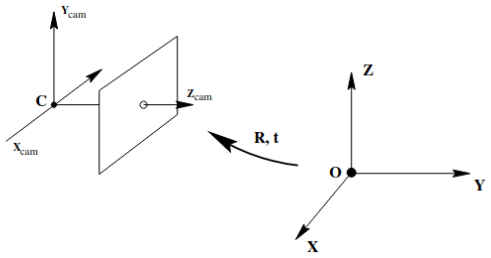


Figure 2.1 The camera extrinsics $(\mathbf{R}, \mathbf{t})$ relates the global coordinate system with the camera coordinate system. The latter is defined with the origin as the camera center and the z-axis facing towards the image plane located at $z = 1$. Image credit: [7].

2.1.1 Triangulation

Assuming known camera parameters and corresponding keypoints $\mathbf{x}_j$ in a set of images, we want to estimate the position of a 3D point $\mathbf{X}$. This problem is known as triangulation [4]. Ideally the solution would be found at the intersection point of the viewing rays, but due to noise this is not guaranteed. One common approach is the direct linear transform (DLT) method [7]. We re-formulate the camera equations eq. (2.1) and combine them for each view $i = 1, \ldots, n$:

$$\underbrace{\begin{bmatrix} \mathbf{P}_1 & -\mathbf{x}_1 & 0 & \ldots & & 0 \\ \mathbf{P}_2 & 0 & -\mathbf{x}_2 & 0 & \ldots & 0 \\ \vdots & & & \ddots & & \vdots \\ \mathbf{P}_{n-1} & 0 & \ldots & & & \\ \mathbf{P}_n & 0 & & & 0 & -\mathbf{x}_n \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} \mathbf{X} \\ \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix}}_{\mathbf{v}} = \mathbf{0}. \tag{2.3}$$

According to the DLT algorithm, the solution minimizing $||\mathbf{A}\mathbf{v}||^2$ subject to $||\mathbf{v}|| = 1$ is given by the last left-singular vector from the singular value decomposition of $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ (i.e. the last column of $\mathbf{V}$). The position of the scene point is thus given by the first three entries of $\mathbf{v}$.

2.1.2 Perspective-n-Point

The Perspective-n-Point (PnP) problem refers to the task of estimating the camera pose from a set of n scene points with 2D correspondences [8]. It has been shown that a minimum of 3 points are needed to obtain a finite set of solutions. This is known as the P3P problem and is highly relevant as outlier contamination can heavily skew the final pose. To combat faulty 2D-3D correspondences it is common to employ a random sample consensus (RANSAC) paradigm. Iteratively, a small subset of data points are randomly sampled and fitted to the problem. Evidently, the sample set must be large enough to obtain a valid solution. The subset of points that are within some defined error tolerance of the model is known as the consensus set. Unless the process terminates prematurely, the solution with the largest consensus set is deemed to be the correct one.

It can be shown that the P3P problem has at most four solutions [8]. As it is an old and heavily studied problem many solutions have been proposed [9]. In this thesis we use the solver proposed by Ding *et al.* [9] which bases its solution on the intersection of two conics.

2.2 Feature Extraction and Matching

Finding locations in pairs of images that refer to the same point is a fundamental part of many computer vision applications [4]. Interesting points in an image are known as keypoints. Each keypoint has an associated feature vector which describes it. One of the most widely used methods is the Scale-invariant feature transform (SIFT) [10]. But in recent years it has become more common to leverage learned features, such as Superpoint[11].

Keypoints can be used to find covisible points between images. Similar to feature extraction, these matches can benefit from deep learning. Examples of such methods include SuperGlue[12] and LightGlue[13]. Nearest neighbor matching is an example of a method which does not use machine learning. For smaller datasets it is possible to exhaustively match every image pair. However, for larger image collections this becomes unfeasible and unnecessary. In such cases it can be useful to leverage global descriptors to only match the most similar images [3].

2.3 Structure-from-Motion

The process of reconstructing 3D structures from a collection of images is referred to as Structure-from-Motion [1]. It typically takes a collection of $i = 1, \ldots, n$ unordered images $\mathcal{I} = \{I_i \in \mathbb{R}^{H_i \times W_i \times 3}\}$ as input and outputs scene points $\mathcal{X} = \{\mathbf{X}_k \in \mathbb{R}^3 \,|\, k = 1, \ldots, N_\mathcal{X}\}$ and a set of registered poses $\mathcal{P} = \{\mathbf{P}_i \in \mathbf{SE}(3) \,|\, i = 1, \ldots, N_\mathcal{P}\}$. If the reconstruction is updated in a

sequential matter it is known as incremental SfM. Global SfM instead jointly estimates all camera poses simultaneously, which are then used to triangulate the scene points followed commonly by a final global refinement process known as global bundle adjustment (GBA) [14]. While both incremental and global SfM pipelines typically begin similarly through image feature extraction and matching, the way they differ in camera geometry and 3D structure estimation make them appropriate in different situations. The need of frequent bundle adjustment to prevent drift related issues benefits incremental SfM pipelines in terms of robustness and accuracy, but at the cost of a loss in scalability. On the contrary, global SfM models are typically more scalable but less robust and accurate. Since incremental SfM has seen a variety of different approaches, we will demonstrate the general idea by summarizing arguably the most widely adopted SfM pipeline COLMAP [1, 15].

2.3.1 Feature Matching and Geometric Verification

As mentioned previously, from a collection of $i = 1, \ldots, n$ unordered images $\mathcal{I} = \{I_i \in \mathbb{R}^{H_i \times W_i \times 3}\}$ we extract a set of features for each image. The feature locations, $\mathbf{x}_j \in \mathbb{R}^2$, are represented by feature vectors $\mathbf{f}_j$, which make up the set of features $\mathcal{F}_i = \{(\mathbf{x}_j, \mathbf{f}_j) \,|\, j = 1, \ldots, N_i\}$. By default, COLMAP uses SIFT features but is compatible with other types of feature extractors [1, 2]. Next, the image features are matched by comparing the set of features between two images. The most similar feature in the second image for each of the features in the first are matched by comparing the feature vectors to some similarity metric. Matching features solely based on visual descriptors do not ensure the matches make sense geometrically. To mitigate this issue it is common to perform geometric verification between the matches for images believed to be overlapping. In COLMAP this is done by estimating transformations based on the spatial configuration between an image pair. By using a RANSAC estimation technique to reduce the influence of outliers, they map keypoints from one image to the other. Mapped keypoints sufficiently close to their correspondences are then counted as inliers and assumed to be geometrically verifiable.

2.3.2 Image Registration

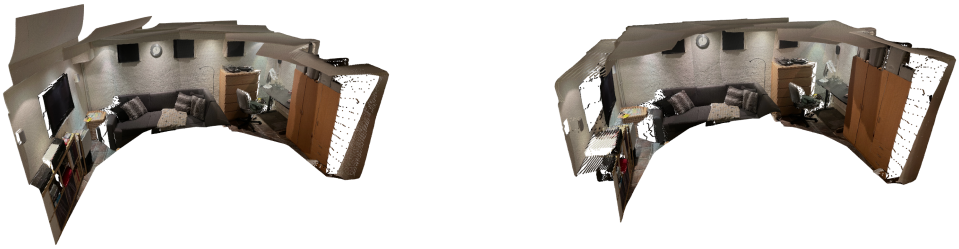
The quality of the final reconstruction is heavily dependent on the initial image pair. To deal with this, COLMAP carefully selects the initial pair based on the number of inlier matches from a location in the image graph with many overlapping images [1]. Starting the process with a two-view reconstruction from a sparser part of the image graph often leads to less robust and accurate reconstructions. Iteratively, based on the current scene points and registered images, a new view has to be selected. As incremental SfM is prone to accumulated errors, selecting a good next image is critical to mitigate such errors while also providing more accurate pose estimates and triangulated scene points. In COLMAP, they consider the images not yet registered which observe a set number of scene points above a certain threshold. Unregistered images with many observations where the 2D correspondences follow a more uniform distribution are ranked higher. In certain cases this allows them to rank images with fewer observation following a more even distribution above images with many observations clustered together. This would not be possible for a ranking structure based solely on the number of visible scene points. Once the next view has been selected the pose is estimated by solving the PnP problem for the set of 2D-3D correspondences between the image keypoints and visible scene points. With the new pose estimated it is possible to re-triangulate existing points or triangulate new ones. For this COLMAP uses an outlier robust RANSAC approach based on the DLT method.

2.3.3 Bundle Adjustment

To prevent the structure from reaching a non-recoverable state as a result of drifting following consecutive image registrations, bundle adjustment is a necessary step in incremental SfM [1], see fig. 2.2. The set of registered images $\mathcal{P}$ and scene points $\mathcal{X}$ are jointly refined by applying a function π that projects the scene points into an image with a 2D correspondence to minimize the following cost function

$$\sum_{i=1}^{N_\mathcal{P}} \sum_{k=1}^{N_\mathcal{X}} \sum_{j} \rho\left(||\pi(\mathbf{K}_i, \mathbf{P}_i, \mathbf{X}_k) - \mathbf{x}_j||^2\right), \tag{2.4}$$

where ρ is a loss function. As registering new images only leads to local changes in incremental SfM, it is redundant to perform global bundle adjustment if not necessary. Therefore, local bundle adjustment (LBA) is often performed. For instance, it can be performed on the set of most-connected images following each newly registered image. Global bundle adjustment is therefore reserved for special cases, such as when the amount of scene points grow by a certain percentage, or periodically after a fixed amount of newly registered images.



(a) Before Bundle Adjustment **(b)** After Bundle Adjustment

Figure 2.2 By periodically refining the geometry through bundle adjustment we reduce accumulated errors.

Bundle adjustment problems have a tendency to grow very large as the number of residuals and parameters is directly linked to the number of keypoint measurements, scene points and images. At a certain point, solving these non-linear least squares problems becomes unfeasible for dense solvers [4]. However, note that each feature location in an image only depends on one scene point and the camera parameters of the image. Therefore, if either the cameras or points are known, the equations for the other become independent. By reordering the variables it is possible to exploit the sparsity of this problem to more effectively solve it.

2.4 Monocular Depth Estimation

Monocular depth estimation (MDE) refers to the task of estimating the depth of each pixel from a single image [6]. The absence of reliable visual cues and stereoscopics makes this an ill-posed problem [16]. By estimating a depth map it is possible to recover the 3D geometry through unprojection using the camera intrinsics [17]. This process is known as monocular geometry estimation. Without strong visual cues, estimating the camera intrinsics remains challenging for a monocular image, leading to distortions of the 3D geometry. Recent work has tried to circumvent this by directly estimating point clouds, omitting estimation of the camera intrinsics altogether when recovering the 3D geometry [18]. However, the use of scale-invariant point maps still makes such models susceptible to focal-distance ambiguity when inferring it from the point cloud [17]. By instead utilizing an affine-invariant point map it is possible to eliminate this ambiguity. From the depth map estimate, it possible to *backproject* the image to get the 3D geometry, see fig. 2.3. For sets of points we also may refer to this as *single-view lifting*, or simply *lifting*.

(a) Original Image **(b)** Estimated Monocular Depth

(c) Backprojected Image

Figure 2.3 The monocular depth estimate of an image can be used to recover the 3D geometry through backprojection.

2.4.1 Related work

In this subsection we go through related work regarding MDE. The MDE model of choice is presented, and a previous approach of incorporating monocular depth priors into an SfM model is described.

MoGe MoGe[17] is a state-of-the-art monocular geometry model which directly predicts an affine-invariant 3D point cloud from a single image. Through global and local loss functions it favors the model in learning consistent and high-quality 3D geometry, while still maintaining optimized inference speed. As it is the MDE model of choice we will go through in detail how the loss functions are designed to achieve better accuracy compared to past methods while maintaining generalizability. The network structure uses a pretrained vision transformer encoder and a lightweight CNN-based upsampler for the decoding.

The coordinate system of the predicted point cloud and the ground truth need to be aligned to be comparable. To determine the global loss, the alignment parameters s and $\mathbf{t}$ are used to transform the former camera space into the latter. Let $\hat{\mathbf{p}}_i$ and $\mathbf{p}_i$ denote the predicted and ground truth 3D points for the i-th pixel. Only the non-infinity region $\mathcal{M}$, is considered when computing the global loss. This means that regions containing skies or objects with undefined geometry are not considered. To handle extreme variations in depth they normalize the distance between the predicted and ground truth 3D points by z_i, the z-coordinate of $\mathbf{p}_i$. The global loss is thus defined as

$$\mathcal{L}_G = \sum_{i \in \mathcal{M}} \frac{1}{z_i} ||s\hat{\mathbf{p}}_i + \mathbf{t} - \mathbf{p}_i||_1, \tag{2.5}$$

where $||\cdot||_1$ is the L_1-norm. The absolute residuals are truncated to improve robustness. Note that the alignment parameters are unknown and have to be estimated. These are simply chosen as the pair which minimizes the global loss

$$(s^*, \mathbf{t}^*) = \arg\min_{s,\mathbf{t}} \mathcal{L}_G, \tag{2.6}$$

where t_x and t_y are equal to 0.

Models solely focused on global alignment neglect ambiguities in relative depth between objects. This results in poor local geometry. MoGe considers spheres with varying radii r_j and uses independent alignment for each 3D point. The region around $\mathbf{p}_j$ is defined as

$$\mathcal{S}_j = \{i \mid ||\mathbf{p}_i - \mathbf{p}_j|| \leq r_j,\, i \in \mathcal{M}\}. \tag{2.7}$$

The local loss is then computed similar to the global loss but over the set of anchor points $\mathcal{H}_\alpha$, where α determines the scale of the sphere

$$\mathcal{L}_{\mathcal{S}(\alpha)} = \sum_{j \in \mathcal{H}_\alpha} \sum_{i \in \mathcal{S}_j} \frac{1}{z_i} ||s_j^* \hat{\mathbf{p}}_i + \mathbf{t}_j^* - \mathbf{p}_i||_1. \tag{2.8}$$

Even with local supervision the surface quality of the geometry may be uneven. The angular difference $\angle(.,.)$ between the estimated normals $\hat{\mathbf{n}}_i$ and ground truth $\mathbf{n}_i$ for each pixel i is computed. The loss function

$$\mathcal{L}_N = \sum_{i \in \mathcal{M}} \angle(\hat{\mathbf{n}}_i, \mathbf{n}_i), \tag{2.9}$$

is used to supervise and improve the surface quality. As mentioned previously, only the masks with labels $\mathcal{M}$ are considered during supervision. To determine these, they formulate the loss function

$$\mathcal{L}_M = ||\hat{\mathbf{M}} - (1 - \mathbf{M}_{\text{inf}})||_2^2, \tag{2.10}$$

where $\hat{\mathbf{M}} \in \mathbb{R}^{H \times W}$ is the predicted mask and $\mathbf{M}_{\text{inf}}$ is the ground truth mask with the infinity labels. The predicted mask is thresholded to obtain the valid points.

From the predicted 3D point cloud it is possible to recover the focal length f and the depth map. The solution is found by minimizing the error of the projection

$$\min_{f, t_z'} \sum_{i=1}^{N} \left(\frac{f x_i}{z_i + t_z'} - u_i \right)^2 + \left(\frac{f y_i}{z_i + t_z'} - v_i \right)^2, \tag{2.11}$$

where (x_i, y_i, z_i) is the predicted 3D points with pixel coordinates (u_i, v_i) and $t_z' = t_z / s$. The depth map in camera space is then recovered by adding t_z' to the z-coordinates of the point cloud.

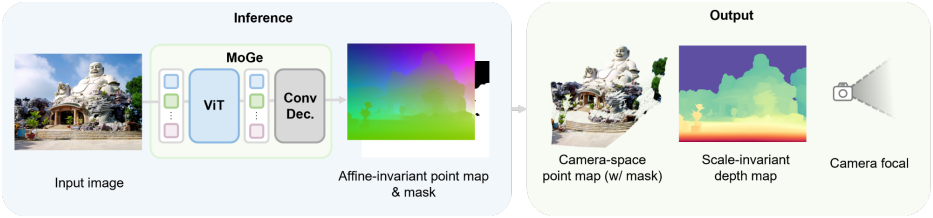


Figure 2.4 Overview of the MoGe pipeline. Image credit: [17]

MADPose Employing affine-invariant transformations for relative pose estimation has previously been explored [19]. The transformation of a depth map $D \in \mathbb{R}^{M \times N}$ is formulated as applying two scalar values α and β:

$$\hat{D} = \alpha D + \beta. \tag{2.12}$$

From a set of 2D keypoint correspondences $(\mathbf{p}_{1j}, \mathbf{p}_{2j})$, the transformed depth values, $\hat{d}_{1j}$ and $\hat{d}_{2j}$, are used to backproject the points

$$\mathbf{P}_{ij} = \mathbf{K}_i^{-1} \begin{bmatrix} \mathbf{p}_{ij} \\ 1 \end{bmatrix} \cdot \hat{d}_{ij}, \quad i \in \{1, 2\}. \tag{2.13}$$

As the two 3D points are related by the rotation and translation given by the relative pose, the transformed depth priors provide additional information that can be utilized when estimating the transformation. The proposed solvers function under the assumption that the pairwise distances δ from the backprojected 3D points should be equal for each view given a set sample

size M of 2D correspondences. For a pair of 3D points (j, k), this means that the distance in view i is defined as

$$\delta_{jk}^{(i)} = \mathbf{P}_{ij} - \mathbf{P}_{ik}, \quad i = \{1, 2\} \quad \text{and} \quad j \neq k. \tag{2.14}$$

Due to the transformation between the two poses being length-preserving, the pairwise distances between all pairs of 3D points are equal in the two views

$$||\delta_{ij}^{(1)}|| = ||\delta_{ij}^{(2)}||. \tag{2.15}$$

Assuming two unknown focal lengths, the resulting system of equation can be solved using four point correspondences to derive the scalings, offsets and focal lengths giving four solutions. Only solutions resulting in positive depth values using real valued scalars are kept, from which they estimate the relative pose using the singular value decomposition method. The solutions are then evaluated iteratively using a Hybrid LO-MSAC framework over truncated pairwise depth induced reprojection errors alongside epipolar Sampson errors. To simplify the computations they compute the relative scaling $a = \alpha_2/\alpha_1$ and estimate the two shifts as $b_1 = \beta_1/\alpha_1$ and $b_2 = \beta_2/\alpha_2$.

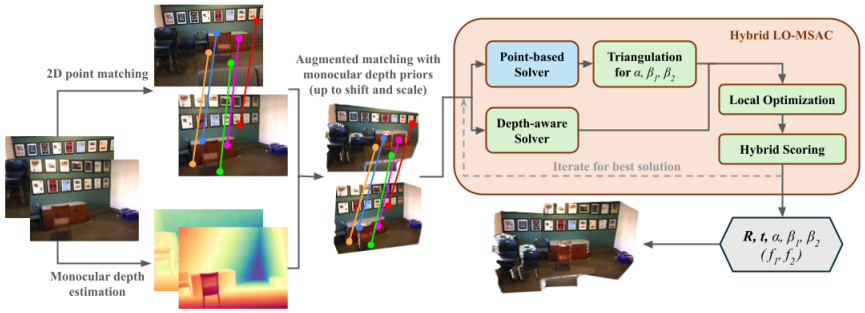


Figure 2.5 The full MADPose pipeline. Image credit: [19]

MP-SfM Incorporating monocular depth priors into an incremental SfM pipeline has previously been studied in the concurrent work "MP-SfM: Monocular Surface Priors for Robust Structure-from-Motion" [15] achieving better pose estimations in low-overlap and low-parallax reconstructions compared to SfM methods relying on three-view overlap for next image registration.

Unlike other methods that rely on a set of unordered images with known camera intrinsics, each image is also associated with an accompanying depth map $D_i \in \mathbb{R}_0^{+H_i \times W_i}$, normal map $N_i \in \mathcal{S}^{2^{H_i \times W_i}}$ and confidence maps $\Sigma_{D_i}, \Sigma_{N_i}$. Either sparse or dense features are extracted for each image, which are then matched and geometrically verified. In the case of image sets with low-overlap, where all possible initial poses are deemed unstable, the method utilizes the depth priors to lift 3D points allowing an initial pose to be found by estimating the absolute pose according to PnP. Inliers with low-parallax are then lifted while the remaining matches are triangulated.

For the depth priors to remain consistent with the point cloud they apply a scaling to the original depth maps

$$D_i^* = D_i \cdot \underset{j,k}{\text{median}} \left(\left\{ \bar{D}_i(\mathbf{X}_k) / D_i(\mathbf{x}_j) \right\} \right), \tag{2.16}$$

where the depth of point $\mathbf{X}_k$ in frame i is denoted $\bar{D}_i(\mathbf{X}_k)$ and $D_i(\mathbf{x}_j)$ is the depth at coordinate $\mathbf{x}_j$. The pose of each new view is estimated using both multi-view triangulated and lifted points. Afterwards, the depths are refined and evaluated through a depth consistency check to determine whether to de-register certain images.

During global and local bundle adjustment they jointly optimize three cost functions

$$\arg \min_{\mathcal{P}, \mathcal{X}, \mathcal{D}^*} C_{BA} + C_{reg} + C_{int}. \tag{2.17}$$

The first term is the standard cost used in bundle adjustment problems, given by the accumulated reprojection error of each 3D point $\pi(\mathbf{K}, \mathbf{P}, \mathbf{X})$ into each view i where it is visible

$$C_{BA} = \sum_{i \in \mathbb{R}} \sum_{j,\, k} \rho_{BA} ||\pi(\mathbf{K}_i, \mathbf{P}_i, \mathbf{X}_k) - \mathbf{x}_j||_{\Sigma_{\mathbf{x}_j}}, \tag{2.18}$$

where the norm is calculated according to Mahalanobis distance and ρ_{BA} is a truncated Smooth-L1 loss. To handle discrepancies between the refined depths and the 3D points they compute the following loss with a robust Cauchy loss ρ_{reg}

$$C_{reg} = \sum_{i \in \mathbb{R}} \sum_{j,\, k} \rho_{reg} \left(||\bar{D}_i(\mathbf{X}_k) - D_i^*(\mathbf{x}_j)|| \right). \tag{2.19}$$

The last cost term focuses on the difference between the refined depth and the estimated surface tangent plane normal $\Delta D_i^* \in \mathcal{S}^{2^{H_i \times W_i}}$ compared to the monocular depth and normal priors

$$C_{int} = \sum_{i \in \mathbb{R}} \sum_{u,\, v} \Big[\rho_{prior} \left(||D_i^*(u,v) - D_i(u,v)||_{\Sigma_{D_i}(u,v)} \right) \tag{2.20}$$

$$+ \rho_{int} \left(||N_i(u,v) - \Delta D_i^*(u,v)||_{\Sigma_{N_i}(u,v)} \right) \Big], \tag{2.21}$$

where (u, v) refer to pixel coordinates and ρ_{prior}, ρ_{int} are truncated L_2 robust losses.

The final step is to possibly de-register images by performing a depth consistency check. Any registered image I_c which has too many inconsistent observations with other existing views will thus be discarded. These observations are determined by reprojecting the refined depth map into overlapping images creating a min-depth buffer $D_{i \leftarrow c}^*$. Similarly, this is done in the opposite direction giving the consistency ratio with confidence γ as

$$\beta_i = \frac{1}{W_i H_i} \sum_{u,\, v} \mathbb{1} \left(\frac{D_i^*(u,v) - D_{i \leftarrow c}^*(u,v)}{\Sigma_{D_i}(u,v) + \Sigma_{D_{i \leftarrow c}}(u,v)} > \gamma \right) \tag{2.22}$$

$$+ \frac{1}{W_c H_c} \sum_{u,v} \mathbb{1} \left(\frac{D_c^*(u,v) - D_{c \leftarrow i}^*}{\Sigma_{D_c}(u,v) + \Sigma_{D_{c \leftarrow i}}(u,v)} > \gamma \right). \tag{2.23}$$

Any registered image with a ratio above a certain threshold is then discarded and the associated structure removed.

3

Methodology

In this chapter, we present each step of our monocular depth based SfM pipeline, seen in fig. 3.1. This includes selecting an initial pair, triangulation of the scene points and how the optimization problem is structured for the bundle adjustment.

As an overview, we briefly summarize our approach. First, we select an initial image pair and estimate the two-view geometry as described in section 3.1. The scene points are found by averaging the lifted keypoints along each track. In section 3.2 it is described how we select the next view and transform the associated depth prior to align with the observed scene points. We utilize a combination of local and global bundle adjustment to reduce accumulated errors. The presented cost function, introduced in section 3.3, is based on reprojection errors and depth map deviations. In section 3.4 we describe our process of rejecting or de-registering views. Lastly, the triangulated points and camera poses are used to generate a more complete reconstruction, which is described in section 3.5.

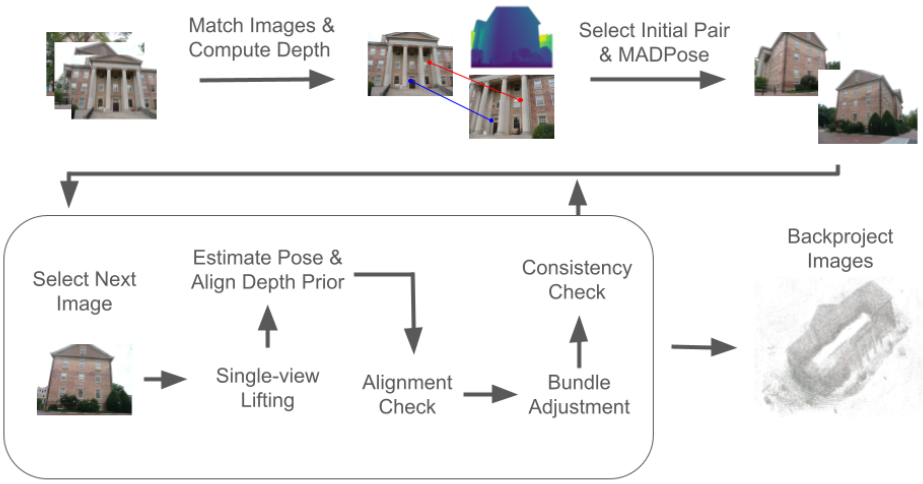


Figure 3.1 Summary of our monocular depth based incremental SfM pipeline.

3.1 Initialization

For a collection of unordered images, we perform feature extraction and matching. We consider two types of 2D-2D correspondences. The first is the default SIFT matches from COLMAP. Furthermore, we consider Superpoint[11] features and LightGlue[13] matches found by the HLOC[2] repository. For the latter, the images are first reshaped to a smaller size to find these correspondences. Lastly, we use the geometric verification pipeline provided by COLMAP to filter out geometrically unfeasible 2D-2D correspondences, which gives us our pairwise inlier matches.

The initial pair is selected as the image pair with the most amount of inlier matches. Each image is also processed through MoGe to get the set of depth maps $\mathcal{D} = \{D_i \in \mathbb{R}^{H_i \times W_i}\}$ associated with the images. The focal lengths from MoGe are used as initial values for the pipeline and we assume a simple pinhole camera model.

We use the two-focal solver presented in MADPose[19] to estimate the relative pose for the initial pair and update the focal length estimates. It also provides the initial scalings $\alpha_1 = 1$, α_2 and offsets β_1, β_2, which we use to transform the depth maps according to eq. (2.12), creating the first entries in the set of transformed depth maps $\hat{\mathcal{D}} = \{\hat{D}_i \in \mathbb{R}^{H_i \times W_i}\}$.

The matches for the initial pair are used to create the set of tracks $\mathcal{T}$, where each scene point $\bar{\mathbf{X}}_i$ is associated with a track $\mathcal{T}_i$. By traversing each track, we are able to locate which images observe the scene point and the corresponding feature locations. The transformed depth maps and tracks are used to find the initial scene points. These are calculated as the mean of the lifted matches using the transformed depths. More specifically, the 3D point for the 2D correspondences along track $\mathcal{T}_i$ is given by

$$\bar{\mathbf{X}}_i = \frac{1}{|\mathcal{T}_i|} \sum_{j \in \mathcal{T}_i} \left[\mathbf{K}_j^{-1} \mathbf{R}_j^T \hat{D}_j(\mathbf{x}_j) \mathbf{x}_j - \mathbf{R}_j^T \mathbf{t}_j \right]. \tag{3.1}$$

3.2 Next View Selection

For next view registration, we consider the view with the most amount of inlier matches with the currently registered images. Similar to COLMAP, the multi-view triangulated points can be used to estimate the pose of the next view. Additionally, the transformed depth priors allows for single-view lifting of 2D-2D correspondences. Unique keypoint matches only observed in the next view and one of the registered views are thus lifted. Using both the new lifted points and the pre-existing scene points we estimate the pose for the next image by solving the PnP problem for the 2D-3D correspondences. This allows us to register new views in low-overlap cases where traditional SfM fails, see fig. 3.2. The solution for the pose is found using the solvers provided by Poselib[20].

Figure 3.2 Single-view lifting allows next-view registration even with no three-view overlap between images. This omission in COLMAP limits it to only be able to register two views and one set of scene points (either red or blue).

To align the depth prior with the observed point cloud, we set up a two point RANSAC solver. As mentioned in the theory section, the camera equations eq. (2.1) states the projection of a scene point $\mathbf{X}$ into image i should equal up to scale with the corresponding pixel coordinate in homogeneous coordinates. Setting $\lambda = \alpha_i D_i(\mathbf{x}) + \beta_i$ lets us re-write this as

$$\begin{bmatrix} \mathbf{R}_i^T D(\mathbf{x}) \mathbf{K}_i^{-1} \mathbf{x} & \mathbf{R}_i^T \mathbf{K}_i^{-1} \mathbf{x} \end{bmatrix} \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} = \mathbf{X} + \mathbf{R}_i^T \mathbf{t}_i. \tag{3.2}$$

As this system is ill-conditioned for a single point, we iteratively sample two 2D-3D correspondences uniformly and compute the least-squares solution. In each iteration, we lift the 2D points and project the resulting 3D points into the registered images. Inliers are defined as the reprojection errors below a given threshold and the solution with the largest amount of inliers is stored.

Simultaneously as the 2D-3D correspondences for the next view are found we construct, or extend, the tracks. While registering a new image there are two possible cases. Firstly we consider the case if a keypoint in the next view has a match which is already associated with a scene point. In such cases we simply extend the pre-existing track. The other case deals with keypoints that have a match in some registered image but have not yet been triangulated. For such cases we utilize the refined depth map of the old image and lift the keypoint creating a new scene point. This creates a brand new track containing two views. Once the depth map of a newly registered image has been transformed it can be used to backproject the image, see fig. 3.3.

(a) Triangulated Keypoints **(b)** Backprojected Images

Figure 3.3 By aligning the depth maps to the triangulated keypoints we are able to recover a more complete reconstruction through backprojection.

3.3 Bundle Adjustment

To solve the bundle adjustment problem we use the following approach. For each track $\mathcal{T}_i$ we uniformly sample a keypoint $\mathbf{x}_i^*$ from one view. We lift the point to get the scene point $\mathbf{X}_i^*$, which we project into the other views along the track. The cost in each view is determined by two components. Firstly, we consider the reprojection error in pixel coordinates. Secondly, we also compute the difference between the observed depth $\bar{D}$ of the scene point and the transformed depth prior $\hat{D}$ for the keypoint match.

Let $\Theta = \{(f_i, \alpha_i, \beta_i)\}$ denote the set containing the focal lengths, scalings and offsets. The total cost is determined over the set of tracks $\mathcal{T}$, camera poses $\mathcal{P}$ and parameters Θ for the registered images. We consider the solution which minimizes the described cost, which we formally define as

$$\arg\min_{\mathcal{P},\Theta} \sum_{i\in\mathcal{T}} \sum_{j\in\mathcal{T}_i} \rho \left|\left| \pi\left(\mathbf{K}_j, \mathbf{P}_j, \mathbf{X}_i^*\right) - \mathbf{x}_j,\ \bar{D}_j(\mathbf{X}_i^*) - \hat{D}_j(\mathbf{x}_j) \right|\right|^2, \tag{3.3}$$

where ρ is a robust Cauchy loss function and π is a function which projects a 3D point into a 2D image.

We solve eq. (3.3) using a sparse Schur solver in Ceres[21]. Lastly, we need to fix seven degrees of freedom to preserve the scale. In GBA, the rotation and translation of the first image in the initial pair are kept constant, which fixes six out of seven degrees of freedom. The last one is chosen as the element of the translation vector with the largest absolute value for the second image.

Similar to other incremental SfM pipelines, we utilize local bundle adjustment as global bundle adjustment becomes costly and inefficient if used too frequently. For a newly registered

image we consider the set of most-connected views, defined as the N registered images with the most pairwise matches with the new view. All tracks associated with the scene points observed in these images are then considered in the LBA. The parameters in all images not in the most-connected views are kept fixed to preserve the scale and prevent drifting from the rest of the scene. If no tracks extend beyond the most-connected views, then the parameters of two views are fixed as described for the GBA.

After bundle adjustment, we use the found solution to update the geometry. In global bundle adjustment, this means that all scenes points and poses are updated. Local bundle adjustment instead only alters the poses of the most-connected images and their observed scene points. Updating these scene points is not limited to the most-connected views, but instead the entire track is considered.

The settings triggering either global and local bundle adjustment can be set depending on the scene. Similar to COLMAP and other incremental SfM pipelines we perform global bundle adjustment when necessary. Once a new image has been registered, we determine whether sufficiently many images has been registered, or the number of scene points has grown by a certain percentage, since the last GBA. If this is the case we perform GBA updating the geometry. LBA can instead be triggered if the scene grows at a certain rate between two image registrations.

3.4 De-Registering Images

Thus far, we have recklessly assumed that we are able to add new views as long as there are matches with any previously registered image. Furthermore, any registered view can not be de-registered. The model then becomes susceptible to cases where there are faulty matches or if we are unable to properly align a depth map. We perform two tests to determine if an image should be de-registered.

Firstly we consider the most recently registered image. The RANSAC solver for the PnP problem yields a set of inliers for the found solution of the estimated pose. Furthermore, aligning the depth map with the observed scene points in the new image provides an additional set of inliers. Consider the average inlier ratio from these two sets. Any newly registered image where this value is less than some pre-defined threshold is considered unreliable. By adding such an image we risk accumulating errors which affect the entire reconstruction. Images that are deemed unreliable are therefore not added to the reconstruction and newly lifted 3D points from matches in other images are not added to the point cloud.

Additionally, we also perform a global test to see if any registered images may be inconsistent with the rest of the reconstruction. For every image we extract the set of observed points and their corresponding tracks. The latter gives us a set of images that share at least one observation with the image in question. We consider the reprojection error found by projecting the lifted 2D correspondences in overlapping images using the transformed depth maps. Similar to the previously described case, the inconsistent images are determined by the inlier ratio. Images below the inlier threshold are de-registered and removed from the tracks. If the track previously only contained two views then the 3D point is discarded.

3.5 Final Reconstruction

Once the pipeline has finished, the estimated triangulated points and camera parameters are used to produce the final reconstruction. As previously mentioned, this is done through back-projection of the transformed depth priors. Only pixels contained in the non-infinity region $\mathcal{M}$ from MoGe are backprojected. Furthermore, any pixel with negative depth is filtered out, yielding the final reconstruction.

4

Results

In the following chapter we present the results of our model and compare it to COLMAP. We examine scenarios regarding the reconstruction quality with ground truth camera parameters as well as pose estimation. Additionally we perform ablation studies motivating the components of the model and their effect on the runtime.

4.1 Reconstruction Results

As our first test we look at the accuracy and completeness of the 3D reconstructions produced by our method compared to COLMAP. For this we use the ETH3D Stereo benchmark[22] with the associated evaluation code. It uses the ground truth point cloud for each dataset to evaluate the quality of a reconstruction. The measurements are evaluated based on distance thresholds which can be selected within the range 1cm to 50cm.

First, the distance from each 3D ground truth point to the closest reconstruction point is estimated [22]. Using this, the completeness is defined as the ratio of ground truth points where this distance is below the evaluation threshold. Additionally, the accuracy is defined as the ratio of reconstruction points with a distance below the evaluation threshold to the closest ground truth point. Lastly, we also report the F_1 score of the reconstruction, which is determined as the harmonic mean of the accuracy and completeness

$$F_1 = 2 \cdot \frac{accuracy \cdot completeness}{accuracy + completeness}. \tag{4.1}$$

We consider evaluation thresholds of 1cm, 2cm and 5cm. For feature extraction and matching we use either SIFT or Superpoint[12] with LightGlue[13] (SP+LG). The ground truth camera parameters are used and are kept fixed for each method. Methods which utilize the same feature extraction and matching method uses the same set of matches. The results can be seen in table 4.1.

Table 4.1 Multi-View Stereo evaluation showcasing completeness, accuracy and F_1 score over different evaluation thresholds. The results are reported as percentages and are taken as the average over all datasets presented in section A.1.

Matches	Method	Evaluation Threshold		
		1cm	2cm	5cm
SIFT	COLMAP	0.97/**86.75**/1.91	2.80/**92.85**/5.38	8.52/**97.26**/15.47
SIFT	ours	**28.44**/22.68/**24.54**	**53.97**/37.05/**43.15**	**78.82**/61.00/**68.11**
SP+LG	COLMAP	0.63/67.03/1.24	2.52/78.81/4.81	10.61/89.66/18.67
SP+LG	ours	28.19/23.48/24.18	49.95/38.26/41.80	71.92/63.36/65.19

4.2 Pose Estimation

The accuracy of the estimated poses for the full SfM scene is evaluated using the ETH3D ground truth poses. For each registered view, we compute the angular error with the corresponding

ground truth rotation and report the metric distance between the estimated camera centers. The angular error ϕ in radians for the rotation is defined as

$$\phi = \left| \arccos \left(\frac{\mathrm{tr}\left(R_{est}^T R_{GT}\right) - 1}{2} \right) \right|, \tag{4.2}$$

where R_{est} is the estimated pose and R_{GT} the ground truth. We report the angular error as the Area Under the recall Curve (AUC) up to 1/5/20°. Furthermore, the ratio of registered images is presented. The compiled result for both our method and COLMAP can be seen in table 4.2.

Table 4.2 Summary of the pose estimation for the ETH3D dataset. The AUC is reported as percentages. The complete results for each dataset can be seen in section A.2. Note, the median camera center distance is found from the average distance for each image set.

Matches	Method	AUC 1/5/20°	Median Camera Center Distance (m)	Images Registered (%)
SIFT	COLMAP	35.39/49.17/60.01	0.60	76.83
SIFT	ours	32.86/70.77/90.76	0.1	**96.50**
SP+LG	COLMAP	**59.54**/78.39/85.01	**0.02**	92.08
SP+LG	ours	38.85/**78.53**/**94.22**	0.08	84.17

4.2.1 Focal Length Estimation

In addition to estimating the camera poses, we have assumed the focal length of each camera to be unknown. With this in mind, we report the median of the average absolute focal length difference for COLMAP and our method compared to the ground truth. Additionally, we report the initial deviation based on the estimates from MoGe, see fig. 4.1.

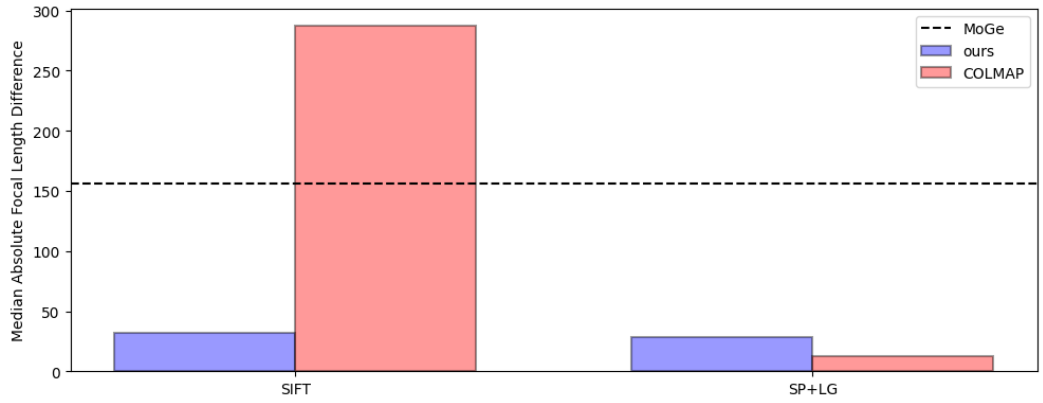


Figure 4.1 Median absolute difference of the predicted focal length and ground truth for the ETH3D dataset. The complete results can be seen in section A.2. Note, the median is computed from the average absolute difference for each image set.

4.3 Low-overlap Pose Estimation

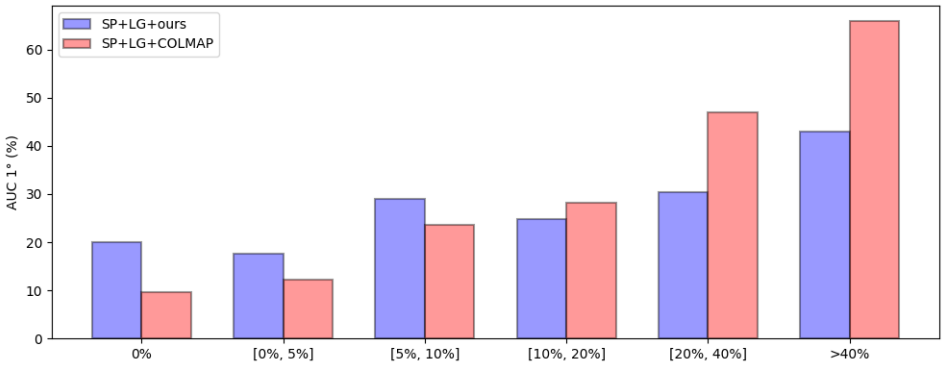
As mentioned previously, traditional SfM pipelines often break down or are unable to continue in low-overlap cases. Since we are able to lift points from a single view without the need for multi-view triangulation our method should be less susceptible to such cases. To see if this is the case we replicate the low-overlap triplet test presented in MP-SfM [15]. From a collection of SfM datasets, consider collections of image triplets. The triplets have been sampled according to a pre-defined maximum collective visual overlap. For image collections with ground truth depth maps, the overlap is defined as the ratio of covisible pixels. If no ground-truth depth is available it is defined as the ratio of covisible feature points from the original full SfM model.

We consider six intervals of increasing overlap and determine the pose accuracy compared

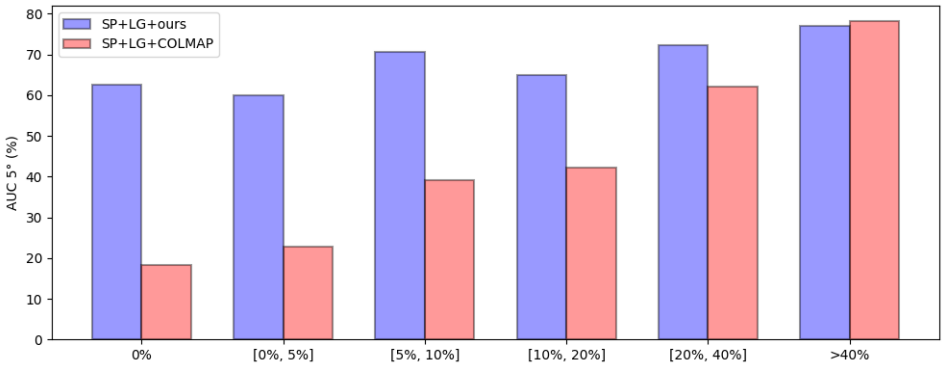
to the ground truth. For each registered image pair we compute the angular error in rotation and translation of the relative pose. The pose error is defined as the maximum of the two. Let t_{est} and t_{GT} be the estimated and ground truth translations. The angular error θ in radians is then defined as

$$\theta = \arccos\left(\frac{t_{est} \cdot t_{GT}}{||t_{est}|| \cdot ||t_{GT}||}\right). \tag{4.3}$$

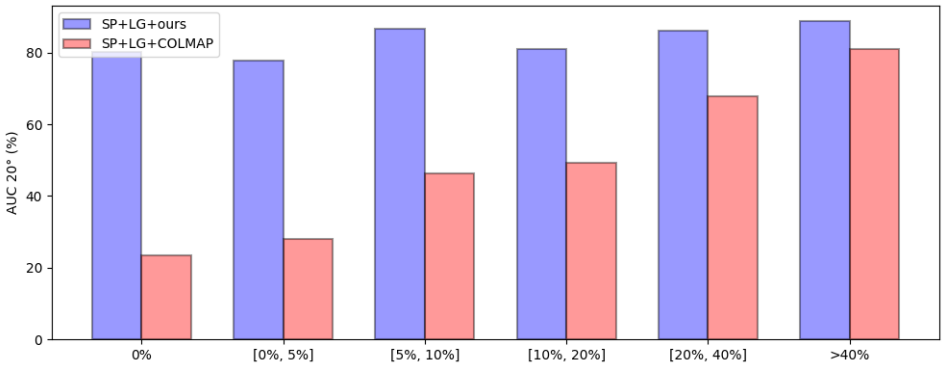
Any image pair not registered is given the maximum pose error of 180°. We report the results as the AUC up to 1/5/20°, see fig. 4.2.



(a) AUC@1°



(b) AUC@5°



(c) AUC@20°

Figure 4.2 Reported AUC up to 1/5/20° for increasing three-view overlap over image triplets.

4.4 Ablation Studies

In this section we present ablation studies for the proposed method. Firstly we consider the effect of introduced components on the overall result. Additionally, we also study the effect of different parts of the pipeline on the overall runtime. The results were computed from one indoor and outdoor scene. These were chosen as *Delivery Area* and *Electro* by uniformly sampling image sets from the ETH3D dataset.

4.4.1 Component Analysis

We perform an ablation study to motivate the inclusion of each component in our model. In total we study the inclusion of five components. Firstly, we both study how the model behaves when only allowed to use the triangulated keypoints, and when the inclusion of single-view lifting is removed. This makes the pipeline more closely resemble COLMAP in regards to the size of the final reconstruction and which scene points we are able to use when registering a new view.

Secondly, we try to motivate the inclusion of the offset parameters in the transformed depth map. Comparing our approach to the monodepth-SfM pipeline presented in MP-SfM[15] we note that the offset is omitted. The assumption that depth maps from different views can be aligned by simple scaling factors is a common assumption [19]. However, as seen in the two-view case presented in MADPose[19] this is not necessarily the best approach. Therefore we try to quantify the inclusion of the offset parameter by removing it entirely from the pipeline.

As the proposed method for computing the loss diverges from convention, we set out to study the effect of this choice on the model. We compare the runtime and results to a cost determined by the reprojection errors:

$$\arg\max_{\mathcal{P},\Theta} \sum_{i\in\mathcal{T}} \sum_{j\in\mathcal{T}_i} \rho \left|\left|\pi\left(\mathbf{K}_j, \mathbf{P}_j, \bar{\mathbf{X}}_i\right) - \mathbf{x}_j\right|\right|^2, \tag{4.4}$$

where ρ is a robust Cauchy loss function and π is a function which projects the 3D point into a 2D image. How this affects the overall runtime of the model can be seen in fig. 4.3.

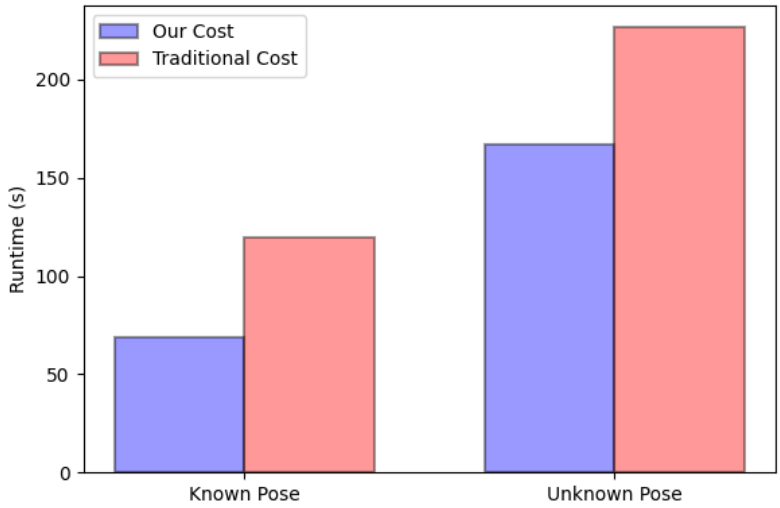


Figure 4.3 Runtime of the model with the proposed cost versus a traditional one. The presented time in seconds is taken as the average over the Delivery Area and Electro image sets and only consider the reconstruction part of the pipeline.

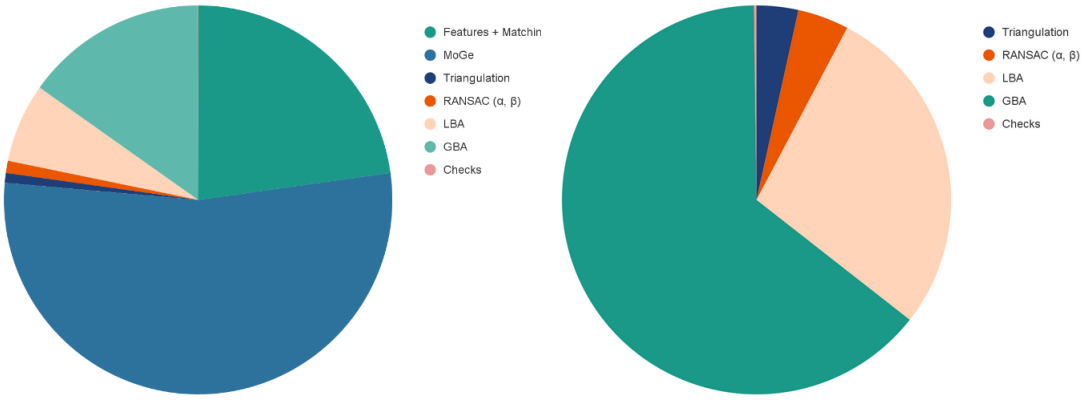
Lastly, we also consider the inclusion of the robust (α, β) estimator to align the original depth map with the observed scene points. Instead, we perform a least-squares fit for all points to align the depth prior. The results can be seen in table 4.3.

Table 4.3 Ablation of our pipeline. We perform Multi-View Stereo evaluation showcasing completeness, accuracy and F_1 score over different evaluation thresholds. The results are reported in percentages and the ability to register images was disabled for this test.

Variant	Evaluation Threshold		
	1cm	2cm	5cm
SIFT+ours	18.05/17.31/**17.75**	41.72/29.49/**34.47**	**72.88**/51.70/**60.38**
keypoints only	0.22/**25.59**/0.43	1.01/**37.47**/1.97	5.81/**56.63**/10.54
no lifting	16.71/15.73/16.20	40.59/26.79/32.27	68.60/48.91/57.09
no offset	16.11/13.66/14.73	36.65/23.82/28.85	62.97/43.57/51.51
no (α, β) RANSAC	**18.13**/17.04/17.57	**41.80**/28.57/33.94	69.12/50.33/58.22
traditional cost	17.35/15.80/16.54	39.27/27.35/32.24	69.33/49.96/58.08

4.4.2 Runtime Analysis

As mentioned previously, bundle adjustment often gets costly which increases the overall runtime substantially. To see if this is also true for this method we study how each part of the pipeline affects the overall runtime, see fig. 4.4. Firstly, we study the feature extraction+matching and MDE. The former are based on SIFT matches found by the default COLMAP pipeline. As these two parts of the pipeline are external, we also present the proportions with these omitted. The remaining parts we consider are the triangulation, (α, β) RANSAC, pose estimation, bundle adjustments and internal checks. Note, both the alignment and global check have been grouped together, and the triangulation includes next image registration, track creation and MADPose. We did not achieve any improvement in the overall runtime compared to COLMAP. This was the case both with and without the MDE included.



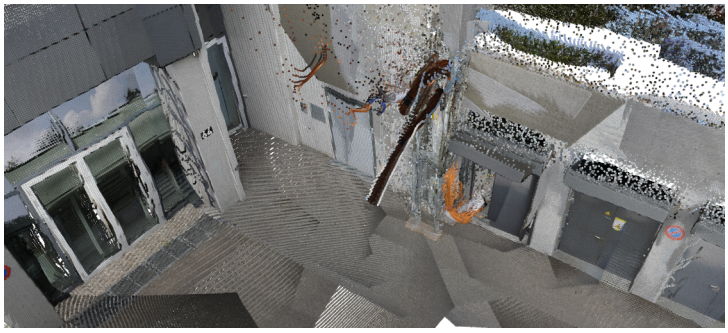
(a) Full pipeline **(b)** Matching and MDE omitted

Figure 4.4 Visualization of the runtime distribution of the model. The percentages are derived from the sum of each component for the Delivery Area and Electro image sets.

4.5 Reconstruction Visualizations

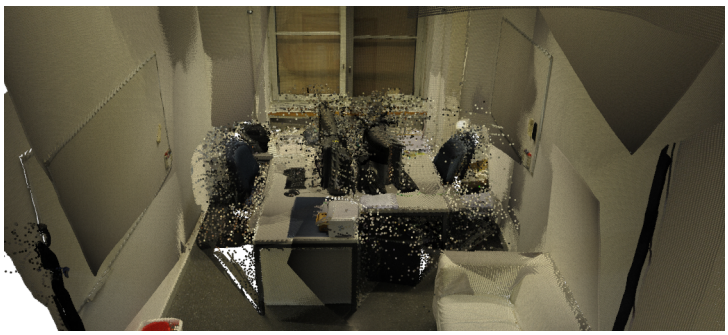
In this section be present visualizations of the final reconstructions for a subset of the ETH3D image sets. We showcase the following image sets: Delivery Area, Electro, Kicker and Office. For the sake of visual clarity we only observe parts of the reconstructions, these can be seen in fig. 4.5.

(a) Delivery Area



(b) Electro

(c) Kicker



(d) Office

Figure 4.5 Examples from final reconstructions for the ETH3D dataset.

5

Discussion

5.1 Accuracy and Robustness

Firstly, we note that our method overall performs worse compared to COLMAP in terms of accuracy. In the of case of the multi-view stereo test seen in table 4.1 this is true across all three evaluation thresholds and both matching methods. This behavior is expected as only triangulated keypoints are evaluated while our method backprojects full images. As a result we are, to a certain extent, limited by the accuracy of the depth priors provided by MoGe. Even considering perfect alignment to the scene points, the model will not be able to accurately predict other areas if the depth priors are faulty to begin with. However, note that our method also performs worse for the pose estimation at AUC 1°, seen in table 4.2. As the pose estimation only depends on the keypoints, this still suggest there is a loss in accuracy compared to COLMAP. Note that we manage to improve this accuracy through the use of SP+LG matches, but it is still not enough to match COLMAP. However, for SP+LG there is a decrease in the percentage of registered images, which may explain the increased accuracy. Triangulating scene points using the DLT method, as done by COLMAP, takes into consideration noise from all views. Our triangulation method of averaging the lifted keypoints does not do this, and can thus accumulate errors which impact the overall accuracy.

On the other hand, we note that for 5/20° our method obtains better results using the same matches. Higher AUC for larger rotation errors suggests our method is more robust at the cost of accuracy. While COLMAP achieves the most accurate camera center placements overall, we note that it is more prone to failure as seen in section A.2.

This tendency is also prevalent in the focal length estimates, especially when it comes to SIFT matches, as seen in fig. 4.1. However, similar to the pose estimates, ignoring the failure cases, we see that COLMAP outperforms our method in terms of accuracy. On the other hand, comparing the final focal lengths to the monocular estimates we note a clear improvement. The additional views provide more clear geometric cues not available in a single image, which would explain this improvement.

One of the core ideas behind single-view lifting is that we should be able to better deal with failure cases for traditional SfM pipelines. This includes low or no three-view overlap, as seen in fig. 3.2. For such cases we see a significant improvement with our method, as seen in fig. 4.2. When it comes to the lowest possible overlap, this change is the most noticeable as COLMAP is only able to register two views, meaning two out of three relative poses have the maximum angular error. As the overlap increases, the performance of COLMAP significantly improves. While COLMAP should be able to register all three views for the lower to middle intervals we still see that our method outperforms it. The ability to leverage single-view lifting is used generate more scene points. This gives us additional information, not accessible to COLMAP, when estimating the pose. With this in mind, it can explain why we still observe an improvement when the overlap grows, which diminishes as the amount of observed points increases. For intervals with larger overlap we note that our method performs worse than COLMAP at times. This which is likely due to accuracy issues described previously, which is the most noticeable for AUC 1°.

5.2 Completeness and Reconstruction Quality

By backprojecting full images using the transformed depth maps, we naturally generate a larger reconstruction compared to only triangulating keypoints. Even with lower accuracy, we should expect a greater completeness compared to COLMAP. As seen in table 4.1 this is true for our method across all evaluation thresholds and matches. Furthermore, despite the loss in accuracy we see that our method still obtains better F_1 scores. This would signify that the increased completeness is able to compensate for this, achieving better reconstruction results overall. The full multi-view stereo results for each image set can be seen in section A.1, where we note that this behavior is consistent.

5.3 Ablation Studies

5.3.1 Component Analysis

By only evaluating the reconstruction based on the triangulated scene points we observe an increased accuracy and decreased completeness. Naturally, this behavior is to be expected as the point cloud is smaller, and bundle adjustment optimizes the residuals along the tracks. Only considering keypoints therefore removes one of the main benefits of this methods. The accuracy being larger than the completeness is more akin to the results for COLMAP. However, comparing it to the average accuracy over the datasets seen in section A.1, we note that it is lower compared to COLMAP.

Single-view lifting is another important aspect of the model. It allows for more scene points for the pose estimation and bundle adjustment. Removing it from the pipeline results in a drop in the overall accuracy. Additionally, the completeness is also affected as the model does not have the ability to register as many images in the case of no to low overlap. The results in table 4.3 supports that this inclusion is useful for the model.

Omitting the offset parameters results in a noticeable reduction in accuracy and completeness for our method. As mentioned previously, the inclusion of this parameter has seen to have a noticeable improvement in relative pose estimation, as reported by MADPose. This coincides with our results that this inclusion improves the depth alignment. Therefore aligning depth priors through simple scaling, such as in MP-SfM, may instead benefit through the use of affine transformations.

Least squares fitting of the initial (α, β) parameters to all observed scene points does not seem to impact the model significantly. The overall reconstruction quality is similar, which is also true for the accuracy and completeness. Further tests would have to be conducted to see how robust estimation of the parameters affects convergence speed in the bundle adjustment. Flawed initial estimates could also lead to de-registration of new images. But as this feature was disabled in the test, no conclusions can be drawn.

As seen in fig. 4.3, using a more traditional cost function, where we compute the reprojection error of the averaged lifted points, increased the overall runtime of the model. Unlike the GBA problem in eq. (2.4), each residual now depends on camera parameters from more than one image. By computing the scene points as the mean of the backprojections we thus reduce the sparsity of the problem. Through choosing one view at random, we reintroduce the sparsity into the optimization problem, which could motivate this improved runtime. Additionally, we observe improved results for our proposed cost function which further favors this decision.

5.3.2 Runtime Analysis

In fig. 4.4, we note the MDE takes up more than half of the runtime. This is troublesome as the remaining pipeline depends on the depth priors. While MoGe is optimized for speed, it is still an external program with a determined runtime that we can not alter. Switching to another model may improve this. However, the results are heavily dependent on the accuracy of the monocular depth priors and focal lengths. As MoGe has shown to achieve better results for these compared to past methods, this tradeoff seems reasonable.

Ignoring the MDE and feature matching, the bundle adjustment occupies most of the runtime, similar to past SfM pipelines. Through LBA, the more computationally expensive GBA is performed only in necessary cases. However, the LBA still takes up a substantial part of the runtime. As the model was not able to achieve a quicker runtime compared to COLMAP, this suggest that the issue resides in the bundle adjustment. Intuitively, optimizing two parameters for each depth prior instead of each scene point should speed up the optimization problem. To identify why this was not the case would require further analysis.

6

Conclusion and Future Work

6.1 Conclusion

Aligning with the goals presented in this thesis, we have presented an incremental SfM pipeline utilizing affine transformations of monocular depth priors. Furthermore, it successfully produces dense reconstructions from sparse 2D-2D matches and overcomes low three-view overlap failure cases in traditional SfM methods. With this in mind, we managed to achieve the goals presented in the problem statement.

In terms of completeness and F_1 score, our method outperforms COLMAP while also achieving better pose estimation for larger AUCs. The main drawback is a substantial reduction in accuracy. Lastly, our method heavily outperforms COLMAP in challenging low-overlap scenes.

6.2 Future Work

Intuitively, optimizing the (α, β) parameters for each image should speed up the process of bundle adjustment compared to optimizing each scene point as in traditional SfM pipelines. However, we did not achieve any improvements in the overall runtime. If this is due to the formulation of the cost function or some other factor is hard to distinguish based on this analysis, but could perhaps motivate further examination.

At the moment, the choice of hyperparameters are quite simple and loosely motivated. The model would benefit from a deeper study on how to effectively choose these based on factors such as image dimensions, overlap and difficulty.

The choice of having (α, β) be shared across the whole image is a restriction made for this thesis. Leveraging different transformation parameters across the images might be another approach worthy of exploration.

Bibliography

[1] J. L. Schönberger and J.-M. Frahm. "Structure-from-Motion Revisited". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[2] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk. "From coarse to fine: robust hierarchical localization at large scale". In: *CVPR*. 2019.

[3] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. *Netvlad: cnn architecture for weakly supervised place recognition*. 2016. arXiv: 1511.07247 [cs.CV]. URL: https://arxiv.org/abs/1511.07247.

[4] R. Szeliski. *Computer Vision Algorithms and Applicatios First Edition*. Spring, 2011.

[5] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg. *Roma: robust dense feature matching*. 2023. arXiv: 2305.15404 [cs.CV]. URL: https://arxiv.org/abs/2305.15404.

[6] A. Bhoi. *Monocular depth estimation: a survey*. 2019. arXiv: 1901.09402 [cs.CV]. URL: https://arxiv.org/abs/1901.09402.

[7] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision Second Edition*. Cambridge University Press, 2004.

[8] M. A. Fischler and R. C. Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: M. A. Fischler et al. (Eds.). *Readings in Computer Vision*. Morgan Kaufmann, San Francisco (CA), 1987, pp. 726–740. ISBN: 978-0-08-051581-6. DOI: https://doi.org/10.1016/B978-0-08-051581-6.50070-2. URL: https://www.sciencedirect.com/science/article/pii/B9780080515816500702.

[9] Ding, Yaqing and Yang, Jian and Larsson, Viktor and Olsson, Carl and Åström, Kalle. "Revisiting the P3P Problem". eng. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2023, 4872–4880. ISBN: 9798350301298. DOI: {10.1109/CVPR52729.2023.00472}. URL: %7Bhttp://dx.doi.org/10.1109/CVPR52729.2023.00472%7D.

[10] D. Lowe. "Object recognition from local scale-invariant features". In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Vol. 2. 1999, 1150–1157 vol.2. DOI: 10.1109/ICCV.1999.790410.

[11] D. DeTone, T. Malisiewicz, and A. Rabinovich. *Superpoint: self-supervised interest point detection and description*. 2018. arXiv: 1712.07629 [cs.CV]. URL: https://arxiv.org/abs/1712.07629.

[12] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. "SuperGlue: learning feature matching with graph neural networks". In: *CVPR*. 2020.

[13] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys. "LightGlue: Local Feature Matching at Light Speed". In: *ICCV*. 2023.

[14] L. Pan, D. Baráth, M. Pollefeys, and J. L. Schönberger. *Global structure-from-motion revisited*. 2024. arXiv: 2407.20219 [cs.CV]. URL: https://arxiv.org/abs/2407.20219.

[15] Z. Pataki, P.-E. Sarlin, J. L. Schönberger, and M. Pollefeys. *Mp-sfm: monocular surface priors for robust structure-from-motion*. 2025. arXiv: 2504.20040 [cs.CV]. URL: https://arxiv.org/abs/2504.20040.

[16] F. Khan, S. Salahuddin, and H. Javidnia. "Deep learning-based monocular depth estimation methods—a state-of-the-art review". *Sensors* **20**:8 (2020). ISSN: 1424-8220. DOI: 10.3390/s20082272. URL: https://www.mdpi.com/1424-8220/20/8/2272.

[17] R. Wang, S. Xu, C. Dai, J. Xiang, Y. Deng, X. Tong, and J. Yang. *Moge: unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision*. 2024. arXiv: 2410.19115 [cs.CV]. URL: https://arxiv.org/abs/2410.19115.

[18] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. *Dust3r: geometric 3d vision made easy*. 2024. arXiv: 2312.14132 [cs.CV]. URL: https://arxiv.org/abs/2312.14132.

[19] Y. Yu, S. Liu, R. Pautrat, M. Pollefeys, and V. Larsson. *Relative pose estimation through affine corrections of monocular depth priors*. 2025. arXiv: 2501.05446 [cs.CV]. URL: https://arxiv.org/abs/2501.05446.

[20] V. Larsson and contributors. *PoseLib - Minimal Solvers for Camera Pose Estimation*. 2020. URL: https://github.com/vlarsson/PoseLib.

[21] S. Agarwal, K. Mierle, and T. C. S. Team. *Ceres Solver*. Version 2.2. 2023. URL: https://github.com/ceres-solver/ceres-solver.

[22] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. "A multi-view stereo benchmark with high-resolution images and multi-camera videos". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

A

Appendix

The complete results for each test presented in the results chapter can be found below. All image sets are retrieved from the ETH3D high-resolution multi-view stereo benchmark. During development, the outdoor and indoor scenes *Courtyard* are *Pipes* were studied. Note, these are included in the final results.

A.1 Complete Reconstruction Results

This sections contains the full results for all image sets used to calculate the result presented in table 4.1. The results are presented as percentages with the ordering completeness, accuracy and F_1 score.

Table A.1 Complete reconstruction result over each evaluation threshold and image set used.

	Evaluation Threshold		
	1cm	2cm	5cm
SIFT+COLMAP			
Courtyard (Outdoor)	0.27/74.92/0.53	1.50/88.26/2.95	8.98/95.62/16.42
Electro (Outdoor)	0.28/84.37/0.56	1.07/92.82/2.11	4.13/96.99/7.92
Terrace (Outdoor)	0.45/87.40/0.90	1.88/92.98/3.68	9.03/96.95/16.52
Meadow (Outdoor)	0.02/69.13/0.05	0.12/80.11/0.25	0.80/92.68/1.59
Playground (Outdoor)	0.58/78.30/1.14	2.47/87.19/4.80	10.94/95.62/19.64
Delivery Area (Indoor)	0.55/82.86/1.09	1.98/90.46/3.88	7.74/96.26/14.33
Kicker (Indoor)	1.30/88.41/2.55	3.51/94.14/6.76	8.72/97.40/16.00
Office (Indoor)	2.17/95.52/4.25	5.62/97.91/10.63	12.20/99.59/21.74
Pipes (Indoor)	0.59/96.26/1.17	1.53/97.85/3.01	4.83/98.99/9.22
Relief (Indoor)	1.85/93.74/3.63	4.77/97.32/9.09	10.31/99.28/18.69
Relief 2 (Indoor)	1.52/94.57/3.00	4.16/97.41/7.98	10.91/98.91/19.65
Terrains (Indoor)	2.04/95.48/4.00	4.95/97.77/9.42	13.64/98.77/23.97
SIFT+ours			
Courtyard (Outdoor)	12.70/19.11/15.26	42.83/33.92/37.86	87.02/58.86/70.22
Electro (Outdoor)	17.93/17.83/17.88	39.12/29.25/33.48	66.15/51.01/57.60
Terrace (Outdoor)	24.45/25.90/25.15	50.69/41.95/45.91	79.38/66.38/72.30
Meadow (Outdoor)	13.79/18.92/15.95	39.94/32.54/35.86	67.99/60.76/64.17
Playground (Outdoor)	15.95/10.02/12.31	34.39/16.87/22.64	57.52/32.04/41.16
Delivery Area (Indoor)	20.07/17.91/18.93	47.62/30.06/36.86	80.12/52.32/63.31
Kicker (Indoor)	44.98/22.10/29.63	77.73/37.08/50.21	96.67/62.37/75.82
Office (Indoor)	56.23/28.02/37.41	77.46/44.13/56.23	95.34/66.59/78.41
Pipes (Indoor)	41.27/30.79/35.27	69.60/48.94/57.47	85.39/73.63/79.07
Relief (Indoor)	29.04/32.92/30.86	47.57/51.09/49.27	70.21/77.51/73.68
Relief 2 (Indoor)	20.06/29.92/29.49	49.75/47.31/48.50	69.78/73.16/71.43
Terrains (Indoor)	44.81/18.67/26.35	70.83/31.41/43.52	90.30/57.33/70.13

	Evaluation Threshold		
	1cm	2cm	5cm
SP+LG+COLMAP			
Courtyard (Outdoor)	0.20/46.43/0.40	1.20/63.77/2.35	8.02/81.48/14.60
Electro (Outdoor)	0.11/64.56/0.22	0.60/77.96/1.19	4.14/88.71/7.91
Terrace (Outdoor)	0.18/63.64/0.36	0.95/77.05/1.87	6.47/89.36/12.07
Meadow (Outdoor)	0.06/35.22/0.12	0.42/52.93/0.84	4.21/76.90/7.99
Playground (Outdoor)	0.41/59.20/0.81	2.04/71.40/3.96	10.87/85.61/19.29
Delivery Area (Indoor)	0.24/61.55/0.47	1.24/75.47/2.44	7.75/87.76/14.23
Kicker (Indoor)	0.64/67.44/1.27	2.94/79.61/5.67	12.70/89.37/22.24
Office (Indoor)	0.53/74.78/1.05	2.27/84.98/4.42	9.74/94.15/17.66
Pipes (Indoor)	0.85/82.57/1.68	3.58/90.10/6.89	14.17/94.48/24.64
Relief (Indoor)	0.88/82.83/1.75	3.95/91.34/7.57	14.00/96.48/24.45
Relief 2 (Indoor)	0.64/78.82/1.27	3.04/88.36/5.88	12.65/95.23/22.33
Terrains (Indoor)	2.83/87.28/5.49	7.98/92.79/14.69	22.61/96.34/36.62
SP+LG+ours			
Courtyard (Outdoor)	12.62/21.00/15.76	43.31/36.28/39.48	88.32/61.61/72.59
Electro (Outdoor)	19.71/19.69/19.70	43.28/32.76/37.29	70.94/57.63/63.60
Terrace (Outdoor)	21.94/19.04/20.38	48.20/31.02/37.75	77.72/51.03/61.61
Meadow (Outdoor)	15.22/21.77/17.92	39.17/38.87/39.02	65.97/70.07/67.96
Playground (Outdoor)	7.83/9.85/8.72	19.21/16.91/17.99	37.32/35.19/36.22
Delivery Area (Indoor)	17.59/19.02/18.28	41.26/32.66/36.46	71.88/57.00/63.58
Kicker (Indoor)	54.32/26.80/35.89	83.58/44.16/57.79	98.31/71.33/82.67
Office (Indoor)	62.49/32.66/42.90	85.22/51.62/64.29	98.47/77.08/86.47
Pipes (Indoor)	41.61/34.47/37.70	65.62/52.47/58.31	82.78/78.56/80.61
Relief (Indoor)	8.37/25.86/12.64	12.08/41.83/18.75	15.67/70.17/25.62
Relief 2 (Indoor)	28.95/27.18/28.04	46.32/41.63/43.85	66.99/65.64/66.31
Terrains (Indoor)	47.60/24.43/32.29	72.19/38.93/50.58	88.70/65.02/75.04

A.2 Complete Pose Estimation Results

The complete table for the result presented in table 4.2 is shown in this section. Additionally, this section contains the complete results for the focal length test.

Table A.2 Complete camera pose results for each image set used.

	AUC 1/5/20°	Mean Camera Center Distance (m)	Images Registered (%)
SIFT+COLMAP			
Courtyard (Outdoor)	84.12/96.82/99.21	0.05	97
Electro (Outdoor)	81.90/96.38/99.10	0.04	89
Terrace (Outdoor)	65.78/89.68/96.77	0.57	100
Meadow (Outdoor)	0.00/0.00/37.70	35.54	33
Playground (Outdoor)	0.00/0.00/0.00	0.32	11
Delivery Area (Indoor)	62.42/90.67/97.32	2.16	100
Kicker (Indoor)	59.44/91.89/97.97	0.04	94
Office (Indoor)	0.00/0.00/0.00	1.50	27
Pipes (Indoor)	69.36/93.87/98.47	0.01	79
Relief (Indoor)	1.67/71.27/91.49	0.63	100
Relief 2 (Indoor)	0.00/0.00/0.00	1.90	97
Terrains (Indoor)	0.00/0.00/2.08	1.43	95

	AUC 1/5/20°	Mean Camera Center Distance (m)	Images Registered (%)
SIFT+ours			
Courtyard (Outdoor)	83.11/96.62/99.16	0.07	100
Electro (Outdoor)	14.78/82.93/95.73	0.10	82
Terrace (Outdoor)	35.78/86.80/96.70	0.09	100
Meadow (Outdoor)	49.20/89.84/97.46	0.22	100
Playground (Outdoor)	0.00/0.00/57.36	0.70	84
Delivery Area (Indoor)	43.27/86.84/96.71	0.10	98
Kicker (Indoor)	47.63/84.23/95.72	0.08	100
Office (Indoor)	0.00/57.88/89.47	0.07	100
Pipes (Indoor)	51.61/90.32/97.58	0.02	100
Relief (Indoor)	36.02/87.22/96.80	0.20	97
Relief 2 (Indoor)	32.92/86.16/96.54	0.20	97
Terrains (Indoor)	0.00/1.20/70.04	0.46	100
SP+LG+COLMAP			
Courtyard (Outdoor)	0.00/0.00/34.99	5.71	100
Electro (Outdoor)	91.67/98.33/99.58	0.02	93
Terrace (Outdoor)	73.97/94.79/98.70	0.03	100
Meadow (Outdoor)	55.11/90.81/97.70	4.85	100
Playground (Outdoor)	67.43/93.49/98.37	0.02	87
Delivery Area (Indoor)	40.91/86.15/96.54	0.12	91
Kicker (Indoor)	78.14/95.63/98.91	0.02	100
Office (Indoor)	73.65/94.73/98.68	0.01	92
Pipes (Indoor)	82.12/96.42/99.11	0.01	100
Relief (Indoor)	0.00/0.00/0.00	128.39	42
Relief 2 (Indoor)	70.76/94.15/98.54	0.02	100
Terrains (Indoor)	80.69/96.14/99.03	0.01	100
SP+LG+ours			
Courtyard (Outdoor)	63.91/92.78/98.20	0.08	84
Electro (Outdoor)	68.60/89.91/93.91	0.34	93
Terrace (Outdoor)	50.79/89.75/97.44	0.07	100
Meadow (Outdoor)	33.56/87.16/96.79	0.23	100
Playground (Outdoor)	0.00/62.70/90.67	0.26	47
Delivery Area (Indoor)	42.15/88.44/97.11	0.06	82
Kicker (Indoor)	54.75/90.87/97.72	0.04	94
Office (Indoor)	0.00/2.30/74.19	0.16	88
Pipes (Indoor)	85.80/97.16/99.29	0.01	100
Relief (Indoor)	14.71/82.09/95.52	0.04	42
Relief 2 (Indoor)	14.21/74.13/93.53	0.12	87
Terrains (Indoor)	37.67/85.06/96.26	0.06	93

Table A.3 The complete results for the focal length test. Each entry is the average absolute focal length distance for the given dataset and method.

	MoGe	SIFT		SP+LG	
		ours	COLMAP	ours	COLMAP
Courtyard (Outdoor)	51.84	18.94	5.00	14.48	21.66
Electro (Outdoor)	217.27	17.57	4.74	26.77	5.00
Terrace (Outdoor)	90.33	18.12	254.73	22.26	15.67
Meadow (Outdoor)	144.90	32.73	21041.24	44.20	14.50
Playground (Outdoor)	234.88	38.97	101.42	29.50	10.11
Delivery Area (Indoor)	119.72	23.92	1171.23	14.00	15.29
Kicker (Indoor)	105.23	31.47	11.35	19.41	24063.35
Office (Indoor)	306.65	42.74	25146258.22	70.50	9.23
Pipes (Indoor)	141.51	20.46	5.00	12.53	4.71
Relief (Indoor)	322.17	38.84	615.59	80.48	2958597.75
Releif 2 (Indoor)	354.77	128.35	319.82	40.20	7.06
Terrains (Indoor)	167.08	636.48	914.42	71.20	9.33

A.3 Complete Low-overlap Pose Estimation Results

In this section we present the full result for each dataset used in the three-view overlap pose estimation test seen in fig. 4.2. The marker X signifies no triplet with three-view overlap in the interval for the given image set.

Table A.4 Complete AUC results for the low overlap three-view relative pose estimation.

	Overlap %					
	0	[0, 5]	[5, 10]	[10, 20]	[20, 40]	> 40
SP+LG+COLMAP						
Courtyard (Outdoor)	5.80/11.94/16.19	8.28/16.13/22.44	15.42/48.16/61.84	36.22/66.19/79.10	63.91/83.83/88.83	77.97/91.38/93.9
Electro (Outdoor)	8.95/17.86/25.21	8.70/16.76/25.05	15.15/28.38/32.10	18.74/33.77/39.21	49.21/63.75/69.19	74.43/87.91/90.4
Terrace (Outdoor)	9.66/28.46/32.11	10.33/25.48/30.42	19.34/44.13/52.70	17.18/26.44/37.68	32.13/48.17/57.49	47.40/55.80/57.37
Meadow (Outdoor)	X	X	X	0.00/6.28/16.55	22.72/35.81/45.30	34.56/63.98/74.35
Playground (Outdoor)	0.00/6.68/15.87	3.52/9.99/18.26	13.68/29.26/32.32	12.29/20.48/25.94	21.75/29.47/34.42	42.44/50.65/53.44
Delivery Area (Indoor)	23.40/31.40/32.85	18.56/28.16/31.94	36.10/49.12/53.94	55.04/69.98/72.90	58.24/71.78/74.61	67.66/72.48/73.38
Kicker (Indoor)	0.00/15.75/20.60	11.84/23.39/26.68	22.21/44.93/55.94	39.59/58.08/66.08	56.57/79.84/86.91	63.48/81.01/84.73
Office (Indoor)	2.58/8.93/19.13	2.52/11.07/20.28	22.47/41.79/53.64	12.62/25.62/35.86	40.63/62.21/71.30	72.21/85.92/88.58
Pipes (Indoor)	25.87/31.84/32.96	25.87/31.84/32.96	31.78/33.02/33.26	33.62/59.50/70.64	64.38/81.20/85.58	90.51/98.10/99.53
Relief (Indoor)	0.00/9.95/19.34	17.26/47.79/53.62	8.27/18.78/21.36	26.67/33.49/36.95	X	65.12/70.57/71.59
Relief 2 (Indoor)	3.66/5.73/10.88	4.66/6.55/12.16	13.27/18.65/28.13	13.74/17.59/21.22	26.67/33.97/36.82	67.39/81.99/84.97
Terrains (Indoor)	25.92/31.92/32.98	24.14/32.45/34.07	62.08/75.33/84.45	72.97/84.61/88.33	80.52/93.12/95.5	86.79/97.36/99.34
SP+LG+ours						
Courtyard (Outdoor)	22.61/68.72/83.87	22.92/66.42/78.87	33.27/76.61/89.42	25.36/68.93/85.27	50.35/89.23/97.31	62.59/89.61/95.66
Electro (Outdoor)	19.87/59.91/81.29	19.43/55.76/77.60	31.37/71.84/93.03	35.15/75.56/93.68	36.49/80.13/94.88	39.53/77.68/94.39
Terrace (Outdoor)	37.93/82.09/95.52	31.83/76.19/91.62	42.15/85.48/96.37	25.11/74.22/93.22	24.03/77.13/94.34	30.65/80.08/95.02
Meadow (Outdoor)	X	X	X	8.65/57.86/89.53	16.71/65.36/91.24	29.41/68.74/91.69
Playground (Outdoor)	26.78/70.17/80.04	18.08/59.19/74.64	27.46/68.18/81.18	18.46/46.89/59.18	24.16/52.72/63.89	28.96/57.46/69.38
Delivery Area (Indoor)	35.02/78.53/92.90	22.28/61.32/80.53	26.33/71.71/85.79	29.16/75.56/88.99	34.27/74.97/86.99	37.14/68.47/78.96
Kicker (Indoor)	18.65/54.76/88.66	11.15/59.81/89.95	35.56/74.24/93.54	22.14/60.03/83.94	31.75/79.49/94.86	43.56/79.81/92.12
Office (Indoor)	8.34/60.59/88.38	13.36/63.43/89.23	11.28/60.68/87.96	22.98/72.76/92.94	23.53/75.61/93.22	28.04/75.53/93.88
Pipes (Indoor)	0.00/69.93/92.48	0.00/71.85/92.96	30.71/85.88/96.47	41.80/83.32/95.83	43.44/85.23/96.31	72.88/94.57/98.64
Relief (Indoor)	12.51/46.16/53.21	17.26/47.79/53.62	29.44/60.84/73.54	14.77/46.90/55.96	X	44.16/64.40/70.05
Relief 2 (Indoor)	17.05/31.79/36.07	16.87/33.41/38.62	26.70/53.07/63.12	17.59/35.45/41.76	18.57/38.13/43.80	44.21/78.14/86.64
Terrains (Indoor)	21.80/65.78/90.35	21.03/65.45/88.18	24.97/68.83/90.97	35.73/78.85/92.08	30.37/75.99/90.79	54.45/88.64/97.16

Master's Theses in Mathematical Sciences 2025:E50
ISSN 1404-6342

LUTFMA-3592-2025

Mathematics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/