Universidad Zaragoza

1542

Escuela de Ingeniería y Arquitectura Universidad Zaragoza

Instituto Universitario de Investigación en Ingeniería de Aragón Universidad Zaragoza

Universidad de Zaragoza

Escuela de Ingeniería y Arquitectura

Final Master's Thesis

Triangulation in Deformable Scenes

Author: Luis Calderón Robustillo
Master: Master's Degree in Robotics, Graphics and Computer Vision

Tutor: Javier Civera Sancho
Research Group: Robotics, Computer Vision and Artificial Intelligence (Ropert)

2024-2025

Abstract

In this work, we address the open problem of two-view triangulation in deformable scenes using monocular cameras with known internal calibration and pose. The goal of triangulation in deformable environments will be established as estimating the 3D structure of the surfaces that are imaged in each camera view. We will formulate such estimation as a deformable model-based non-linear optimization that we will iteratively optimize from an initial seed solution.

In this investigation we evaluate several deformation models, that encode different types of deformations between surfaces, which makes our solutions effective in scenes that exhibit near isometric deformations. In contrast, they are less effective when the surface deformations follow more random patterns, something that we evaluated in simulation. Our model-based optimization, in general, excel on frames that present significant camera translation and rigid deformation. We leveraged the sparse structure of the optimization in our implementation, achieving very low computational loads.

Finally, we also propose in this thesis the use of single-view depth neural networks to model additional constraints for the ill-posed problem of triangulation of deformable scenes. Specifically, we evaluated two distinct cases: perfectly known metric scale for the predicted scale, which fully constraints the problem but may be unrealistic in real setups, and the most realistic one of unknown metric scales. Our experimental results, obtained from both synthetic and real-world datasets, characterize the performance of our solutions, show their effectiveness and give insights for future research on this problem.

Keywords: Two-view triangulation, deformable scenes, monocular 3D reconstruction, model-based optimization, single-view depth.

Contents

1 Introduction **3**

2 Related Work **4**

3 Triangulation methods in deformable scenes **6**
3.1 Notation 6
3.2 Deformation models 7
3.3 Initial seed for the triangulation 9
3.4 Non-rigid optimization 13
3.4.1 Image constraints 13
3.4.2 Depth constraints 15
3.4.3 Up-to-scale depth constraints 15

4 Results **15**
4.1 Datasets 16
4.2 Implementation details 17
4.3 Analysis 18
4.3.1 Synthetic data 18
4.3.2 Real data 26

5 Conclusions **36**

1 Introduction

Triangulation refers to the process of determining the position of a 3D point from its projections onto two or more cameras. Triangulation is a fundamental block within computer vision, that appears in a wide array of tasks, and specifically plays a pivotal role in stereo vision [10, 8], simultaneous localization and mapping (SLAM) [4], and structure-from-motion (SfM) [10, 36]. In the computer vision literature, triangulation solutions are typically defined as those minimizing a function of the re-projection residual, mostly in the image [12, 25, 38, 15, 22] but also in the bearing space [9, 11, 19]. All the methods in the literature, however, assume a rigid scene, which is a typical assumption in SfM and SLAM. They are hence suboptimal when the scene is deforming, a case that has been addressed in several works [30, 34]. Observing a gap in the scientific literature for specific studies of two-view triangulation in deforming scenes, the goal of this thesis is formulating specific solutions to it, that can be used in non-rigid SfM and SLAM.

Reconstructing the 3D structure of dynamic or deforming scenes is a major challenge in computer vision, but with very relevant applications in robotics, medicine, autonomous driving, and augmented reality. Scenarios involving moving objects, such as pedestrians in front of robots or autonomous vehicles, or deformable surfaces, such as internal organs inside the human body seen from an endoscopic cameras (see Figure 1), are well-known challenging cases for which traditional approaches assuming rigidity will fail, and emerging approaches that model the non-rigidities lack accuracy and robustness.

This project introduces a two-step methodology: first, the alternative selection of 3D positions based on ray information in section 3.3, and second, an optimization process presented in section 3.4 leveraging deformable models to refine the initial guess. The study investigates how varying approaches to 3D point positioning impact the precision of the final reconstruction. In addition, to enhance the initial estimates, a thorough review of existing deformable models in the literature is conducted in section 3.2. Finally, it is culminated with the integration and application of model-based nonlinear optimization techniques.

Through the analysis made in section 4, we will explore how, as in many other fields, increased solution complexity does not necessarily yield better results. Additionally, we will examine how factors such as the degree of deformation and the distance between the camera and the points create distinct scenarios for study. Furthermore, we will assess how the models enhance the estimation when deformations are aligned with predefined patterns represented by the models, while demonstrating limited effectiveness for random, unpredictable deformations.

Therefore, by implementing novel nonlinear methods and validating them on both synthetic and real-world datasets, the study aims to provide a robust solution that enhances the performance of current SLAM (Simultaneous Localization and Mapping) and 3D vision systems in deformable environments.

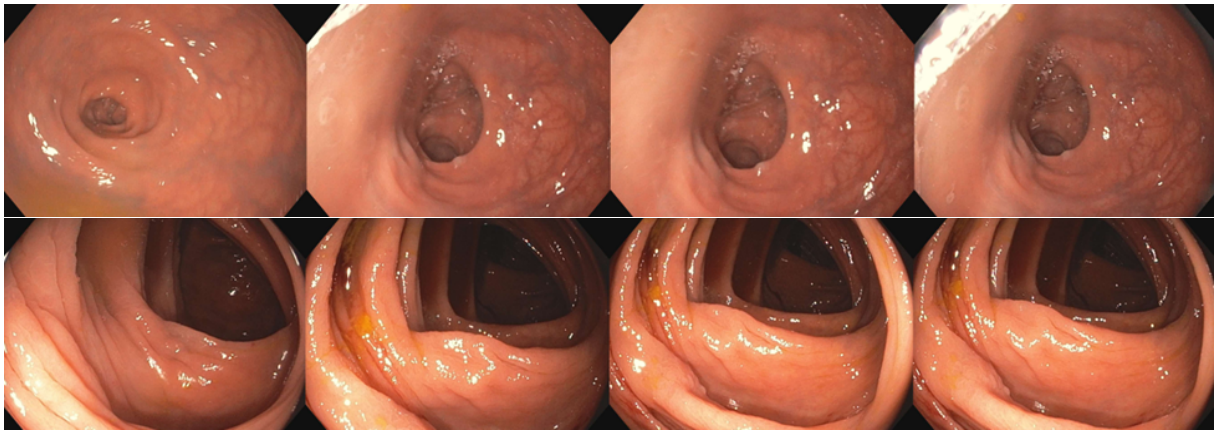

Figure 1: Different sequences of colonoscopy frames of the same patient extracted from [1], as an example of the deformable environment targeted at this master thesis.

2 Related Work

Rigid triangulation is a traditional topic within computer vision for which many effective solutions already exist, although research for new methods still continues. Triangulation basically takes image matches between two or more views and, assuming known extrinsics, estimates the 3D position of the point by minimizing a function of the reprojection error. Variations refer to the specific function of the error, e.g., the L2 [12], L1 [22] or L∞ [9] norms, to image or angular reprojection errors [19], and to two or more views [38]. While the recent global structure from motion pipeline GLOMAP [26] avoids the triangulation stage for 3D reconstruction, incremental structure from motion [36] uses it, as well as visual SLAM [5].

Up-to-scale single-view depth predictions can be nowadays available from deep networks such as [42, 29, 3], and have been used in the literature for relative motion estimation [13], but their use for triangulation as we do has been never explored in the literature.

3D reconstruction and localization in non-rigid scenes remain as open research challenges. Their complexity arises from the ill-posed nature of the problem with monocular data: multi-view matches do not fully constraint all degrees of freedom of the system, being needed to add further constraints in the deformation model to find a unique solution. In the last decade, however, visual SLAM pipelines have been expanding their applicability from rigid to deformable environments.

DynamicFusion [24] is one of the most referenced approaches able to achieve 3D reconstructions of deforming scenes by fusing RGB-D scans. The literature on non-rigid real-time reconstruction using RGB-D sensors includes many other works, for example, [20, 21, 44, 43]. Depth measurements make the estimation problem observable, and hence all these works present impressive results. Triangulation is straightforward from the depth measurement of just one view. However, depth sensing is not typically available in many potential application domains, such as colonoscopies.

There is significant research works on the topic denoted as non-rigid structure from motion [6, 7, 27, 39, 40]. They typically make use of isometric constraints to constraint the deformations of the scene, and jointly and globally solve for the scene structure and the camera motion. In general, existing datasets, such as [14], image small-scale scenes

for which global structure from motion is solved globally and not incrementally, and triangulation does not bear a high relevance. In the structure from motion literature, we have not found explicit works on triangulation in deforming scenes.

DefSLAM [17] is the first SLAM pipeline designed for deformable scenes using only monocular images. However, it does not perform triangulation between frames. Instead, it initializes the map representation from the first frame as a planar template at some depth, and uses the multi-view constraints from posterior frames to make it converge to appropriate values.

NR-SLAM [33] does perform two-view triangulation in deforming scenes as part of its mapping pipeline. The procedure involves first an attempt for rigid triangulation, acknowledging the instability of non-rigid triangulation and quasi-rigidity in certain parts of the colonoscopies they use. If rigid triangulation has a low reprojection error, they stick to this solution. If the reprojection error is high, they use neighboring points already in the representation to initialize the point with an average depth. This method, while effective, assumes the existence of a previous 3D map and hence cannot be used from two views from scratch, as the triangulation problem is typically defined.

In this work, we address the gap in the literature for non-rigid triangulation from two view with known intrinsics and extrinsics, just from image matches and without any other explicit knowledge about the scene structure. We believe this is a new problem for which ours is just a first step, that deserves further investigation and will be of interest for non-rigid structure from motion and SLAM pipelines.

3 Triangulation methods in deformable scenes

Throughout this section, we will begin by defining the notation used at the thesis to ensure clarity and facilitate understanding. Secondly, we will review some existing deformable models available in the literature. The goal of this task is to identify potential optimization terms that could serve as valid constraints for our specific scene characteristics. Next, we will explore the challenges of determining an initial 3D position in space using only ray information, in order to use it as initial seed to our non-linear optimization, introducing three distinct solutions to address this problem. However, these solutions are subject to two distinct cases that significantly impact the final solution's accuracy. Finally, we will present the non-linear optimization formulations that target to refine the initial estimates into an optimal one in the cost functions defined, hopefully reducing the initial triangulation error.

3.1 Notation

We are triangulating from two views, taken at different time instants $k = 0$ and $k = 1$. We will denote the two cameras at $k = 0$ and $k = 1$ as C_0 and C_1.

The pose of the two cameras is modeled as a rigid transformation consisting of a rotation matrix $R_{wk} \in SO(3)$ and a translation vector $t_{wk} \in \mathbb{R}^3$. These transformations describe how each camera C_k is positioned relative to the world reference frame w, and transforms 3D points from C_k's reference frame to world. The rigid transformation matrix of a camera k with respect to the world is denoted as $T_{kw} \in SE(3)$, encapsulating both rotation and translation.

The 3D structure of the scene is represented by a cloud of points in the world frame. For a given point j, its 3D position in the world frame w is denoted as $X_{wj} \in \mathbb{R}^3$. In a deformable scene, the position of this point may vary over time, so we introduce X_{wkj} to represent the 3D position of point j at time instant k.

From feature extraction and matching, we may have 2D observations of these 3D points in the images taken by both cameras. For a given image taken at time k and a given point j, the observed 2D keypoint in the image is denoted as $u_{kj} \in \mathbb{R}^2$. These image points are related to their corresponding 3D points through the camera projection function $u_{kj} = \pi_k(X_{wkj}, T_{kw})$, which projects a 3D point X_{wkj} to its 2D image coordinates u_{kj} according to the camera's extrinsic and intrinsic parameters, T_{kw}.

In order to handle deformations for triangulating points, we introduce energy terms into our formulation, which will be applied to the 3D meshes constructed from the points at each time instance k. These terms ensure that the relative distances between points in the 3D mesh remain as similar as possible, thereby regularizing our solutions towards the smallest possible degree of non-rigidity but always complying with the reprojection error. We define weights ϕ_r and ϕ_d to relate the relative importance of the reprojection error with respect to the deformation constraints we impose. Within this deformation model-based optimization we assign weights w_{ji} to pairs of points j and i, computed using the angles $\theta_{j,i}$ opposite the edge formed by these points in the mesh. These weights influence the deformation constraints.

We assume that all our residuals, r_{jk}, follow a Gaussian distribution, $r_{jk} \sim \mathcal{N}(0, \Sigma_k)$. The covariance matrix Σ_k is typically defined as isotropic, $\Sigma_k = I_2$, where I_2 is the

2×2 identity matrix. In the case of the reprojection error constraints, this reflects the uncertainty in the observed keypoints, measured in pixel units.

The cost function of the optimization will be evaluated using 3D meshes formed by the points at a given instant k. These meshes, constructed using Delaunay triangulation as described in Section 3.2, will define the structure of the point set in order to compute the deformations. In this context, $N(j)$ represents the set of neighboring vertices of point j within each triangle of the mesh.

For the sake of clarity, we summarize below the notation for the main concepts that we introduced in this section and will appear in the next ones.

- C_0, C_1: Cameras at two different time instants $k = 0$ and $k = 1$.
- k: The frame index or time instance.
- $T_{kw} \in SE(3)$, $R_{kw} \in SO(3)$ and $t_{kw} \in \mathbb{R}^3$: Transformation matrix (camera pose), rotation matrix and translation vector of camera C_k with respect to the world.
- $X_{wkj} \in \mathbb{R}^3$: 3D position of point j in the world frame at instant k.
- $u_{kj} \in \mathbb{R}^2$: Observed 2D keypoint corresponding to 3D point X_{wkj}.
- $\pi_k(\cdot)$: Camera projection function for camera k, mapping 3D points to 2D image points.
- ϕ_r and ϕ_d: Weights to balance, respectively, the reprojection error and deformation error terms in the optimization.
- w_{ji}: Weight associated with the pair of points j and i.
- $\theta_{j,i}$: Angles opposite to the edge formed by the points X_{wkj} and X_{wki} in the mesh.
- r_{jk}: Residuals assumed to follow a Gaussian distribution $r_{jk} \sim \mathcal{N}(0, \Sigma_k)$.
- Σ_k: Covariance matrix of the residuals in the image domain.
- $N(j)$: Set of neighboring vertices of point j in triangle of the mesh.

3.2 Deformation models

The technical literature offers a vast array of deformation models to choose from, varying in complexity, and that encode different material deformation properties. Deformation models are widely applied in various tasks such as image registration [41], image recognition [16], and simultaneous localization and mapping (SLAM) [33]. In these projects, common model constraints are utilized, including elastic tensors such as those proposed in [17], which relate the stretching energy applied to the surface to the resulting deformation. Hyperelastic models, such as the Mooney-Rivlin model [23] [32] or the Ogden model [2], are employed to describe materials that undergo large deformations, like rubbers and biological tissues. Viscous and viscoelastic models, such as the Maxwell or Zener models surveyed in [35], combine a spring (elasticity) and a damper (viscosity) in series, making them suitable for materials that exhibit initial elastic behavior followed by flow.

Furthermore, viscoplastic models, such as those presented in [28], describe materials that undergo unrecoverable deformations once a critical load level is exceeded.

In this thesis, we have implemented and tested the As-Rigid-As-Possible (ARAP) [37] model in addition to various elastic and viscoelastic terms. For all of them, we have used a *Delaunay* triangulation. In this section, we detail the specific formulations the energy terms corresponding to the deformation models we have evaluated.

- **As-rigid-as-possible (ARAP) model**: It models the rigidity of the surface mesh by summing up the deviations from rigidity for each cell. The energy term is specifically given by:

$$E_{\text{ARAP}} = \sum_{j=1}^{n} \sum_{i \in N(j)} w_{ji} \|(p'_j - p'_i) - R_j(p_j - p_i)\|^2 \tag{1}$$

  where $N(j)$ denotes the set of neighboring vertices of point j in each triangle of the mesh, p_j and p_i are the 3D positions of the points in the original mesh, p'_j and p'_i are the same points in the mesh of the deformed surface and R_j is the rotation matrix for each triangle.

  The weights w_{ji} are defined as:

$$w_{ji} = \frac{1}{2}\left(\cot\theta_{j,i} + \cot\theta_{j,i}\right) \tag{2}$$

  where $\theta_{j,i}$ and $\theta_{j,i}$ are the angles opposite to the edge formed by the points p_j and p_i.

- **Elastic model**: The difference in the length l_e^t of each edge e in each triangle of the mesh in the frame t with respect to its length l_e^k in the shape-at-rest of T_k is given by:

$$E_{\text{Elastic}} = \sum_{j=1}^{n} \sum_{i \in N(j)} \left(\frac{l_e^1 - l_e^0}{l_e^0}\right)^2 \tag{3}$$

  where l_e^0 and l_e^1 are:

$$l_e^0 = p_j - p_i, \quad l_e^1 = p'_j - p'_i$$

- **Hyperelastic Odgen**: In mechanics of rubber, the representation of the Ogden model depends on the principal stretches. The general form of the energy function is given as a convergent series of the principal stretches:

$$E_{\text{Hyperelastic Odgen}}(\lambda_1, \lambda_2, \lambda_3) = \sum_{p=1}^{N} \frac{\mu_p}{\alpha_p}\left(\lambda_1^{\alpha_p} + \lambda_2^{\alpha_p} + \lambda_3^{\alpha_p} - 3\right),$$

  where α_p and μ_p are real numbers characteristic of the material, such that:

$$\mu_p \cdot \alpha_p > 0.$$

For equal-biaxial tension, $\sigma_1 = \sigma_2 = \sigma$ and $\sigma_3 = 0$. So we get:

$$\lambda_1 = \lambda_2 = \lambda, \lambda_3 = \frac{1}{\lambda^2}.$$

Thus, the strain energy function becomes:

$$E_{\text{Hyperelastic Odgen}} = \sum_{p=1}^{N} \frac{\mu_p}{\alpha_p} \left(2\lambda^{\alpha_p} + \lambda^{-2\alpha_p} - 3\right). \tag{4}$$

[2] defines various standard α_p and μ_p tested in LMA method. In the other hand, the stretch factor λ_{ij}, to be applicable to the 3D meshes formed by the points, can then be defined as:

$$\lambda_{ji} = \frac{\|p'_j - p'_i\|}{\|p_j - p_i\|}.$$

3.3 Initial seed for the triangulation

In the traditional triangulation problem for a rigid scene, various challenging situations are addressed, such as the parallax angle between cameras, data association, and outlier detection. Once these issues are resolved, a 3D position must be selected for the points matched between frames to determine the precise locations of objects in the scene relative to the cameras.

In the monocular case, where camera poses are known, only the rays' information is available to determine the 3D points' positions. Using an SVD-based triangulation of the image matches, the initial solution is often chosen as the 3D points computed from the cross-product of the orthonormal vector derived from the SVD solution and each projection ray (see Figure 2). Alternatives, such as the method proposed in [18], refine this selection to better handle low-parallax scenarios, and it will be the final implemented equation.

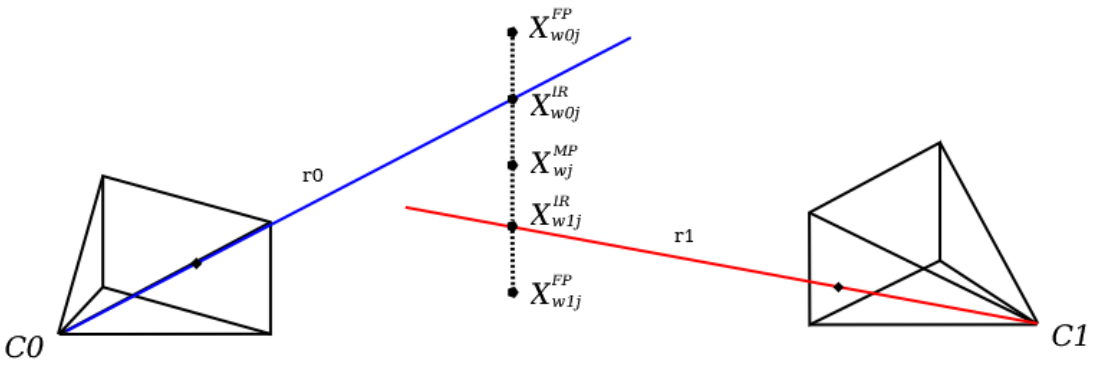

Figure 2: Illustration of the three different initial seeds evaluated for two-view non-rigid triangulation.

From these triangulation methods, for rigid scenes, it is logical to select the midpoint between the rays, assuming a single 3D point observed from both cameras with some

reprojection error. However, in deformable scenes, the 3D point matched between frames has a position in each camera's frame. Thus, two distinct solutions must be chosen for each match, one for each image. Figure 2 illustrates three possible approaches:

The **MidPoint** solution, denoted as X^{MD}_{w0j}, assumes no deformable motion between the 3D points, maintaining a consistent reprojection error across both frames. In contrast, the **InRays** solution, X^{IR}_{w0j}, provides an initial estimation with zero reprojection error for each point and hence allowing for some relative motion between them. Finally, the **FarPoints** solution, X^{FP}_{w0j}, generates a seed using the distance between the rays and X^{MP}_{wj}, while traversing the orthonormal vector in the opposite direction. Consequently, each point exhibits a combination of reprojection error relative to the images and a degree of motion between frames.

To calculate these solutions, we utilize equations extracted from [18]. First, we consider two input ray vectors f_0 and f_1, which are normalized as follows:

$$\hat{f_0} = \frac{f_0}{\|f_0\|}, \quad \hat{f_1} = \frac{f_1}{\|f_1\|}. \tag{5}$$

The input camera poses T_{0w} and T_{1w} are used to calculate T_{10}, which transforms all operations to the C_1 reference frame:

$$T_{10} = T_{1w} T_{0w}^{-1},$$

where T_{10} contains the translation vector t and the rotation matrix R (we will omit subindices 0 and 1 for clarity in the rest of this section).

Next, we compute auxiliary vectors:

$$p = (R\hat{f_0}) \times \hat{f_1}, \quad q = (R\hat{f_0}) \times t, \quad r = \hat{f_1} \times t. \tag{6}$$

The depths λ_0 and λ_1, given by the Generalized Weighted Midpoint (GWM) Method, are calculated as:

$$\lambda_0 = \frac{\|r\|}{\|p\|}, \quad \lambda_1 = \frac{\|q\|}{\|p\|}. \tag{7}$$

Using these depths, we compute the intermediate points along the rays, s_0 and s_1, defined as:

$$s_0 = \lambda_0 R\hat{f_0}, \quad s_1 = \lambda_1 \hat{f_1}. \tag{8}$$

Finally, we compute the Inverse Depth Weighted MidPoint x_1:

$$x_1 = \frac{\|q\|}{\|q\| + \|r\|} \left(t + \frac{\|r\|}{\|p\|} (R\hat{f_0} + \hat{f_1}) \right). \tag{9}$$

From (9), the final 3D points X_{w0j} and X_{w1j} in the **MidPoints** case are determined as:

$$X^{\text{MP}}_{w0j} = x_1, \quad X^{\text{MP}}_{w1j} = x_1. \tag{10}$$

For the **InRays** solution, we directly use the points from (8):

$$X^{\text{IR}}_{w0j} = t + s_0, \quad X^{\text{IR}}_{w1j} = s_1. \tag{11}$$

Finally, the **FarPoints** solution is computed as follows:

$$X_{w0j}^{\mathrm{FP}} = X_{w0j}^{\mathrm{IR}} + (s_0 - x_1), \quad X_{w1j}^{\mathrm{FP}} = X_{w1j}^{\mathrm{IR}} + (s_1 - x_1). \tag{12}$$

Finally, the points are transformed back to the world frame:

$$X_{w0j} = T_{1w}^{-1} X_{w0j}, \quad X_{w1j} = T_{1w}^{-1} X_{w1j}. \tag{13}$$

Once the three proposed initial seeds are introduced, the focus shifts to analysing the available information to derive the optimal final solution. Two key observations guide this process: First, certain deformations may occur along the space between the frames 0 and 1. Second, inlier matches are subject to noise that we will assume as $r_{jk} \sim \mathcal{N}(0, \Sigma_k)$, introducing an uncertainty of approximately 1 pixel standard deviation in each image axis.

Considering these factors, we identify two distinct scenarios that depend on the distance between the cameras and the 3D points, as well as the degree of deformation between frames 0 and 1. These scenarios are illustrated in Figures 3 and 4.

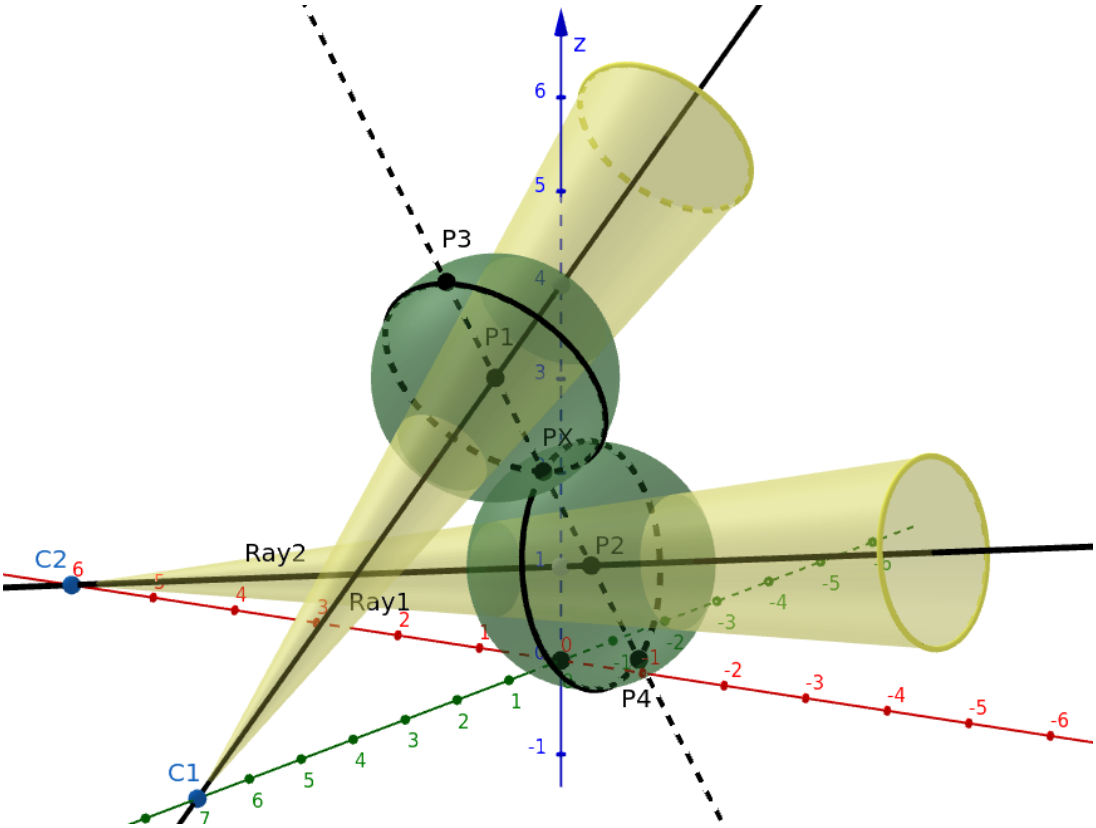

Figure 3: 3D visualization of Case 1 triangulation. Yellow cones represent the uncertainty corresponding to a 1-pixel reprojection error, while green spheres have radii equal to the distances between P_x and P_1 or P_2, respectively.

In the first scenario, depicted in Figure 3, the short distance between the camera and the points and/or the significant deformation in the 3D positions leads to well-separated uncertainty regions for the rays, represented as yellow cones.

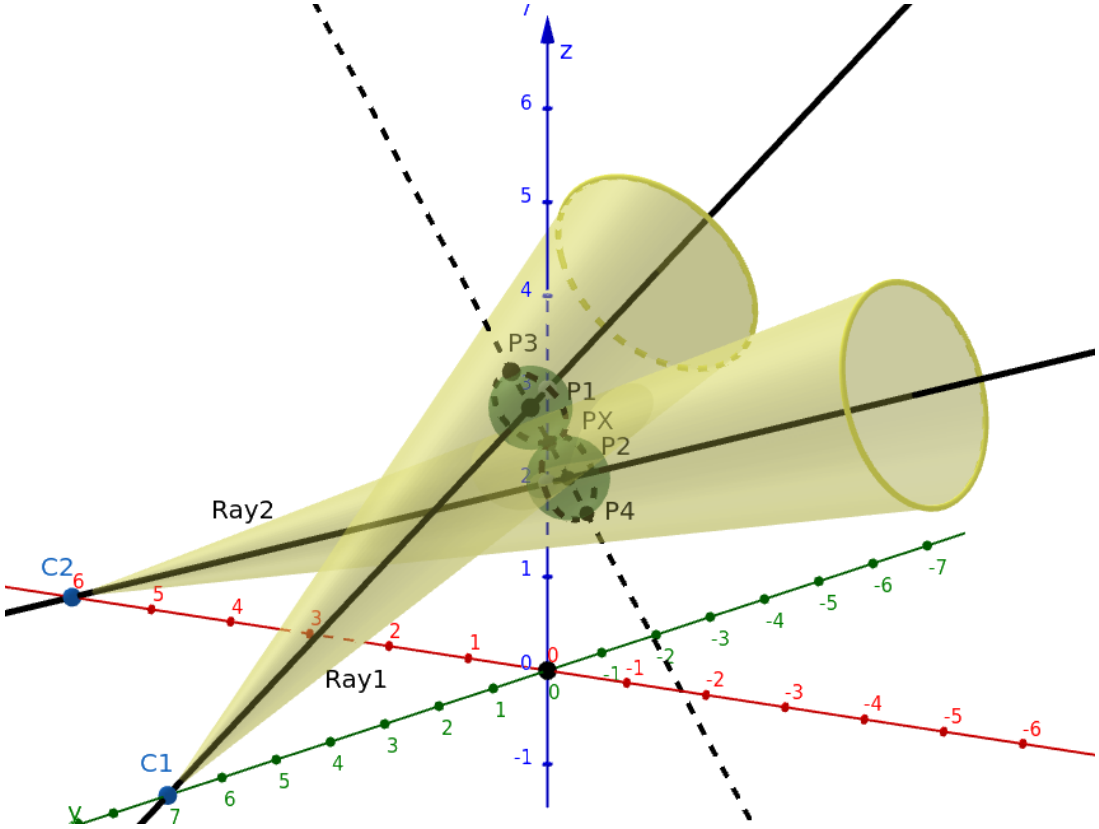

Figure 4: 3D visualization of Case 2 triangulation. Yellow cones represent the uncertainty corresponding to a 1-pixel reprojection error, while green spheres have radii equal to the distances between P_x and P_1 or P_2, respectively.

The triangulated 3D points are expected to have a 1-pixel reprojection uncertainty and lie close to their corresponding rays. Therefore, probable solutions are located on the convergence surfaces formed by the intersection of the green spheres and yellow cones. Based on this representation, for Figure 3:

- **MidPoints** (PX) and **FarPoints** ($P3$ and $P4$) must approach their respective rays to converge toward the solution within the convergence surface.
- In contrast, **InRays** ($P1$ and $P2$) must move away from the rays to achieve the same goal.

This process ensures a closer alignment to the most probable solution among the infinite possibilities on the convergence surface.

In the second scenario, represented in Figure 4, the long distance between the cameras and the points and/or the small deformation over 0 and 1 3D points leads to joined uncertainty regions. In this case, the spheres of approximation to the rays based on the initial seed are no longer valids to stablish a convergence surface of possible solutions. Therefore, all the initial guesses pass trough the idea of moving away from the rays to select a solution closer to the probable solution among the infinite possibilities, but we can already determinate that this is a problematic case where the uncertainty is much more bigger than the known information.

3.4 Non-rigid optimization

3.4.1 Image constraints

From the initial triangulation seeds defined in Section 3.3, here we aim to define a non-linear optimization problem that, by minimizing energy terms corresponding to deformation patterns and the reprojection error, refines such initial guesses to approach the most probable solution among the infinite possibilities in the space.

Let us assume two cameras, C_0 and C_1, at two different time instances, $k = 0$ and $k = 1$. Fixing the reference frame w in the local frame of C_0, the two cameras are related by a rigid transformation $R \in SO(3)$, $t \in \mathbb{R}^3$.

The reprojection error term, which is the only one that is minimized for rigid triangulation, can be formulated as:

$$\min_{X_{wj}} \sum_{k,j} \|u_{kj} - \pi_k \left(X_{wj}, T_{kw}\right)\|^2 \tag{14}$$

Here, u_{kj} denotes the observed keypoint corresponding to the 3D point X_{wj} in frame k, $\pi_k(\cdot)$ models the camera projection function for camera k, T_{kw} is the transformation matrix (camera pose) of camera k with respect to the world, and X_{wj} represents the 3D point j in space. With this formulation, there are 6 unknowns (degrees of freedom) for each camera and 3 unknowns for each point (position). However, for each image match between cameras (a 3D point), there are 4 new equations (x, y in two images). Therefore, for a sufficient number of matches, the problem becomes solvable.

In the case of a deformable scene, however, there are infinite solutions to choose from because there is a 3D point X_{wj} at each instance, and the point X_{wj} at time 0 may have moved at time 1.

To address this problem in a monocular camera, we propose a new formulation of the optimization equation that provides a valid constraint, selecting a solution from the infinite possibilities using the reprojection error and establishing a deformation energy to optimize. The modified optimization is expressed as follows:

$$\min_{0,1} \ \phi_r E_r + \phi_d E_{mov} \tag{15}$$

Thus, the formulation in (15) is established for each pair of frames 0 and 1, incorporating two distinct terms: the reprojection error term and the deformation term.

This formulation also introduces two weighting factors, ϕ_r and ϕ_d, which control the relative contributions of these terms in the optimization process. In this case, we assume a Gaussian distribution for the residuals $r_{jk} \sim \mathcal{N}(0, \Sigma_k)$. The covariance is typically defined as isotropic I_2 in the image domain, and it is given in pixel units. Therefore, ϕ_r is assigned a fixed value of 1, while ϕ_d is optimized through an external non-linear optimization process to approximate a reprojection error standard deviation of 1 pixel in the final solution.

The first term, related to the reprojection error, is defined as:

$$E_r = \sum_{0,1,j} \|u_{0j} - \pi_0 \left(X_{w0j}, T_{0w}\right)\|^2 + \|u_{1,j} - \pi_1 \left(X_{w1j}, T_{1w}\right)\|^2 \tag{16}$$

As previously discussed, the camera pose T_{kw} is known to restrict the possible solutions for the 3D positions of the points. The only change in the reprojection term from the traditional BA in (14) is the introduction of X_{w0j} and X_{w1j}.

The next step is defining the E_{mov} term of (15). To do this, we will use the deformation models presented in Section 3.2 to restrict the deformation energy following one of these movement patterns. However, these models will apply constraints on the relative distances between the 3D vertices of the meshes formed by the points in each instance 0 and 1. For instance, the ARAP model forces the meshes to maintain equal relative distances, without accounting for a potential absolute motion between the meshes in space. The same principle applies to the other models. This results in an additional term that represents the global transformation between surfaces, which minimizes the rigid movement between the 3D meshes:

$$E_{mov} = E_{non-rigid} + E_{rigid} \tag{17}$$

The global transformation proposed is an extra term included in the energy movement, represented as:

$$E_{rigid} = \min \sum_{0,1,j} \| \left(R_g X_{w1j} - t_g\right) - X_{w0j} \|^2 \tag{18}$$

Here, we aim to obtain the values of the 3D points as well as the global transformation T_g, with $R_g \in SO(3)$ and $t_g \in \mathbb{R}^3$ between the frames points meshes.

In the case of the ARAP term, for a 3D point mesh in space, Equation (1) ensures that the relative distances between points in consecutive frames remain as consistent as possible (see [37] for further details). The model is formulated as:

$$E_{non-rigid} = \sum_{0,1,j,i} w_{ji} \| \left(X_{w0j} - X_{w0i}\right) - R_j \left(X_{w1j} - X_{w1i}\right) \|^2 \tag{19}$$

Similar substitutions have been applied for the remaining models described by Equations (3) and (4).

However, it is important to highlight that, in the ARAP case, the rotations R_j for fixed C_0 and C_1 must be computed in a preliminary step prior to the entire optimization process, as indicated in [37]. Consequently, these rotations will be treated as fixed values during the method's execution.

To calculate their values, we define for C_0:

$$e_{ji} := X_{w0j} - X_{w0i},$$

and similarly, e'_{ji} for the deformed cell C_1. Let S_j denote the covariance matrix:

$$S_j = \sum_{j,i} w_{ji} e_{ji} e'^T_{ji} = P_j D_j P'^T_j,$$

where D_j is a diagonal matrix containing the weights w_{ji}, P_j is the $3 \times N(j)$ matrix containing e_{ji}'s as its columns, and P'_j is similarly defined for e'_{ji}. The rotation R_j can then be derived from the singular value decomposition of $S_j = U_j \Sigma_j V_j^T$:

$$R_j = V_j U_j^T,$$

ensuring that the column of U_i corresponding to the smallest singular value is adjusted, if necessary, to satisfy $\det(R_j) > 0$. For further details, refer to [37].

3.4.2 Depth constraints

If depth measurements are available, we can include the following additional term in the non-linear optimization of the equation 15, in order to minimize the depth values of each 3D point with respect to its depth measurement

$$E_{depth} = \phi_{depth} \sum_{k,j} d_{kj} - z_{kj} \tag{20}$$

where d_{kj} is the depth measurement in instant k of point j and z_j is the z-component (depth) of the transformed point j in the camera frame. In this case, the problem would be fully observable and would offer the best results. However, having depth measurement is not realistic in many situations, such as endoscopic sequences.

3.4.3 Up-to-scale depth constraints

In the other hand, normally depth measurements from monocular cameras images are taken from AI tools. Therefore, a common case pass through having depth measuremnets up-to-scale. Therefore, an extra term would be added to optimize a global scale value for each frame:

$$E_{depth} = \phi_{depth} \sum_{k,j} d_{kj} - z_{kj} s_k \tag{21}$$

where s_k is the optimizable scale for each instant k.

4 Results

This section presents an evaluation of the proposed approaches, detailing first both the synthetic and real-world scenarios used for validation and testings and some implementation details. The simulation framework provides precise control over parameters such as camera poses, keypoint matches, and deformation, enabling accurate error quantification in selected setups. Conversely, the more realistic Drunkard's dataset that we also use, introduces effects such as unknown matching noise and outliers, thus offering deeper insights into the robustness and practical applicability of our developments.

Our analysis begins by describing the datasets and implementation details of our experiments. The proposed pipeline for generating synthetic data is outlined in Figure 7, together with the triangulation pipeline. The triangulation pipeline is exactly the same for the synthetically generated data and for the image matches extracted from the Drunkard's dataset sequences. After data considerations, we delve into the quantitative and qualitative assessment of the equations' performance under various conditions, demonstrating the validity and effectiveness of the approach.

4.1 Datasets

While synthetic data enables complete control over the dataset, more realistic images allow us to evaluate our method closer to application setups. Therefore, the first dataset was created by us, considering various deformation scenarios and different camera setups. The more realistic (although still synthetic) dataset selected for this evaluation comes from [31].

For the synthetic scenarios generated, we simulated camera parameters with sufficient parallax for precise point triangulation. However, we varied four key characteristics, which are presented in the following. The first item represents the primary characteristic of the environment, and there are cases from this value combined with all cases from the subsequent items.

1. **Distance of points from cameras**: 20 cm, 80 cm, and 150 cm. Examples in Figure 5.
2. **Deformation shape**: The rigid deformation can be **planar**, where the entire surface undergoes the same movement, or **gradual**, where movement varies across the surface and increases progressively through the mesh. Examples in Figure 6. Curved deformations have not been tested in these simulations, however, they are present in the most realistic Drunkard's dataset that we also use.
3. **Deformation pattern**: **Rigid** deformation refers to uniform deformation of the entire surface in a single direction, while **Gaussian** deformation describes aleatoric deformation of each point based on its initial distribution. Both rigid and Gaussian deformations have been tested, as well as their simultaneous application.
4. **Deformation magnitude**: Deformation magnitudes of 2.5 mm and 10 mm were tested for all previous cases.

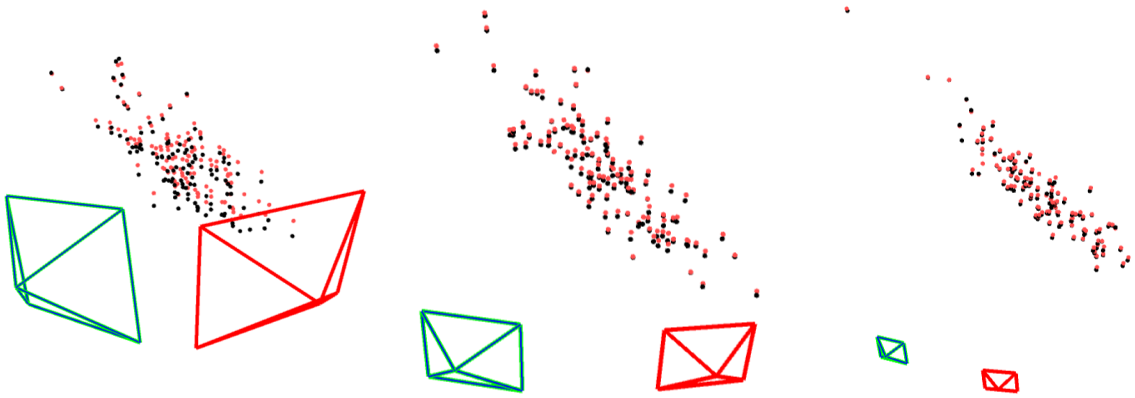

Figure 5: Triangulation examples with different points-to-camera distance: 20cm (left), 80cm (middle) and 150cm (right). Both camera poses are represented by green and red pyramids. Red points correspond to 3D positions observed from C_0 and black points correspond to points observed from C_1

On the other hand, the Drunkard's dataset [31] was selected due to the difficulty of finding real ones that include ground truth and varying levels of deformation. Drunkard's

allows us to thoroughly test our implementation and analyze our approach. Specifically, it contains 19 scenes with four levels of deformation difficulty. All scenes are available in resolutions of 1024×1024 or 320×320 and provide camera poses throughout the entire trajectory, along with color images, depth maps, surface normals, and optical flow data.

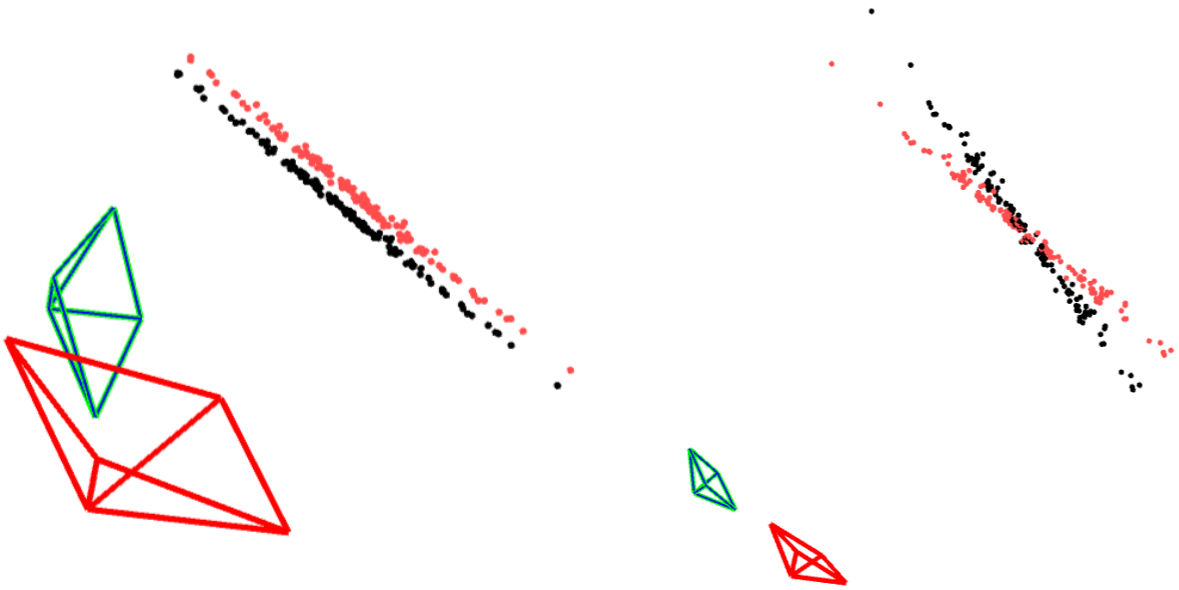

Figure 6: Triangulation examples with different shape deformation: Planar (P) (left) and Gradual (G) (right). Both camera poses are represented by green and red squares pointing to the points. Red points correspond to 3D positions observed from C_0 and black points correspond to points observed from C_1.

4.2 Implementation details

To test and validate our formulation, we have firstly developed a simulation environment to use synthetic data with known camera poses, matches, deformation and 3D points positions solution.

Figure 7 presents the environment pipeline. First, the simulation data is created with customisable parameters and values (deformation sizes and reprojection error), then the key points/matches are the starting point of the triangulation block.

In this way, the various proposed equations can be implemented and evaluated on simulated data and thus exact measurements of the errors obtained. In the other hand, for real data the second triangulation pipeline is followed using same methods but introducing new steps to process and find matches in real images.

The implementation of the project has been developed entirely in C++17. Key libraries such as OpenCV, Sophus, and Eigen were extensively utilized to handle various computer vision tasks. For the creation and computation of Delaunay meshes, Open3D and Qhull were employed. Additionally, the g2o library was used for the proposed non-linear optimization, while the NLopt library handled the external non-linear optimization of the weight ϕ_d.

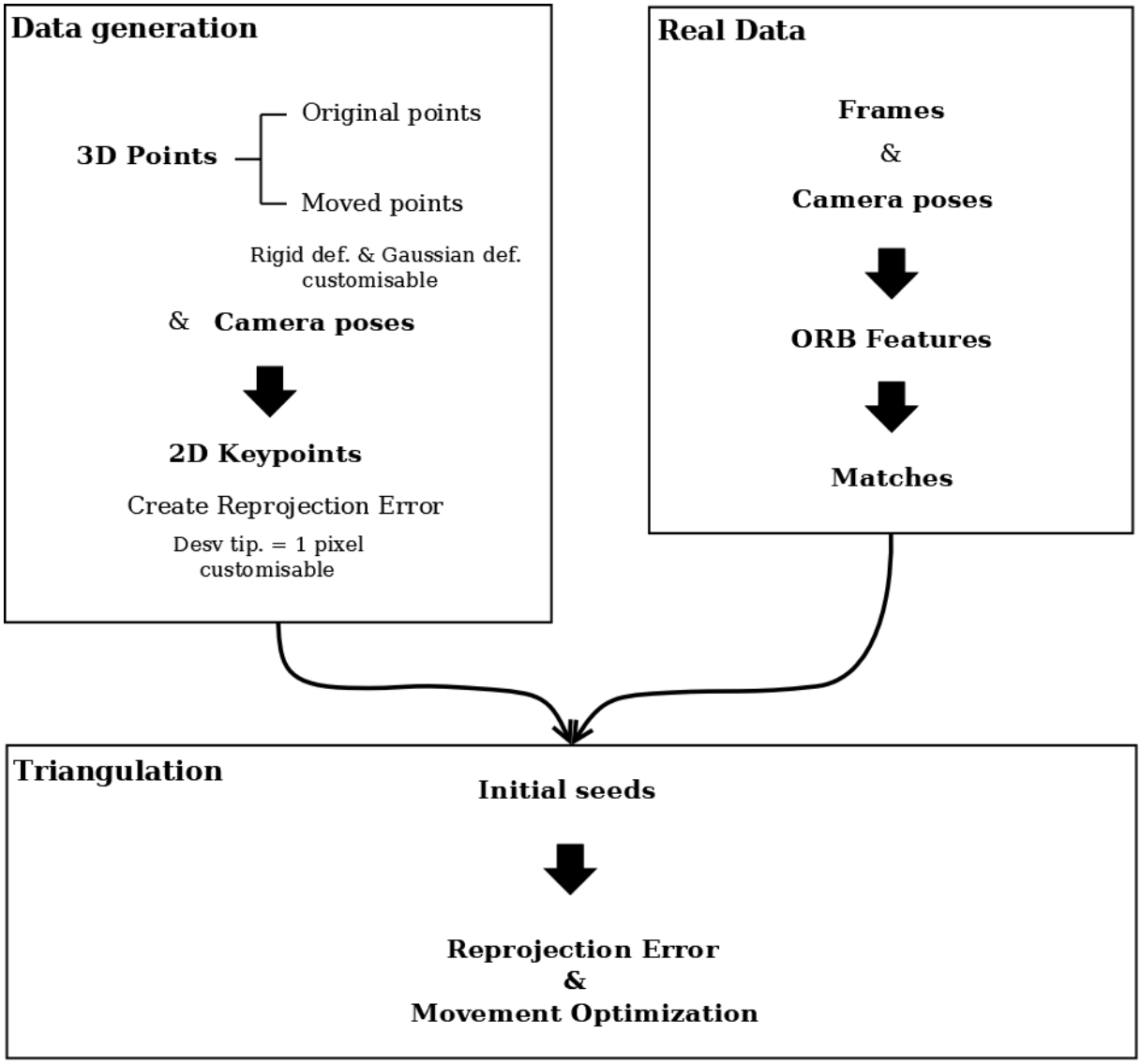

Figure 7: Pipeline depicting the process from data generation (simulation) and extraction (real images) to the 3D triangulation of the scene.

4.3 Analysis

4.3.1 Synthetic data

Initially, Table 1 presents a comparative analysis of the final error in millimeters across three modeling approaches: the ARAP approach 3.2, elastic model 3.2, and hyperelastic model 3.2. In all three cases, the equations incorporate an additional global transformation term, detailed in Equation 18, to address non-relative deformations. To illustrate its effectiveness, the table includes a fourth column displaying the ARAP equation results obtained without utilizing the global transformation term.

From Table 1, several key conclusions can be drawn. Most notably, the "as rigid as possible" (ARAP) model with the global transformation term demonstrates superior performance across all experiments, establishing itself as the preferred approach for subsequent stages. This outcome is substantiated by experimental data that unequivocally

validates the ARAP model as the most closely representing rigid deformation patterns compared to elastic and hyperelastic approaches. Moreover, the hyperelastic model introduces additional complexity through its requirement of manual material parameter selection.

The analysis reveals that aleatoric movement inherently introduces noise to predictive models due to the absence of a discernible pattern. Consequently, the optimal results are observed in the ARAP model with global transformation term, particularly for rigid deformations.

After selecting the ARAP approach, Table 2 presents the final results for this model using different initialization seeds (InRays, MidPoints, and FarPoints). In scenarios with camera-to-points distances relatively low compared to the movement magnitude (specifically the 20 cm distance cases), the initial seed selection shows a small impact. These cases are represented in Figure 3, and the randomness behind which seed provides the best result, and the variances observed between experiments indicate that optimization's effectiveness when changing seed positions is largely attributable to chance in these scenarios.

Conversely, when camera-to-point distances increase while maintaining movement magnitudes (hence, for lower parallaxes), the scenarios are illustrated in Figure 4. In these instances, the average 3D errors increase with movement, and the FarPoints approach emerges as the most effective seed initialization strategy. Therefore, we will use FarPoints seed in next stages although we must remember that in the first scenario of high parallax, the selection of the seed was not relevant.

After selecting the initial seed approach, Table 3 presents a comparative analysis of initial versus final 3D average errors. In low-parallax scenarios (camera-to-points distances of 80 and 150 cm), there are higher initial and final error values compared to 20 cm camera-to-points distance situations, but consistent improvement is observed. Conversely, in 20 cm camera-to-points distance scenarios, performance improvements are predominantly evident in cases involving rigid deformations, aligning with our prior expectations. In these experiments, the average error remains consistently lower than the applied movement, and the standard deviation between experiments is minimal, demonstrating the method's stability. In contrast, experiments involving substantial deformations—specifically the "$20-Planar-10g10r$" and "$20-Gradual-10g10r$" cases—exhibit not only increased initial errors but also a pronounced degradation in the optimization process. The substantial standard deviations observed under these extreme deformation conditions indicate that these are the sole experiments revealing system instability.

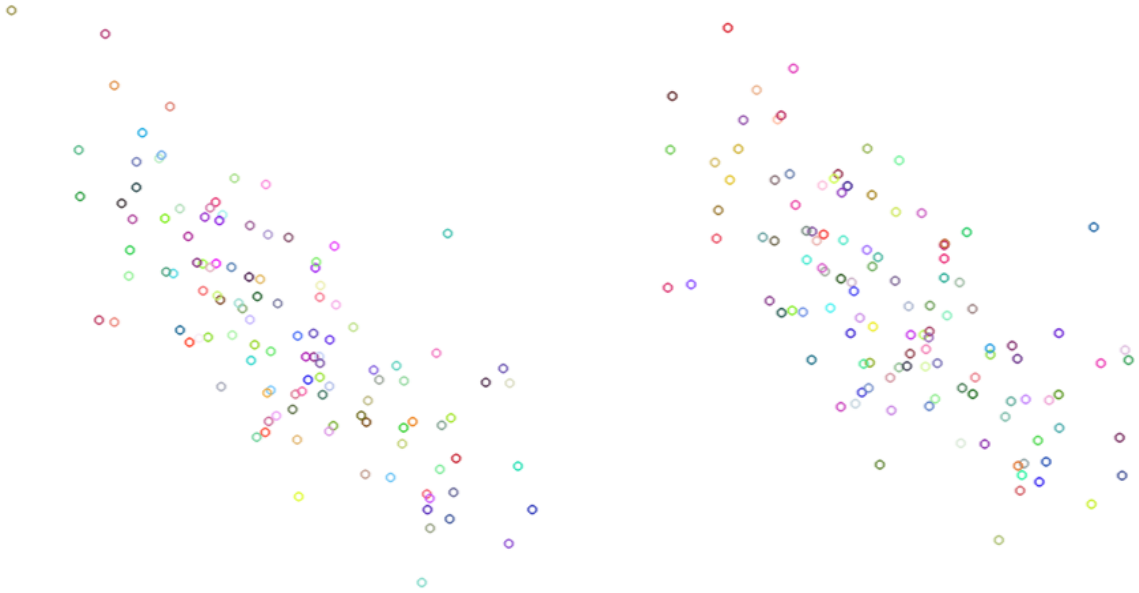

Figure 8: Matches between frames in 150cm points-to-camera distance case. Keypoints of C_0 (left) and keypoints of C_1 (right).

Synthetic data		**ARAP**	**Elastic**	**HyperElastic**	**ARAP NG**
Mov. (mm)	D(cm) - t - def(mm)	Av. 3D error (mm)			
$4,02 \pm 0,11$	20 - P - 2.5 g	$\mathbf{3,14 \pm 0,09}$	$3,35 \pm 0,25$	$4,59 \pm 0,13$	$3,37 \pm 0,24$
$2,5 \pm 0$	20 - P - 2.5 r	$\mathbf{1,46 \pm 0,13}$	$1,51 \pm 0,06$	$1,49 \pm 0,09$	$1,49 \pm 0,08$
$4,8 \pm 0,12$	20 - P - 2.5 g 2.5 r	$\mathbf{2,94 \pm 0,18}$	$3,5 \pm 0,12$	$4,86 \pm 0,46$	$3,47 \pm 0,13$
$15,97 \pm 0,66$	20 - P - 10 g	$\mathbf{11,98 \pm 1,57}$	$19,84 \pm 3,34$	$39,75 \pm 3,53$	$14,44 \pm 0,81$
10 ± 0	20 - P - 10 r	$\mathbf{4,22 \pm 0,57}$	$4,76 \pm 0,18$	$4,44 \pm 0,12$	$4,42 \pm 0,44$
$18,5 \pm 0,69$	20 - P - 10 g 10 r	$\mathbf{12,38 \pm 2,37}$	$19,13 \pm 4,5$	$36,84 \pm 0,68$	$14,71 \pm 1,16$
$3,12 \pm 0,11$	20 - G - 2.5 r	$\mathbf{1,39 \pm 0,16}$	$1,64 \pm 0,05$	$1,54 \pm 0,06$	$1,54 \pm 0,04$
$5,29 \pm 0,08$	20 - G - 2.5 g 2.5 r	$\mathbf{3,24 \pm 0,32}$	$3,67 \pm 0,01$	$4,98 \pm 0,22$	$3,62 \pm 0,14$
$10,39 \pm 0,37$	20 - G - 10 r	$\mathbf{3,09 \pm 1,26}$	$4,36 \pm 0,19$	$3,91 \pm 0,15$	$4,13 \pm 0,34$
$19,93 \pm 0,84$	20 - G - 10 g 10 r	$\mathbf{12,26 \pm 1,47}$	$22,39 \pm 4,76$	$43,56 \pm 4,67$	$14,58 \pm 0,99$
$4,02 \pm 0,25$	80 - P - 2.5 g	$\mathbf{9,38 \pm 1,34}$	$11,01 \pm 1,26$	$11,01 \pm 1,26$	$10,95 \pm 1,24$
$2,5 \pm 0$	80 - P - 2.5 r	$\mathbf{7,22 \pm 0,74}$	$8,03 \pm 0,67$	$8,01 \pm 0,67$	$8 \pm 0,67$
$4,69 \pm 0,01$	80 - P - 2.5 g 2.5 r	$\mathbf{8,96 \pm 1,4}$	$10,42 \pm 0,38$	$10,4 \pm 0,37$	$10,41 \pm 0,37$
$15,81 \pm 0,5$	80 - P - 10 g	$\mathbf{23,25 \pm 3,4}$	$28,16 \pm 2,17$	$34,08 \pm 4,12$	$26,89 \pm 1,97$
10 ± 0	80 - P - 10 r	$\mathbf{13,15 \pm 1}$	$14,31 \pm 0,31$	$14,29 \pm 0,29$	$14,21 \pm 0,13$
$18,99 \pm 1,17$	80 - P - 10 g 10 r	$\mathbf{26,82 \pm 8,64}$	$34,76 \pm 2,82$	$39,23 \pm 3,93$	$31,99 \pm 4,57$
$22,48 \pm 1,15$	80 - G - 2.5 r	$\mathbf{24,26 \pm 3,33}$	$27,77 \pm 1,58$	$27,02 \pm 1,34$	$26,83 \pm 1,8$
$23,66 \pm 1,24$	80 - G - 2.5 g 2.5 r	$\mathbf{26,1 \pm 4,43}$	$31,26 \pm 1,48$	$30,97 \pm 1,45$	$30,28 \pm 1,27$
$42,35 \pm 3,48$	80 - G - 10 r	$\mathbf{42,33 \pm 6,96}$	$48,06 \pm 3,38$	$47,39 \pm 3,34$	$47,04 \pm 3,99$
$48,6 \pm 2,82$	80 - G - 10 g 10 r	$\mathbf{46,7 \pm 6,42}$	$57,65 \pm 4,45$	$57,26 \pm 4,42$	$55,37 \pm 4,75$
$4,06 \pm 0,08$	150 - P - 2.5 g	$\mathbf{16,99 \pm 0,98}$	$19,1 \pm 0,83$	$19,1 \pm 0,9$	$18,88 \pm 0,81$
$2,5 \pm 0$	150 - P - 2.5 r	$\mathbf{15,5 \pm 1,61}$	$17,46 \pm 0,68$	$17,43 \pm 0,69$	$17,22 \pm 0,48$
$4,64 \pm 0,14$	150 - P - 2.5 g 2.5 r	$\mathbf{17,71 \pm 1,7}$	$19,77 \pm 0,61$	$19,77 \pm 0,6$	$19,41 \pm 0,62$
$15,89 \pm 0,46$	150 - P - 10 g	$\mathbf{36,52 \pm 5,04}$	$42,06 \pm 1,55$	$42,07 \pm 1,55$	$40,36 \pm 2,89$
10 ± 0	150 - P - 10 r	$\mathbf{19,38 \pm 1,18}$	$20,43 \pm 1,03$	$20,58 \pm 0,66$	$20,46 \pm 0,78$
$19,27 \pm 0,76$	150 - P - 10 g 10 r	$\mathbf{39,31 \pm 5,09}$	$47,04 \pm 3,13$	$47,12 \pm 3,05$	$43,99 \pm 3,86$
$42,13 \pm 3,14$	150 - G - 2.5 r	$\mathbf{60,41 \pm 7,69}$	$62,42 \pm 7,36$	$61,64 \pm 7,72$	$61,64 \pm 7,07$
$43,15 \pm 0,87$	150 - G - 2.5 g 2.5 r	$\mathbf{60,45 \pm 9,95}$	$65,12 \pm 7,93$	$63,82 \pm 7,77$	$64,11 \pm 7,61$
$75,49 \pm 4,85$	150 - G - 10 r	$\mathbf{93,28 \pm 11,79}$	$96,76 \pm 10,44$	$94,95 \pm 10,49$	$95,56 \pm 10,31$
$82,08 \pm 4,99$	150 - G - 10 g 10 r	$\mathbf{102,15 \pm 10,34}$	$111,74 \pm 5,08$	$109,69 \pm 3,87$	$110,71 \pm 5,39$

Table 1: Average 3D errpr and standard desviation of 3 different triangulation results using ARAP, elastic, and hyperelastic models with the proposed global deformation term, and ARAP without it. Points, viewed with significant parallax at camera-to-points distances of 20, 80, and 150 cm, exhibit planar (P) and gradual (G) deformations under movements alternating between Gaussian (g), rigid (r), and a combination of both. Details in section 4.1. Note that the ARAP model stands out as the one offering lower triangulation errors.

Synthetic data		**ARAP**		
		InRays	MidPoints	FarPoints
Mov. (mm)	D(cm) - t - def(mm)	Av. 3D error (mm)		
$4,02 \pm 0,11$	20 - P - 2.5 g	$3,17 \pm 0,1$	$\mathbf{3,14 \pm 0,09}$	$3,22 \pm 0,43$
$2,5 \pm 0$	20 - P - 2.5 r	$\mathbf{1,46 \pm 0,13}$	$1,52 \pm 0,14$	$1,47 \pm 0,12$
$4,8 \pm 0,12$	20 - P - 2.5 g 2.5 r	$3,22 \pm 0,24$	$3,26 \pm 0,25$	$\mathbf{2,94 \pm 0,18}$
$15,97 \pm 0,66$	20 - P - 10 g	$12,04 \pm 1,6$	$\mathbf{11,98 \pm 1,57}$	$14,97 \pm 2,04$
10 ± 0	20 - P - 10 r	$4,52 \pm 0,14$	$4,8 \pm 0,23$	$\mathbf{4,22 \pm 0,57}$
$18,5 \pm 0,69$	20 - P - 10 g 10 r	$12,56 \pm 2,01$	$\mathbf{12,38 \pm 2,37}$	$18,93 \pm 8,26$
$3,12 \pm 0,11$	20 - G - 2.5 r	$1,69 \pm 0,16$	$1,79 \pm 0,12$	$\mathbf{1,39 \pm 0,16}$
$5,29 \pm 0,08$	20 - G - 2.5 g 2.5 r	$3,43 \pm 0,34$	$3,5 \pm 0,22$	$\mathbf{3,24 \pm 0,32}$
$10,39 \pm 0,37$	20 - G - 10 r	$\mathbf{3,09 \pm 1,26}$	$3,66 \pm 0,63$	$3,55 \pm 0,26$
$19,93 \pm 0,84$	20 - G - 10 g 10 r	$\mathbf{12,26 \pm 1,47}$	$12,31 \pm 1,32$	$66,26 \pm 83,74$
$4,02 \pm 0,25$	80 - P - 2.5 g	$9,99 \pm 0,71$	$10,04 \pm 0,71$	$\mathbf{9,38 \pm 1,34}$
$2,5 \pm 0$	80 - P - 2.5 r	$7,66 \pm 1,07$	$7,65 \pm 1,09$	$\mathbf{7,22 \pm 0,74}$
$4,69 \pm 0,01$	80 - P - 2.5 g 2.5 r	$9,71 \pm 1,35$	$9,77 \pm 1,35$	$\mathbf{8,96 \pm 1,4}$
$15,81 \pm 0,5$	80 - P - 10 g	$25,73 \pm 5,23$	$25,77 \pm 5,17$	$\mathbf{23,25 \pm 3,4}$
10 ± 0	80 - P - 10 r	$13,67 \pm 0,76$	$13,77 \pm 0,68$	$\mathbf{13,15 \pm 1}$
$18,99 \pm 1,17$	80 - P - 10 g 10 r	$31,87 \pm 6,97$	$31,98 \pm 6,83$	$\mathbf{26,82 \pm 8,64}$
$22,48 \pm 1,15$	80 - G - 2.5 r	$25,55 \pm 2,51$	$26,17 \pm 2,56$	$\mathbf{24,26 \pm 3,33}$
$23,66 \pm 1,24$	80 - G - 2.5 g 2.5 r	$28,54 \pm 3,41$	$29,32 \pm 2,66$	$\mathbf{26,10 \pm 4,43}$
$42,35 \pm 3,48$	80 - G - 10 r	$45,76 \pm 3,67$	$45,7 \pm 4,07$	$\mathbf{42,33 \pm 6,96}$
$48,6 \pm 2,82$	80 - G - 10 g 10 r	$53,59 \pm 3,37$	$53,37 \pm 3,79$	$\mathbf{46,7 \pm 6,42}$
$4,06 \pm 0,08$	150 - P - 2.5 g	$18,05 \pm 1,71$	$18,06 \pm 1,7$	$\mathbf{16,99 \pm 0,98}$
$2,5 \pm 0$	150 - P - 2.5 r	$16,39 \pm 1,61$	$16,51 \pm 1,6$	$\mathbf{15,50 \pm 1,61}$
$4,64 \pm 0,14$	150 - P - 2.5 g 2.5 r	$18,73 \pm 1,58$	$18,79 \pm 1,58$	$\mathbf{17,71 \pm 1,7}$
$15,89 \pm 0,46$	150 - P - 10 g	$39,28 \pm 3,67$	$39,36 \pm 3,58$	$\mathbf{36,52 \pm 5,04}$
10 ± 0	150 - P - 10 r	$19,92 \pm 0,53$	$20,30 \pm 0,5$	$\mathbf{19,38 \pm 1,18}$
$19,27 \pm 0,76$	150 - P - 10 g 10 r	$43,96 \pm 7,15$	$44,51 \pm 7,03$	$\mathbf{39,31 \pm 5,09}$
$42,13 \pm 3,14$	150 - G - 2.5 r	$61,32 \pm 7,34$	$61,24 \pm 8,1$	$\mathbf{60,41 \pm 7,69}$
$43,15 \pm 0,87$	150 - G - 2.5 g 2.5 r	$62,88 \pm 8,32$	$63,99 \pm 7,86$	$\mathbf{60,45 \pm 9,95}$
$75,49 \pm 4,85$	150 - G - 10 r	$95,4 \pm 11,5$	$95,32 \pm 12,27$	$\mathbf{93,28 \pm 11,79}$
$82,08 \pm 4,99$	150 - G - 10 g 10 r	$106,68 \pm 8$	$107,69 \pm 6,59$	$\mathbf{102,15 \pm 10,34}$

Table 2: Average and standard desviation of 3 different triangulation results using ARAP with the three proposed initial seeds. Points, viewed with significant parallax at camera-to-points distances of 20, 80, and 150 cm, exhibit planar (P) and gradual (G) deformations under movements alternating between Gaussian (g), rigid (r), and a combination of both. Details in section 4.1.

Synthetic data		**ARAP - FarPoints**	
Mov. (mm)	D(cm) - t - def(mm)	Av. 3D error (mm)	
		Initial	Final
$4,02 \pm 0,11$	20 - P - 2.5 g	$\mathbf{3,17 \pm 0,26}$	$3,22 \pm 0,43$
$2,5 \pm 0$	20 - P - 2.5 r	$1,93 \pm 0,14$	$\mathbf{1,47 \pm 0,12}$
$4,8 \pm 0,12$	20 - P - 2.5 g 2.5 r	$3,47 \pm 0,1$	$\mathbf{2,94 \pm 0,18}$
$15,97 \pm 0,66$	20 - P - 10 g	$\mathbf{12,04 \pm 1,6}$	$14,97 \pm 2,04$
10 ± 0	20 - P - 10 r	$5,95 \pm 0,28$	$\mathbf{4,22 \pm 0,57}$
$18,5 \pm 0,69$	20 - P - 10 g 10 r	$\mathbf{13,21 \pm 1,45}$	$18,93 \pm 8,26$
$3,12 \pm 0,11$	20 - G - 2.5 r	$2,22 \pm 0,06$	$\mathbf{1,39 \pm 0,16}$
$5,29 \pm 0,08$	20 - G - 2.5 g 2.5 r	$3,79 \pm 0,25$	$\mathbf{3,24 \pm 0,32}$
$10,39 \pm 0,37$	20 - G - 10 r	$6,07 \pm 0,26$	$\mathbf{3,55 \pm 0,26}$
$19,93 \pm 0,84$	20 - G - 10 g 10 r	$\mathbf{13,73 \pm 0,99}$	$66,26 \pm 83,74$
$4,02 \pm 0,25$	80 - P - 2.5 g	$9,98 \pm 1,33$	$\mathbf{9,38 \pm 1,34}$
$2,5 \pm 0$	80 - P - 2.5 r	$7,71 \pm 0,43$	$\mathbf{7,22 \pm 0,74}$
$4,69 \pm 0,01$	80 - P - 2.5 g 2.5 r	$9,59 \pm 1,44$	$\mathbf{8,96 \pm 1,4}$
$15,81 \pm 0,5$	80 - P - 10 g	$23,48 \pm 3,1$	$\mathbf{23,25 \pm 3,4}$
10 ± 0	80 - P - 10 r	$13,83 \pm 0,89$	$\mathbf{13,15 \pm 1}$
$18,99 \pm 1,17$	80 - P - 10 g 10 r	$30,2 \pm 7,35$	$\mathbf{26,82 \pm 8,64}$
$22,48 \pm 1,15$	80 - G - 2.5 r	$27,67 \pm 3,03$	$\mathbf{24,26 \pm 3,33}$
$23,66 \pm 1,24$	80 - G - 2.5 g 2.5 r	$30,02 \pm 4,02$	$\mathbf{26,1 \pm 4,43}$
$42,35 \pm 3,48$	80 - G - 10 r	$48,6 \pm 5,92$	$\mathbf{42,33 \pm 6,96}$
$48,6 \pm 2,82$	80 - G - 10 g 10 r	$56,38 \pm 6,24$	$\mathbf{46,7 \pm 6,42}$
$4,06 \pm 0,08$	150 - P - 2.5 g	$18,02 \pm 0,92$	$\mathbf{16,99 \pm 0,98}$
$2,5 \pm 0$	150 - P - 2.5 r	$16,81 \pm 1,88$	$\mathbf{15,5 \pm 1,61}$
$4,64 \pm 0,14$	150 - P - 2.5 g 2.5 r	$18,81 \pm 2,08$	$\mathbf{17,71 \pm 1,7}$
$15,89 \pm 0,46$	150 - P - 10 g	$37,51 \pm 4,87$	$\mathbf{36,52 \pm 5,04}$
10 ± 0	150 - P - 10 r	$20,51 \pm 1,35$	$\mathbf{19,38 \pm 1,18}$
$19,27 \pm 0,76$	150 - P - 10 g 10 r	$42,21 \pm 4,44$	$\mathbf{39,31 \pm 5,09}$
$42,13 \pm 3,14$	150 - G - 2.5 r	$64,69 \pm 8,11$	$\mathbf{60,41 \pm 7,69}$
$43,15 \pm 0,87$	150 - G - 2.5 g 2.5 r	$67,23 \pm 8,39$	$\mathbf{60,45 \pm 9,95}$
$75,49 \pm 4,85$	150 - G - 10 r	$100,83 \pm 13,41$	$\mathbf{93,28 \pm 11,79}$
$82,08 \pm 4,99$	150 - G - 10 g 10 r	$113,16 \pm 10,11$	$\mathbf{102,15 \pm 10,34}$

Table 3: Average and standard desviation of the initial and final average 3D errors over 3 different experiments using ARAP with FarPoints initial seed. Points, viewed with significant parallax at camera-to-points distances of 20, 80, and 150 cm, exhibit planar (P) and gradual (G) deformations under movements alternating between Gaussian (g), rigid (r), and a combination of both. Details in section 4.1.

To further characterize potential triangulation methods, we explored scenarios where depth information is available through additional sensors (metric detph) or deep neural networks capable of predicting up-to-scale depth measurements for single views. As an initial investigation, Table 4 presents the average 3D errors at 20 cm camera-to-point distances before and after applying our optimization technique to scaled depth measurements with varying measurement uncertainties (1mm, 3mm, and 8mm).

We can appreciate various interesting insights. With highly accurate depth measurements (1mm), the average errors remain minimal relative to movement, and the optimization provides little benefits (or none). The upgrading process holds particularly true for scenarios involving purely rigid deformations, but the optimization degrades even in purely rigid gradual (G) shape transformations.

In contrast, in scenarios with moderate and high depth uncertainties (3 mm and 8 mm), the optimization process sometimes reduces the initial errors, particularly in cases of purely rigid deformations. An exception is noted in the scenario $P - 2.5g2.5r$ with 8 mm uncertainty, but the high standard deviation across experiments indicates instances of deterioration. Summing up, when accurate metric depth is available, the problem is fully observable and our proposed optimization, enforcing assumptions that may not hold, worsens the results. However, as the depth measurements become noisier, our proposed optimization may improve in some cases the final triangulation errors. Note in any case, that having depth sensors is not feasible in all situations. The next case that we analyze, in which up-to-scale depth information comes from the prediction of a deep network, is more relevant as it operates on the input of a RGB image.

Synthetic - Depth 20 cm		**ARAP**					
		Depth uncertainty 1 mm		Depth uncertainty 3 mm		Depth uncertainty 8 mm	
Mov. (mm)	t - def(mm)	Av. 3D error (mm)		Av. 3D error (mm)		Av. 3D error (mm)	
		Initial	Final	Initial	Final	Initial	Final
$4,02 \pm 0,11$	P - 2.5 g	$\mathbf{1,08 \pm 0,01}$	$3,27 \pm 0,28$	$\mathbf{2,56 \pm 0,02}$	$3,35 \pm 0,31$	$\mathbf{6,5 \pm 0,04}$	$7,63 \pm 3,31$
$4,02 \pm 0,11$	P - 2.5 g	$\mathbf{1,08 \pm 0,01}$	$3,27 \pm 0,28$	$\mathbf{2,56 \pm 0,02}$	$3,35 \pm 0,31$	$\mathbf{6,5 \pm 0,04}$	$7,63 \pm 3,31$
$2,5 \pm 0$	P - 2.5 r	$1,09 \pm 0,02$	$\mathbf{0,86 \pm 0,07}$	$2,58 \pm 0,04$	$\mathbf{1,39 \pm 0,05}$	$6,55 \pm 0,12$	$\mathbf{2,58} \pm 1,1$
$4,8 \pm 0,12$	P - 2.5 g 2.5 r	$\mathbf{1,07 \pm 0,01}$	$3,35 \pm 0,3$	$\mathbf{2,6 \pm 0,04}$	$3,52 \pm 0,42$	$6,64 \pm 0,3$	$\mathbf{6,51 \pm 2,8}$
$15,97 \pm 0,66$	P - 10 g	$\mathbf{1,08 \pm 0,01}$	$18,15 \pm 0,94$	$\mathbf{2,56} \pm 0,02$	$25,56 \pm 16,63$	$\mathbf{6,49 \pm 0,04}$	$41,86 \pm 22,77$
10 ± 0	P - 10 r	$1,06 \pm 0,01$	$\mathbf{0,71 \pm 0,05}$	$2,55 \pm 0,03$	$\mathbf{0,9 \pm 0,07}$	$6,53 \pm 0,08$	$\mathbf{4,32 \pm 0,89}$
$18,5 \pm 0,69$	P - 10 g 10 r	$\mathbf{1,06 \pm 0,01}$	$24,76 \pm 8,71$	$\mathbf{2,54 \pm 0,01}$	$74,4 \pm 38,81$	$\mathbf{6,5} \pm 0,01$	$70,67 \pm 53,26$
$3,12 \pm 0,11$	G - 2.5 r	$\mathbf{1,06 \pm 0,01}$	$1,19 \pm 0,11$	$2,54 \pm 0,01$	$\mathbf{1,67 \pm 0,4}$	$6,49 \pm 0,01$	$\mathbf{3,81 \pm 1,99}$
$5,29 \pm 0,08$	G - 2.5 g 2.5 r	$\mathbf{1,07 \pm 0,01}$	$3,66 \pm 0,35$	$\mathbf{2,56 \pm 0,02}$	$4,53 \pm 1,11$	$\mathbf{6,53 \pm 0,06}$	$8,17 \pm 0,12$
$10,39 \pm 0,37$	G - 10 r	$\mathbf{1,06 \pm 0,02}$	$2,24 \pm 0,31$	$2,56 \pm 0,02$	$\mathbf{2,21 \pm 0,28}$	$6,54 \pm 0,04$	$\mathbf{5,46 \pm 0,96}$
$19,93 \pm 0,84$	G - 10 g 10 r	$\mathbf{1,05 \pm 0,03}$	$30,64 \pm 11,57$	$\mathbf{2,58 \pm 0,04}$	$46,79 \pm 28,20$	$\mathbf{6,61 \pm 0,08}$	$70,8 \pm 31,85$

Table 4: Average and standard desviation of the initial and final average 3D errors over 3 different experiments using ARAP optimization and 3 different uncertainties over depth measurements used in the initial seed and the optimization. Points, viewed with significant parallax at distances of 20cm from the cameras, exhibit planar (P) and gradual (G) deformations under movements alternating between Gaussian (g), rigid (r), and a combination of both. Details in section 4.1. Observe how, obviously, our deformation models worsen in general the performance of directly triangulating with the depth data, particularly for small depth uncertainties.

Using the middlepoint scenario of 3mm of depth uncertainty, we compare three scenarios: ARAP with no depth measurements (from Table 3), ARAP with depth measurements with unknown scale, and depth measurements with known scale (from Table 4).

These comparisons are summarized in Table 5. From these results, we observe that, when using depth measurements up to scale, we must initialize the optimization process similarly to the case where no depth information is available. Therefore, the initial errors for the first two approaches exhibit the same values. However, we can appreciate that incorporating depth information with optimizable scale values into the optimization process leads to more accurate final errors in almost all cases compared to using no depth information. Finally, by comparing with the last column, we can conclude that, although the method achieves a good performance in both cases up-to-scale and known-scale depth measurements, the initial errors are significantly different, making the known-scale case the most favorable.

Synthetic - Depth 20 cm		**ARAP - Depth uncertainty 3 mm**					
		No depth		Depth up to scale		Depth	
Mov. (mm)	t - def(mm)	Av. 3D error (mm)		Av. 3D error (mm)		Av. 3D error (mm)	
		Initial	Final	Initial	Final	Initial	Final
$4,02 \pm 0,11$	P - 2.5 g	$\mathbf{3,17 \pm 0,26}$	$3,22 \pm 0,43$	$3,17 \pm 0,26$	$\mathbf{3,12 \pm 0,19}$	$\mathbf{2,56 \pm 0,02}$	$3,35 \pm 0,31$
$2,5 \pm 0$	P - 2.5 r	$1,93 \pm 0,14$	$\mathbf{1,47 \pm 0,12}$	$1,93 \pm 0,14$	$\mathbf{1,45 \pm 0,11}$	$2,58 \pm 0,04$	$\mathbf{1,39 \pm 0,05}$
$4,8 \pm 0,12$	P - 2.5 g 2.5 r	$3,47 \pm 0,1$	$\mathbf{2,94 \pm 0,18}$	$3,47 \pm 0,1$	$\mathbf{2,88 \pm 0,28}$	$\mathbf{2,6 \pm 0,1}$	$3,52 \pm 0,42$
$15,97 \pm 0,66$	P - 10 g	$\mathbf{12,04 \pm 1,6}$	$14,97 \pm 2,04$	$12,04 \pm 1,6$	$\mathbf{8,06 \pm 3,16}$	$\mathbf{2,56 \pm 0,02}$	$25,56 \pm 16,63$
10 ± 0	P - 10 r	$5,95 \pm 0,28$	$\mathbf{4,22 \pm 0,57}$	$5,95 \pm 0,28$	$\mathbf{3,28 \pm 0,38}$	$2,55 \pm 0,03$	$\mathbf{0,9 \pm 0,07}$
$18,5 \pm 0,69$	P - 10 g 10 r	$\mathbf{13,21 \pm 1,45}$	$18,93 \pm 8,26$	$13,21 \pm 1,45$	$\mathbf{8,57 \pm 4,69}$	$\mathbf{2,54 \pm 0,01}$	$74,40 \pm 38,81$
$3,12 \pm 0,11$	G - 2.5 r	$2,22 \pm 0,06$	$\mathbf{1,39 \pm 0,16}$	$2,22 \pm 0,06$	$\mathbf{1,29 \pm 0,23}$	$2,54 \pm 0,01$	$\mathbf{1,67 \pm 0,4}$
$5,29 \pm 0,08$	G - 2.5 g 2.5 r	$3,79 \pm 0,25$	$\mathbf{3,24 \pm 0,32}$	$3,79 \pm 0,25$	$\mathbf{3,02 \pm 0,08}$	$\mathbf{2,56 \pm 0,02}$	$4,53 \pm 1,11$
$10,39 \pm 0,37$	G - 10 r	$6,07 \pm 0,26$	$\mathbf{3,55 \pm 0,26}$	$6,07 \pm 0,26$	$\mathbf{1,9 \pm 0,48}$	$2,56 \pm 0,02$	$\mathbf{2,21 \pm 0,28}$
$19,93 \pm 0,84$	G - 10 g 10 r	$\mathbf{13,73 \pm 0,99}$	$66,26 \pm 83,74$	$13,73 \pm 0,99$	$\mathbf{11,09 \pm 0,22}$	$\mathbf{2,58 \pm 0,04}$	$46,79 \pm 28,20$

Table 5: Average and standard desviation of the initial and final triangulation errors over 3 different experiments using: ARAP as in Table 3 without depth information, ARAP with depth measurements up to scale, and ARAP with metric scale depth as in Table 4. Points, viewed with significant parallax at distances of 20cm from the cameras, exhibit planar (P) and gradual (G) deformations under movements alternating between Gaussian (g), rigid (r), and a combination of both. Details in section 4.1. Observe how, for this high parallax experiments, having metric depth and up-to-scale depth measurements offers a certain but not very big improvement, as parallax is a bigger source of information.

4.3.2 Real data

To validate our findings from synthetic data experiments in realistic scenarios, we evaluated our method on real images. These scenarios introduce challenges such as imperfect feature matching, data association challenges, outliers, and low parallax.

Table 6 presents results on four image pairs extracted from different segments of the Drunkard dataset [31]. The first two pairs, $320\_0\_1975-1983$ and $320\_0\_2500-2513$, have a resolution of 320x320 pixels, they come from scene 0, and they are extracted between frames 1975-1983 and 2500-2513, respectively. The other two pairs, $1024\_1\_110-120$ and $1024\_0\_1229-1236$, have a resolution of 1024x1024 pixels, they come from scenes 1 and 0, respectively, and they are extracted between frames 110-120 and 1229-1236, respectively.

These scene segments were chosen to include sufficient camera translation and objects with high contrast, mitigating issues related to outliers, low parallax, and insufficient keypoints.

In scenarios involving deformations, outlier detection using the Essential matrix with RANSAC and reprojection error filtering becomes unreliable. To evaluate our method without these factors influencing the results, Table 6 compares the performance with and without these checks. While both approaches yield a similar number of triangulated points, the case without checks exhibits larger errors in camera translation estimation.

Given the low number of outliers observed in Table 6 and their minimal impact on the overall error, we removed the outlier checks from our implementation to handle deformable situations where these checks are no longer valid. Table 7 presents the results of applying

Real - Rigid scenes		**ARAP - FarPoints**		
Pair	Checks	Av. 3D error normalized by $\|t\|$ (%)	Parallax (^{o})	nPs
320_0_1975-1983	checks	$8,22 \pm 0$	$0,77 \pm 0$	410 ± 0
	no checks	$8,29 \pm 0$	$0,77 \pm 0$	412 ± 0
320_0_2500-2513	checks	$2,52 \pm 0$	$1,69 \pm 0$	272 ± 0
	no checks	$3,14 \pm 0$	$1,71 \pm 0$	286 ± 0
1024_1_110-120	checks	$2,67 \pm 0,01$	$0,69 \pm 0$	185 ± 2
	no checks	$2,88 \pm 0,01$	$0,74 \pm 0$	189 ± 3
1024_0_1229-1236	checks	$0,29 \pm 0$	$1,96 \pm 0$	215 ± 1
	no checks	$0,46 \pm 0$	2 ± 0	218 ± 1

Table 6: Average and standard desviation of triangulation 3D errors normalized by the translation norm between camera poses over 5 different experiments. Pair of frames extracted from Drunkard dataset: each pair has the format "resolution_scene_framesZone". Checks involve verification with the Essential matrix and RANSAC and also verifications of reprojection error. From this table, we observe that outliers may cause problem in our triangulation methods. However, in most cases the effect is limited, so we leave outlier detection in non-rigid triangulation for future work.

our method to real image pairs with varying levels of deformation.

Lee et al. [18] demonstrated that for parallax values between 0.5 and 2.5, their triangulation implementation achieves an average 3D error normalized by the camera translation vector between 25% and 5%. Our rigid cases (level0) consistently meet this expectation, with notably low error percentages in the 1024-resolution case and high-parallax pairs.

In cases with different levels of deformation, our method consistently improves upon the initial error percentages. Notably, the largest improvements are observed in cases with initially high error values. Furthermore, the method does not degrade dramatically the initial solution in rigid cases (level0), in accordance with the magnitude of the errors in each case.

Real deformed		**ARAP - FarPoints**			
Pair	Level	Av. 3D error normalized by $\|t\|$ (%)		Parallax (^{o})	nPs
		Initial	Final		
320_0_1975-1983	level0	$\mathbf{8{,}29 \pm 0}$	$9,6 \pm 0$	$0,77 \pm 0$	412 ± 0
	level1	$21,46 \pm 0$	$\mathbf{20{,}76 \pm 0}$	$0,69 \pm 0$	330 ± 0
	level2	$16,16 \pm 0$	$\mathbf{15{,}42 \pm 0}$	$0,62 \pm 0$	142 ± 0
	level3	$26,18 \pm 0$	$\mathbf{24{,}07 \pm 0}$	$0,66 \pm 0$	248 ± 0
320_0_2500-2513	level0	$3,14 \pm 0$	$\mathbf{2{,}95 \pm 0.01}$	$1,71 \pm 0$	286 ± 0
	level1	$2,9 \pm 0$	$\mathbf{2{,}75 \pm 0}$	$1,62 \pm 0$	340 ± 0
	level2	$\mathbf{13{,}61 \pm 0.07}$	$13,7 \pm 0,01$	$1,71 \pm 0,01$	234 ± 1
	level3	$3,06 \pm 0,07$	$\mathbf{3 \pm 0{,}07}$	$1,77 \pm 0,05$	90 ± 4
1024_1_110-120	level0	$2,88 \pm 0$	$2,88 \pm 0$	$0,74 \pm 0$	186 ± 0
	level1	$5,86 \pm 0,06$	$\mathbf{5{,}85 \pm 0{,}06}$	$0,8 \pm 0$	197 ± 1
	level2	$8,68 \pm 0,14$	$\mathbf{8{,}65 \pm 0{,}11}$	$0.64 \pm 0,01$	147 ± 1
	level3	$15,64 \pm 0$	$\mathbf{15{,}24 \pm 0{,}22}$	0.79 ± 0	148 ± 0
1024_0_1229-1236	level0	$\mathbf{0{,}46 \pm 0}$	$0,47 \pm 0.01$	2 ± 0	219 ± 1
	level1	$1,94 \pm 0$	$\mathbf{1{,}93 \pm 0}$	$2,05 \pm 0$	168 ± 0
	level2	$4,43 \pm 0,09$	$\mathbf{4{,}42 \pm 0{,}09}$	$2,06 \pm 0,01$	161 ± 1
	level3	$3,89 \pm 0$	$\mathbf{3{,}78 \pm 0{,}13}$	$2,72 \pm 0$	230 ± 0

Table 7: Average and standard desviation of initial average 3D errors normalized by the translation norm between cameras poses over 5 different experiments. Pair of frames extracted from Drunkard dataset: each pair has the format "resolution_scene_framesZone". Level involve case of deformation going from rigid level0 to level3.

In addition, we tested real-world scenarios that were previously evaluated in simulation where depth information is available but with an unknown scale. As shown in Table 8, our method proves effective in these scenarios, as expected from simulation results, consistently reducing initial errors in most cases, particularly in those with significant deformations. Furthermore, we conclude that up-to-scale depth information is a crucial factor, significantly improving results from Table 7. Some reconstructions results, from matches observed in Figures 9 and 12, are represented in Figure 14.

Finally, to conclude our experiments, Figure 15 compiles various execution times of the entire process that yields these results. This includes multiple iterations of the external loop that optimizes ϕ_d to balance the weight of the ARAP minimization, as well as multiple iterations of our Levenberg-Marquardt optimization process proposed in this work. From these values, we can appreciate that the number of triangulated points is not a critical factor, as our method leverages sparsity. It is important to note that these execution times can be improved in future work by exploring various possibilities, such as

calculating the Jacobians analytically instead of numerically, or developing an alternative process for obtaining ϕ_d.

Real		**ARAP with depth up to scale - FarPoints**			
Pair	Level	Av. 3D error normalized by $\|t\|$ (%)		Parallax (^{o})	nPs
		Initial	Final		
320_0_1975-1983	level0	**8,29 ± 0**	$42,12 \pm 0,88$	$0,77 \pm 0$	412 ± 0
	level1	$21,46 \pm 0$	**5,04 ± 0**	$0,69 \pm 0$	330 ± 0
	level2	$16,77 \pm 0$	**16,62 ± 0**	$0,62 \pm 0$	138 ± 0
	level3	$26,19 \pm 0$	**10,54 ± 1,98**	$0,66 \pm 0$	245 ± 2
320_0_2500-2513	level0	$3,14 \pm 0$	**1,03 ± 0**	$1,71 \pm 0$	286 ± 0
	level1	$2,9 \pm 0$	**2,13 ± 0**	$1,62 \pm 0$	340 ± 0
	level2	$13,59 \pm 0,06$	**12,93 ± 0,33**	$1,71 \pm 0,01$	234 ± 0
	level3	$3,08 \pm 0,13$	**2,44 ± 0,14**	$1,78 \pm 0,07$	91 ± 4
1024_1_110-120	level0	$2,89 \pm 0,01$	**1,33 ± 1,13**	$0,74 \pm 0$	188 ± 2
	level1	$5,85 \pm 0,05$	**3,61 ± 1,17**	$0,8 \pm 0$	198 ± 1
	level2	$8,72 \pm 0,2$	**6,92 ± 2,2**	$0,62 \pm 0,02$	145 ± 2
	level3	$15,64 \pm 0$	**12,68 ± 3,79**	$0,79 \pm 0$	148 ± 0
1024_0_1229-1236	level0	$0,46 \pm 0$	**0,33 ± 0,03**	2 ± 0	218 ± 1
	level1	$1,11 \pm 0$	**0,23 ± 0**	$2,05 \pm 0$	152 ± 0
	level2	$4,17 \pm 0,12$	**4,02 ± 0,23**	$2,05 \pm 0,01$	152 ± 1
	level3	$4,36 \pm 0,01$	**3,6 ± 1,04**	$1,98 \pm 0$	229 ± 1

Table 8: Average and standard desviation of initial average 3D errors normalized by the translation norm between cameras poses over 5 different experiments using depth maps without scale. Our optimization improves the initial triangulation guess. Pair of frames extracted from Drunkard's dataset: each pair has the format "resolution_scene_framesZone". Level involve case of deformation going from rigid level0 to level3.

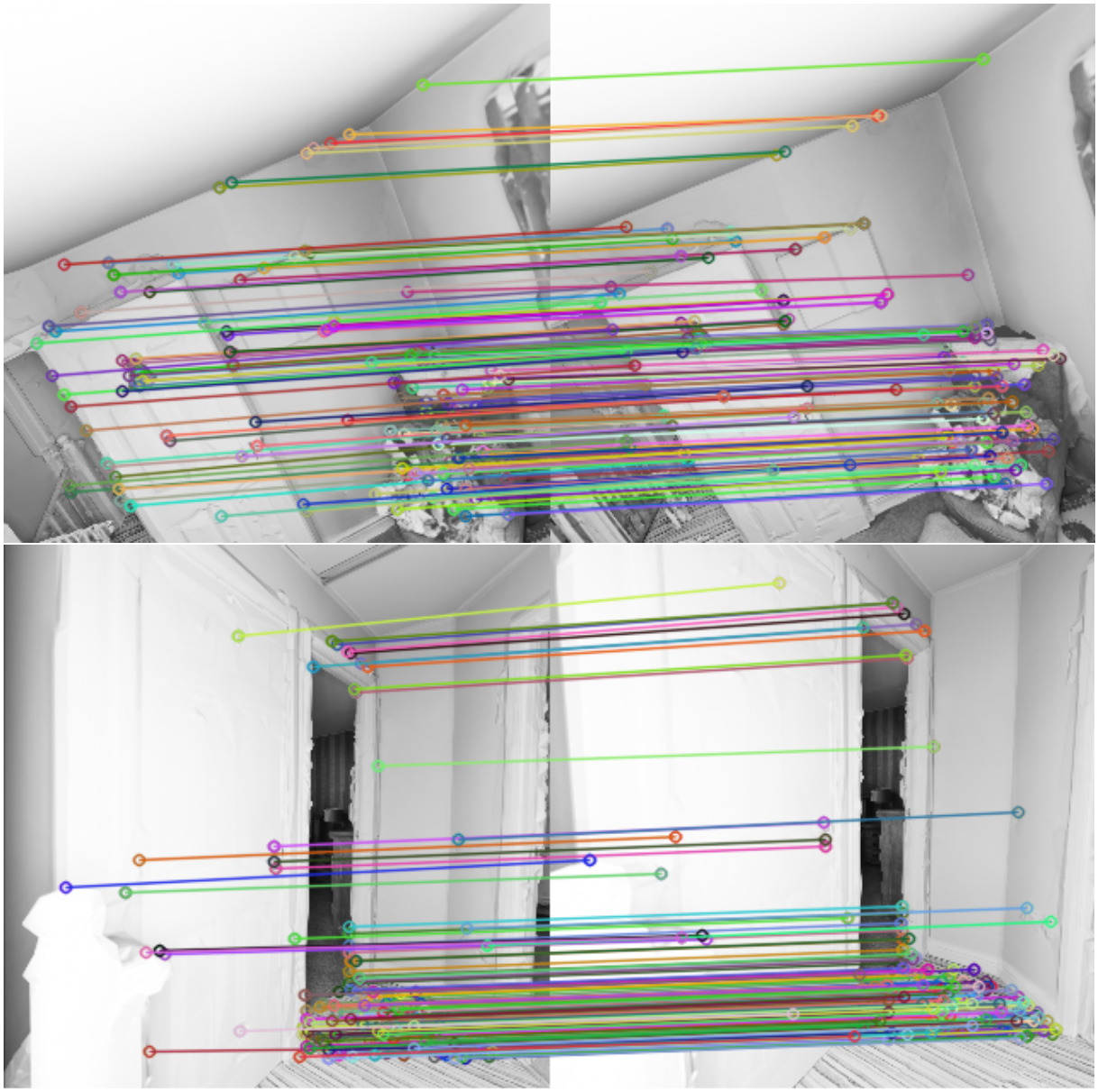

Figure 9: Matches between frames in 320x320 pixels resolution in pairs 1975-1983 (up) and 2500-2513 (bottom).

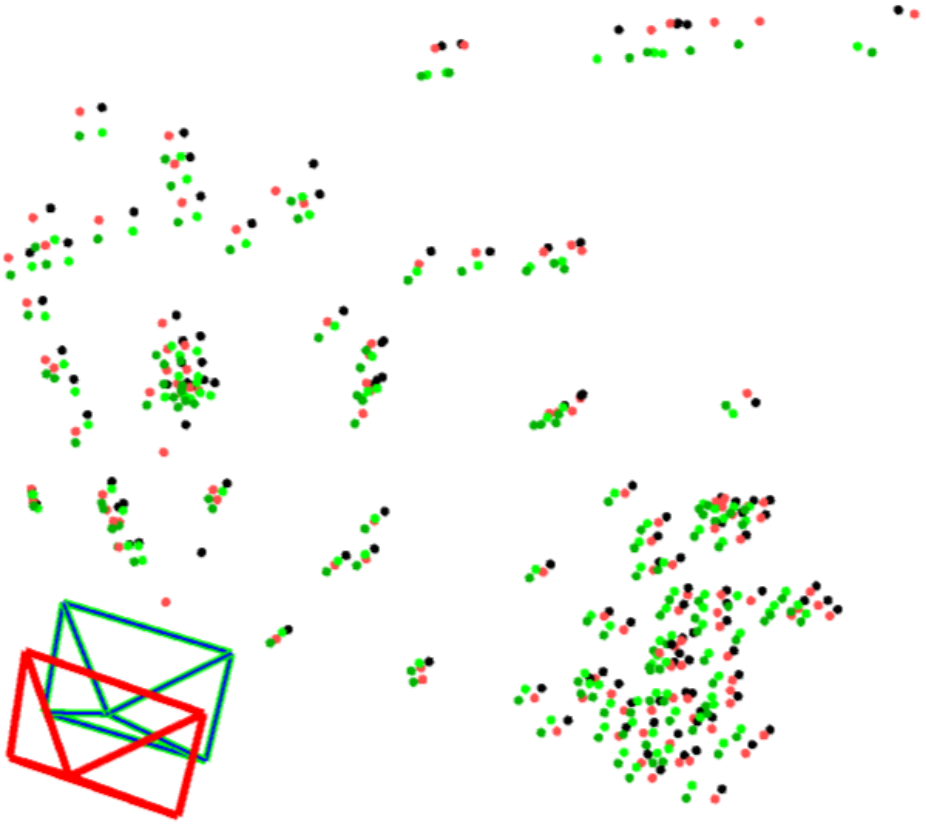

Figure 10: Triangulation results using up-to-scale depth information at level3 level of deformation for view pair 1975-1983. Both camera poses are represented by green and red pyramids. Red points correspond to 3D points at $k = 0$, black points correspond to points observed from $k = 1$, dark green points correspond to our solution for $k = 0$ and light green correspond to our solution for $k = 1$.

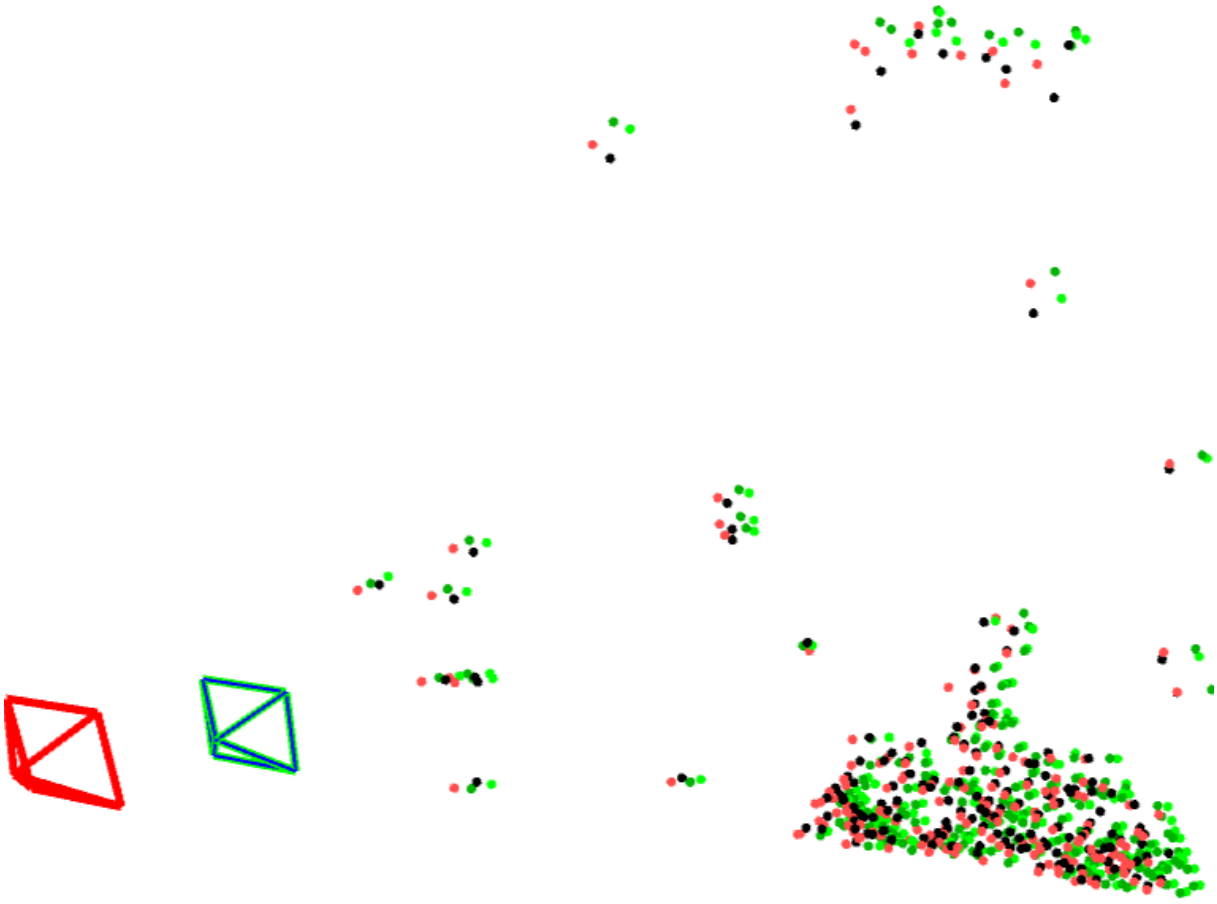

Figure 11: Triangulation results using up-to-scale depth information at level1 level of deformations (level1, level2 and level3) for view pair 2500-2513. Both camera poses are represented by green and red pyramids. Red points correspond to 3D points at $k = 0$, black points correspond to points observed from $k = 1$, dark green points correspond to our solution for $k = 0$ and light green correspond to our solution for $k = 1$.

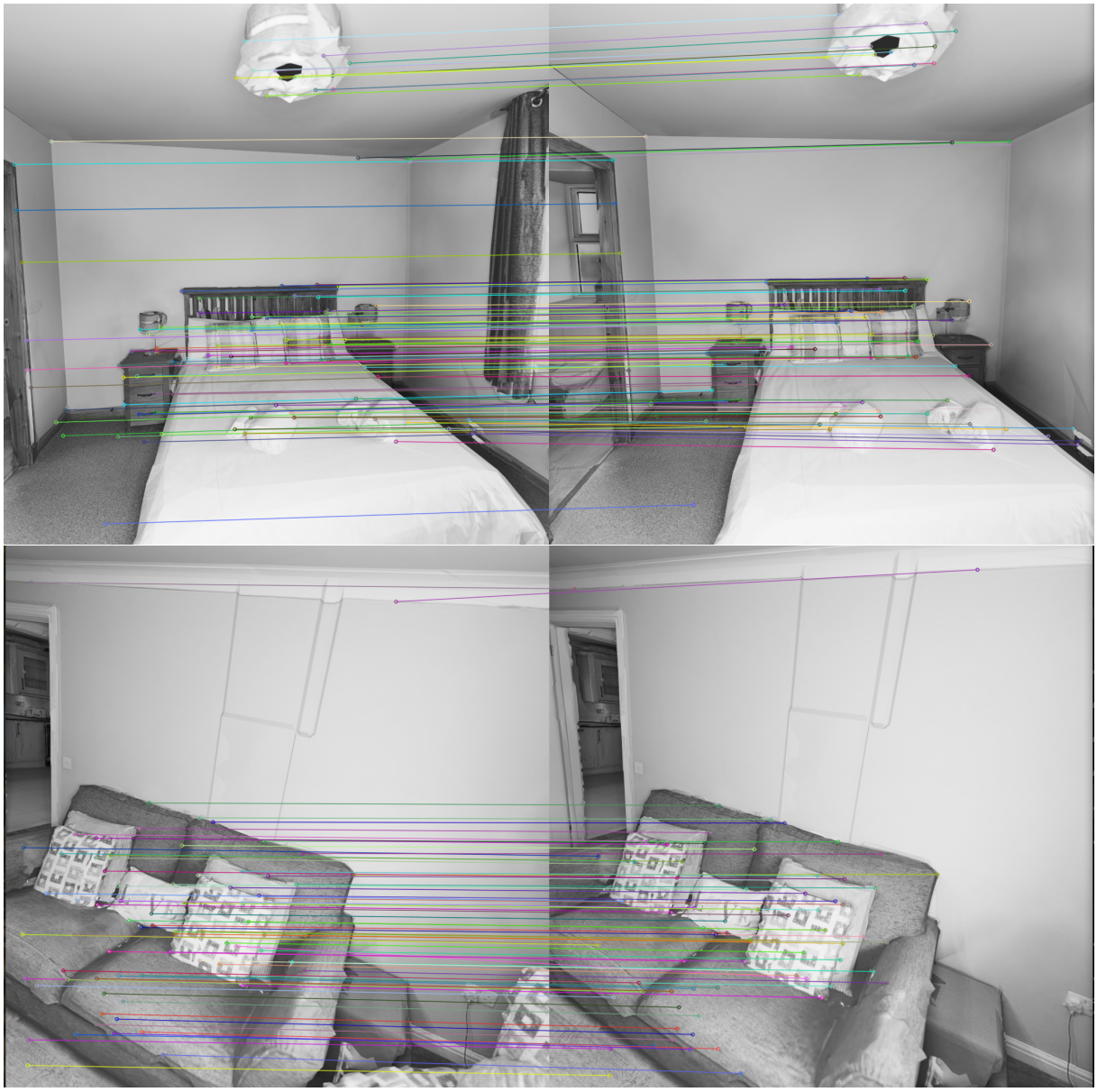

Figure 12: Matches between frames in 1024x1024 pixels resolution in pairs 110-120 (up) and 1229-1236 (bottom).

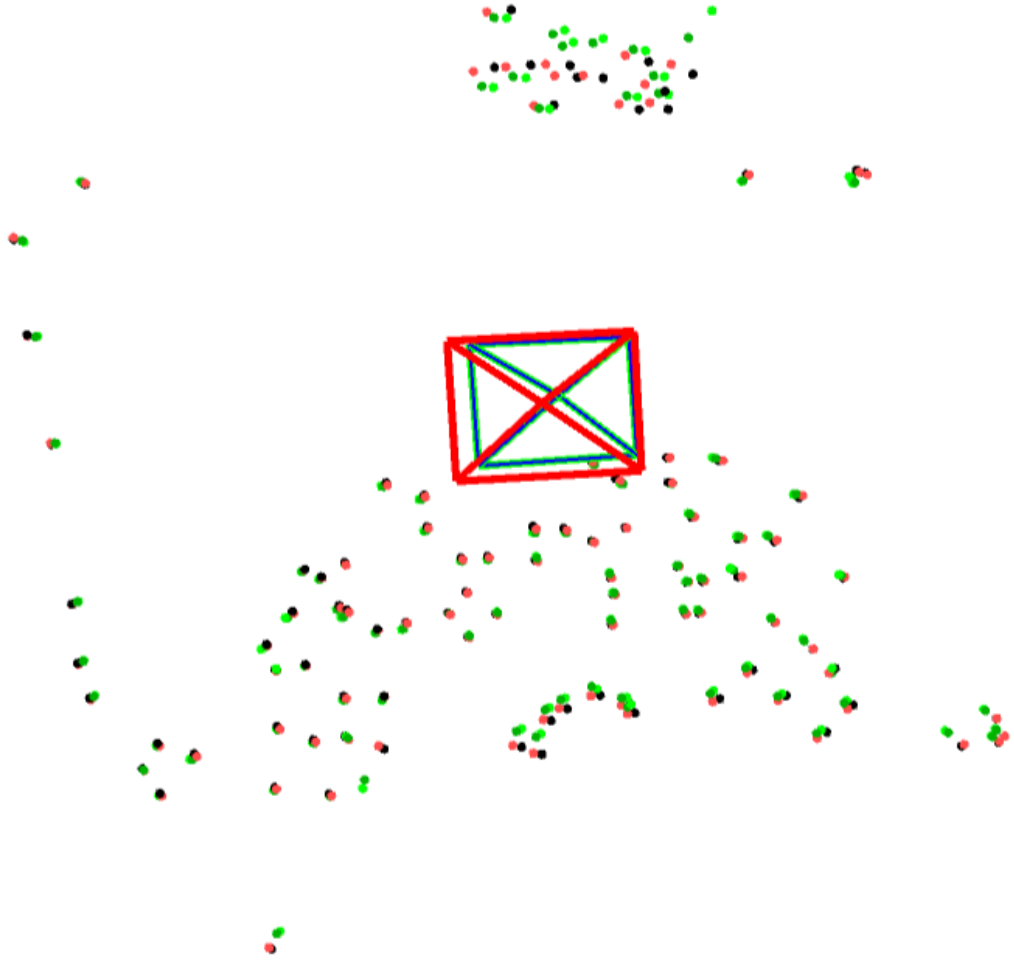

Figure 13: Triangulation results using up-to-scale depth information at level2 level of deformation for view pair 110-120. Both camera poses are represented by green and red pyramids. Red points correspond to 3D points at $k = 0$, black points correspond to points observed from $k = 1$, dark green points correspond to our solution for $k = 0$ and light green correspond to our solution for $k = 1$.

Figure 14: Triangulation results using up-to-scale depth information at level1 level of deformation for view pair 1229-1236. Both camera poses are represented by green and red pyramids. Red points correspond to 3D points at $k = 0$, black points correspond to points observed from $k = 1$, dark green points correspond to our solution for $k = 0$ and light green correspond to our solution for $k = 1$.

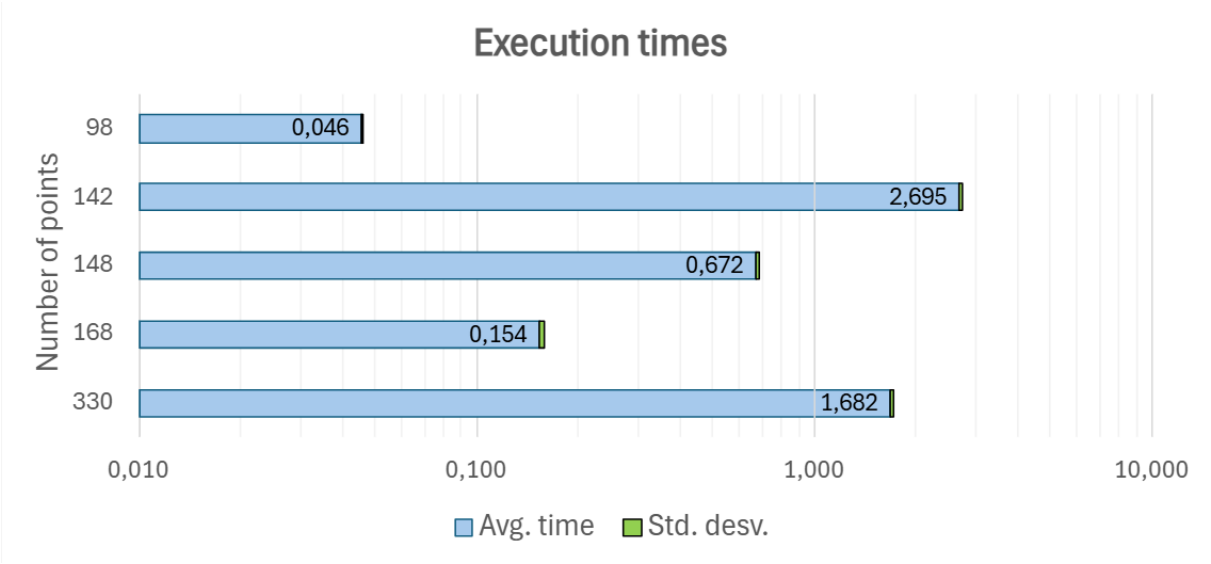

Figure 15: Average and standar desviation of 5 experiments execution times when applying the method proposed through various iterations, including the external loop that balance the weight of the deformation model ϕ_d. Note highly variable convergence times with the number of points, but small variance between runs of the same experiments, suggesting that the convergence time depends on the specific setup but is quite stable.

5 Conclusions

Deformable scenes, in general, remain a challenging problem in computer vision, with several specific open problems. Triangulation, a well studied problem in rigid scenes and with many methods available, has not been addressed from a generic perspective for non-rigid environments. In this thesis we target we addressed such problem by formulating joint energy functions related to the reprojection error and scene deformation models in two views with known intrinsics and extrinsics. Among the different deformation models we evaluated, the As-Rigid-As-Possible one has emerged as the most suitable option for the data considered. This method has not only demonstrated improvements from its initial guesses for multiple deformation cases, but it has also shown enhanced performance when incorporating up-to-scale depth predictions, which could be reasonably obtained by single-view depth neural networks.

Our experiments suggest that the method proposed is a promising technique that successfully refines initial seeds to yield more accurate 3D reconstructions of deformable scenes, while maintaining stability in rigid scenes as well. Although aleatoric movement of points is considered noise, as it does not adhere to the model pattern, the method remains accurate when dealing with known rigid deformations, particularly when up-to-scale depth is integrated. This is promising, as planar and gradual rigid deformations are common in many use cases and real-world scenarios.

This work has been partially funded by the Grants and Scholarships Programme of the Aragon Institute for Engineering Research (I3A) and by the Cátedra XX.

References

[1] Pablo Azagra et al. “Endomapper dataset of complete calibrated endoscopy procedures”. In: *Scientific Data* 10.1 (2023), p. 671 (cit. on p. 4).

[2] Bale Baidi Blaise, Gambo Betchewe, and Tibi Beda. “Optimization of the model of Ogden energy by the genetic algorithm method”. In: *Applied Rheology* 29.1 (2019), pp. 21–29 (cit. on pp. 7, 9).

[3] Aleksei Bochkovskii et al. “Depth pro: Sharp monocular metric depth in less than a second”. In: *arXiv preprint arXiv:2410.02073* (2024) (cit. on p. 4).

[4] Cesar Cadena et al. “Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age”. In: *IEEE Transactions on robotics* 32.6 (2016), pp. 1309–1332 (cit. on p. 3).

[5] Carlos Campos et al. “Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam”. In: *IEEE Transactions on Robotics* 37.6 (2021), pp. 1874–1890 (cit. on p. 4).

[6] Ajad Chhatkuli, Daniel Pizarro, and Adrien Bartoli. “Non-Rigid Shape-from-Motion for Isometric Surfaces using Infinitesimal Planarity.” In: *BMVC*. 2014 (cit. on p. 4).

[7] Ajad Chhatkuli et al. “Inextensible non-rigid shape-from-motion by second-order cone programming”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1719–1727 (cit. on p. 4).

[8] Yasutaka Furukawa, Carlos Hernández, et al. “Multi-view stereo: A tutorial”. In: *Foundations and Trends® in Computer Graphics and Vision* 9.1-2 (2015), pp. 1–148 (cit. on p. 3).

[9] Richard Hartley and Frederik Schaffalitzky. “L/sub/spl infin//minimization in geometric reconstruction problems”. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* Vol. 1. IEEE. 2004, pp. I–I (cit. on pp. 3, 4).

[10] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003 (cit. on p. 3).

[11] Richard Hartley et al. “Verifying global minima for L 2 minimization problems in multiple view geometry”. In: *International Journal of Computer Vision* 101 (2013), pp. 288–304 (cit. on p. 3).

[12] Richard I Hartley and Peter Sturm. “Triangulation”. In: *Computer vision and image understanding* 68.2 (1997), pp. 146–157 (cit. on pp. 3, 4).

[13] Petr Hruby, Marc Pollefeys, and Daniel Barath. “Semicalibrated Relative Pose from an Affine Correspondence and Monodepth”. In: *European Conference on Computer Vision*. Springer. 2025, pp. 39–57 (cit. on p. 4).

[14] Sebastian Hoppe Nesgaard Jensen et al. “A benchmark and evaluation of non-rigid structure from motion”. In: *International Journal of Computer Vision* 129.4 (2021), pp. 882–899 (cit. on p. 4).

[15] Kenichi Kanatani, Yasuyuki Sugaya, and Hirotaka Niitsuma. "Triangulation from two views revisited: Hartley-Sturm vs. optimal correction". In: *practice* 4.5 (2008), p. 99 (cit. on p. 3).

[16] Daniel Keysers et al. "Deformation models for image recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.8 (2007), pp. 1422–1435 (cit. on p. 7).

[17] Jesús Lamarca et al. "Defslam: Tracking and mapping of deforming scenes from monocular sequences". In: *IEEE Transactions on Robotics* 37.1 (2020), pp. 291–303 (cit. on pp. 5, 7).

[18] Seong Hun Lee and Javier Civera. "Triangulation: why optimize?" In: *arXiv preprint arXiv:1907.11917* (2019) (cit. on pp. 9, 10, 27).

[19] Seunghwan Lee and Javier Civera. "Closed-form optimal two-view triangulation based on angular errors". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 2681–2689 (cit. on pp. 3, 4).

[20] Hao Li et al. "3D self-portraits". In: *ACM Transactions on Graphics (TOG)* 32.6 (2013), pp. 1–9 (cit. on p. 4).

[21] Hao Li et al. "Temporally coherent completion of dynamic shapes". In: *ACM Transactions on Graphics (TOG)* 31.1 (2012), pp. 1–11 (cit. on p. 4).

[22] Peter Lindstrom. "Triangulation made easy". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, pp. 1554–1561 (cit. on pp. 3, 4).

[23] Melvin Mooney. "A theory of large elastic deformation". In: *Journal of applied physics* 11.9 (1940), pp. 582–592 (cit. on p. 7).

[24] Richard A Newcombe, Dieter Fox, and Steven M Seitz. "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 343–352 (cit. on p. 4).

[25] John Oliensis. "Exact two-image structure from motion". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.12 (2002), pp. 1618–1633 (cit. on p. 3).

[26] Linfei Pan et al. "Global structure-from-motion revisited". In: *European Conference on Computer Vision*. Springer. 2025, pp. 58–77 (cit. on p. 4).

[27] Shaifali Parashar, Daniel Pizarro, and Adrien Bartoli. "Isometric non-rigid shape-from-motion with riemannian geometry solved in linear time". In: *IEEE transactions on pattern analysis and machine intelligence* 40.10 (2017), pp. 2442–2454 (cit. on p. 4).

[28] Piotr Perzyna. "Fundamental problems in viscoplasticity". In: *Advances in applied mechanics* 9 (1966), pp. 243–377 (cit. on p. 8).

[29] Luigi Piccinelli et al. "UniDepth: Universal Monocular Metric Depth Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 10106–10116 (cit. on p. 4).

[30] Vincent Rabaud and Serge Belongie. "Re-thinking non-rigid structure from motion". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, pp. 1–8 (cit. on p. 3).

[31] David Recasens Lafuente et al. "The Drunkard's Odometry: Estimating Camera Motion in Deforming Scenes". In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on pp. 16, 26).

[32] Ronald S Rivlin. "Large elastic deformations of isotropic materials IV. Further developments of the general theory". In: *Philosophical transactions of the royal society of London. Series A, Mathematical and physical sciences* 241.835 (1948), pp. 379–397 (cit. on p. 7).

[33] José Javier González Rodríguez, José María Montiel, and Juan D. Tardós. "Nr-slam: Non-rigid monocular slam". In: *IEEE Transactions on Robotics* (2024) (cit. on pp. 5, 7).

[34] Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. "Nr-slam: Non-rigid monocular slam". In: *IEEE Transactions on Robotics* (2024) (cit. on p. 3).

[35] H Schiessel et al. "Generalized viscoelastic models: their fractional equations with solutions". In: *Journal of physics A: Mathematical and General* 28.23 (1995), p. 6567 (cit. on p. 7).

[36] Johannes L Schonberger and Jan-Michael Frahm. "Structure-from-motion revisited". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4104–4113 (cit. on pp. 3, 4).

[37] Olga Sorkine and Marc Alexa. "As-rigid-as-possible surface modeling". In: *Symposium on Geometry processing*. Vol. 4. July 2007, pp. 109–116 (cit. on pp. 8, 14, 15).

[38] Henrik Stewénius, Frederik Schaffalitzky, and David Nistér. "How hard is 3-view triangulation really?" In: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. Vol. 1. IEEE. 2005, pp. 686–693 (cit. on pp. 3, 4).

[39] Jonathan Taylor, Allan D Jepson, and Kiriakos N Kutulakos. "Non-rigid structure from locally-rigid motion". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, pp. 2761–2768 (cit. on p. 4).

[40] Sara Vicente and Lourdes Agapito. "Soft inextensibility constraints for template-free non-rigid reconstruction". In: *European conference on computer vision*. Springer. 2012, pp. 426–440 (cit. on p. 4).

[41] Monan Wang and Pengcheng Li. "A review of deformation models in medical image registration". In: *Journal of Medical and Biological Engineering* 39 (2019), pp. 1–17 (cit. on p. 7).

[42] Lihe Yang et al. "Depth anything: Unleashing the power of large-scale unlabeled data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 10371–10381 (cit. on p. 4).

[43] Ming Zeng et al. "Templateless quasi-rigid shape modeling with implicit loop-closure". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 145–152 (cit. on p. 4).

[44] Michael Zollhöfer et al. “Real-time non-rigid reconstruction using an RGB-D camera”. In: *ACM Transactions on Graphics (ToG)* 33.4 (2014), pp. 1–12 (cit. on p. 4).