NR-SLAM: Non-Rigid Monocular SLAM

Juan J. Gómez Rodríguez, José M.M. Montiel, *Member, IEEE* and Juan D. Tardós, *Fellow, IEEE*

arXiv:2308.04036v1 [cs.RO] 1 Aug 2023

Abstract—In this paper we present NR-SLAM, a novel non-rigid monocular SLAM system founded on the combination of a Dynamic Deformation Graph with a Visco-Elastic deformation model. The former enables our system to represent the dynamics of the deforming environment as the camera explores, while the later allows us to model general deformations in a simple way.

The presented system is able to automatically initialize and extend a map modeled by a sparse point cloud in deforming environments, that is refined with a sliding-window Deformable Bundle Adjustment. This map serves as base for the estimation of the camera motion and deformation and enables us to represent arbitrary surface topologies, overcoming the limitations of previous methods.

To assess the performance of our system in challenging deforming scenarios, we evaluate it in several representative medical datasets. In our experiments, NR-SLAM outperforms previous deformable SLAM systems, achieving millimeter reconstruction accuracy and bringing automated medical intervention closer. For the benefit of the community, we make the source code public.

I. Introduction

Visual Simultaneous Localization and Mapping (V-SLAM) techniques have been used in the last decade in a wide range of applications to locate an agent, from autonomous robots to augmented/virtual reality devices, in unknown environments. While these application may seem very different, they share a basic assumption that is crucial for V-SLAM techniques: the rigidity of the environment. While simple, this assumption allows to apply multi-view geometry to reconstruct the environment and locate the camera.

However, there are multiple applications in which the environment cannot be assumed to be stationary. Imagine a Minimal Invasive Surgery (MIS) procedure that aims to remove a polyp. By applying V-SLAM algorithms, the surgeon could be guided to the exact polyp location, easing the intervention. However, medical imagery presents scenarios in which the environment can freely deform due to the intervention, breathing or heartbeats. This simple change in the application domain renders current V-SLAM systems unable to produce accurate and faithful reconstructions. Moreover, most medical imagery is collected with a single monocular camera, which makes the problem of V-SLAM in deformable scenarios even harder, as no geometrical 3D information can be collected from the sensor. MIS imagery poses several additional issues, like poor texture and challenging illumination, worsening the data association step in any V-SLAM system.

If that was not enough, deformable V-SLAM poses another challenge: how to represent in a compact and general way
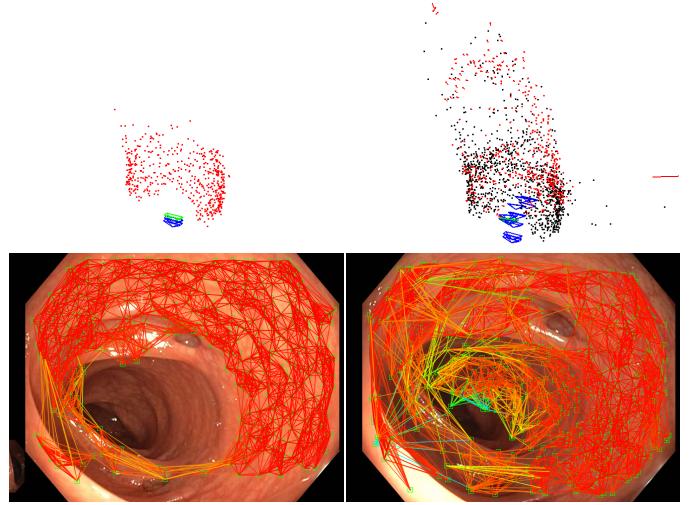
Figure 1: Reconstruction example of NR-SLAM in a real colonoscopy from the Endomapper dataset [1].

deforming surfaces. Some approaches use Signed Distance Functions (SDFs) implemented with 3D voxels, which do not scale well while exploring. Other approaches use a triangle mesh that scales better to bigger maps, but assumes planar topology, excluding tubes or surfaces with holes. Finally, most previous works assume isometric of quasi-isometric deformations, which is very questionable in many medical applications. All these challenges make monocular V-SLAM in medical sequences an open problem without a known general solution.

In this work we propose NR-SLAM, a novel non-rigid monocular SLAM system that copes with the above limitations and can be splitted in 3 main components (Fig. 2): deformable tracking, deformable mapping, and the map. The map represents the deformable environment with a time changing sparse point cloud that is easy to process and to scale when exploring new areas. Point maps are interconnected in a graph structure that relates points belonging to the same surface and that undergo similar deformations. On top of that, we model deformations as a set of pair-wise deformations with a simple mathematical representation. All these in conjunction with a robust and accurate data association based on optical flow make NR-SLAM able to reconstruct MIS imagery with unprecedented accuracy. To summarize, our contributions are:

- A complete monocular deformable SLAM system free from any topological and isometric assumptions.
- A map composed of a sparse set of points and a novel Dynamic Deformation Graph (DDG) that relates map points that deform together (Fig. 1), with a simple visco-elastic deformation model.
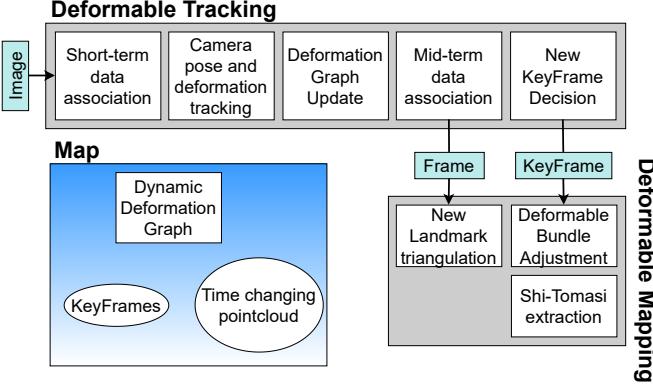- A robust and accurate semi-direct method for camera

Figure 2: Main system components of NR-SLAM.

tracking and deformation estimation based on the DDG.
- A deformable mapping method able to initialize, extend and refine the map estimations made by the deformable tracking.
- Experimental validation in a relevant set of medical datasets. See the accompanying video for examples.
- For the benefit of the community, we release NR-SLAM as an open source library[1].

II. RELATED WORK

In this section, we introduce a summary of the main V-SLAM algorithms for deformable environments (Table I). We first introduce systems that use 3D cameras (stereo or RGB-D) and then pure monocular systems.

A. Non-Monocular deformable SLAM

The use of 3D sensor is a common practice in literature to solve deformable SLAM, as the depth provided greatly reduces the complexity of the problem. DynamicFusion [2] presented a deformable SLAM system with just a RGB-D camera. It builds a canonical model of the scene, i.e. its shape at rest, and deforms it to fit the current observed depth. For that, it combines a form of a Embedded Deformation graph model (ED) [12] with an as-rigid-as-possible regularizer. Fundamentally, ED models build a discretization of the deformations space in a graph structure, speeding up dense reconstructions. DynamicFusion inspired new works like VolumeDeform [4], where the introduction of photometric image alignment improved the quality of the deformable reconstructions. Later, KillingFusion [3] proposed a new system in which deformations were required to be smooth and nearly isometric by using approximate Killing vector fields. MIS-SLAM [5] presented a stereo deformable SLAM system based on as-rigid-as-possible deformations in conjunction with an ED model to solve for the deformations. However, the above methods present a common critical limitation: the map representation as Signed Distance Functions (SDFs). While useful, SDFs scale poorly with the size of the map, rendering the methods not viable for exploration setups. This was addressed in Surfelwarp [6]

where a surfel representation was used instead of the classical SDF.

B. Monocular deformable SLAM

Monocular deformable SLAM is a much more difficult problem, as Structure from Motion (SfM) with a set of freely moving points is unconstrained. The field has been dominated mainly by 2 group of methods: Shape-from-Template (SfT) and Non-Rigid Structure-from-Motion (NRSfM). The former aims to recover the deformed shape of an object imaged by a monocular camera with respect to a a shape-at-rest (template). The most effective approach in SfT is to assume isometric deformations as it has been proven to provide accurate real time solutions [13], [14]. The other family of methods, NRSfM, makes reconstructions of a deforming surface from a stream of monocular images without a known template. First solutions to this problem were formulated using orthographic cameras and statistical models in which the shape of the surface is a linear combination of low-dimensional basis models [15], [16]. As the use of orthographic cameras is not feasible for real applications, later methods proposed to extend the isometric assumption with the use of perspective cameras [17], [18] producing excellent results. The recent work presented in [19] relaxes the isometric constraint to an area-preserving (equiareal) deformation constraint. However, the isometric and equiareal assumptions are not applicable to a wide range of scenarios like living tissue in endoscopies. Moreover, SfT and NRSfM methods assume a static camera, not being directly applicable to deformable SLAM scenarios.

The first deformable monocular SLAM system, able to build and extend a deformable map, was DefSLAM [7]. It was built on ORB-SLAM [20] by changing some parts of the pipeline to adapt it to the deformable case. First, a NRSfM algorithm [18] was integrated in the mapping thread to compute a surface template at keyframe rate. Secondly, those templates were aligned in the tracking thread and a SfT method [21] was used to jointly estimate the camera pose and deformations at frame rate. As these methods are very sensitive to spurious data, SD-DefSLAM [8] proposed to solve the data association step in the SLAM pipeline using optical flow techniques, greatly improving the robustness and accuracy of the system, making it usable in medical sequences. However, these family of methods have two main limitations: the map representation as a triangle mesh and the isometric assumption. The triangle mesh makes these methods unable to reconstruct surfaces with discontinuities or holes. This was addressed in DSDT [10] by formulating the camera tracking and deformation estimation as a photometric alignment of isometric surfels. Nevertheless isometry does not model certain types of surfaces like living tissue and thus this method is not suitable for endoscopic use. In [11] we proposed to tackle both limitations by solving the camera tracking and deformation estimation problem by modelling surfaces as sparse point clouds restricted by as-rigid-as-possible deformations and an ED model.

Although not a deformable SLAM system, it is worth mentioning the interesting work of RNN-SLAM [9], that combines DSO [22] with a recurrent neural network that

Table I: Summary of the most representative methods for deformable SLAM, in chronological order.

	Sensor	Tracking?	Mapping?	Map Initialization	Map Extension	Map representation
DynamicFusion [2]	RGB-D	✓	✓	Depth	Depth	Voxels
KillingFusion [3]	RGB-D	✓	✓	Depth	Depth	SDFs
VolumeDeform [4]	RGB-D	✓	✓	Depth	Depth	SDFs
MIS-SLAM [5]	Stereo	✓	✓	Stereo	Stereo	Sparse point cloud
SurfelWarp [6]	RGB-D	✓	✓	Depth	Depth	Surfels
DefSLAM [7]	Monocular	✓	✓	Plane at unit distance	Isometric NRSfM	Triangle mesh
SD-DefSLAM [8]	Monocular	✓	✓	Plane at unit distance	Isometric NRSfM	Triangle mesh
RNN-SLAM [9]	Monocular + depth network	✓	✓	Depth network	Depth network	Dense point cloud
DSDT [10]	Monocular*	✓	✗	Stereo	-	Surfels
MCPD [11]	Monocular	✓	✗	Essential matrix	-	Sparse point cloud
NR-SLAM (ours)	Monocular	✓	✓	Essential matrix	Rigid/deformable triangulation	Sparse point cloud + Dynamic Deformation Graph

estimates pose and depth from monocular images, obtaining the best reconstructions of colon sections so far. However, the method is specific for colonoscopies and ignores deformations.

Following [11] and in contrast with previous approaches, we solve here the full monocular deformable SLAM problem by integrating a Dynamic Deformation Graph (DDG), an evolved form of ED models, that allows us to model general geometries even with discontinuities, in conjunction with visco-elastic deformations, to model better the deformations occurring in real medical scenarios.

III. NR-SLAM DEFORMATION MODEL

This section is devoted to the deformation model used in NR-SLAM, which is one of the main insights of the system as it is the central piece of all components in NR-SLAM (Fig. 2): the deformable tracking uses it to compute the current camera pose and the environment deformations in all images, the deformable mapping employs it to extend and refine the map, that keeps track of which points deform together in a graph data structure.

A. Visco-Elastic deformation Model

Modeling general deformations is a complex problem as it directly depends on the problem domain to be solved. Too general models present observability problems while specific ones can loose capabilities when the observed deformations deviate from the assumptions made.

For those reasons, we design our deformation model to be general enough to represent different deformations and topologies and, at the same time, be specific enough to our application domain in order to get accurate results. Our deformation model takes into account that in medical scenarios deformations tend to happen slowly and are locally similar in direction and magnitude. Here we present our Visco-Elastic deformation model, representing a spring and a dumper connecting two surface points. On this way, the deformation model is general enough to represent arbitrary surfaces while constraining enough the problem to yield reasonable reconstructions.

First, we make a Camera-over-Deformation (CoD) assumption, that is we assume that most of the image innovation comes from camera motion rather than from deformations of the environment. This assumption, a relaxation of the rigidity assumption of conventional SLAM systems, is usually true when working with temporally close images as in a video sequence from an endoscope. This allows our system to have a reasonable initial seed when trying to initialize and extend the map. Note that this assumption is in sharp contrast with most NRSfM methods that assume a static camera.

Second, we base our deformable model in a set of local deformations modeled as 3D displacements. We consider that small local surface areas tend to behave as almost rigid, while in a more global context deformations can be larger. This is a reasonable prior in medical scenarios, where close points that belong to the same tissue suffer small changes in their relative distances (for example, with muscle stretching), while the relative distance between points that are further apart in the same organ or that belong to different organs suffer bigger changes due to body motions or breathing. Again, this is in contrast with previous NRSfM methods that assume isometric or equiareal deformations.

In all optimizations performed in our system (deformable tracking and Deformable Bundle Adjustment), we encode our deformation model as a set of two regularizers added to the optimization cost function. First, we use an *Elastic regularizer* that penalizes changes in the distance between two locally close surface points i and j formulated as:

$$E^t_{ij,elas} = k \frac{(d^t_{ij} - d^0_{ij})^2}{d^0_{ij}} \quad (1)$$

where k is a global hyperparameter that controls how strong is the regularizer and $d^t_{ij} = \|\mathbf{x}^t_i - \mathbf{x}^t_j\|_2$ is the Euclidean distance between points i and j at time t.

We also add a *Viscous regularizer* that penalizes locally close points that move with different velocities:

$$E^t_{ij,visc} = b^t_{ij}\|\boldsymbol{\delta}^t_i - \boldsymbol{\delta}^t_j\|^2 \quad (2)$$

where $\boldsymbol{\delta}^t_i = \mathbf{x}^t_i - \mathbf{x}^{t-1}_i$ is the motion suffered by point i at time t and b^t_{ij} is a pairwise weight between points i and j that models the cross influence between this pair of points computed as:

$$b^t_{ij} = \exp\left(\frac{-d^2_{ij,max}}{2\sigma^2}\right) \quad (3)$$

where $d_{ij,max}^2$ is the maximum distance observed between map points i and j and σ is a radial basis that controls the influence between points with respect to their relative distance. This weight is directly inspired from the one used in DynamicFusion [2] to regularize the deformation nodes and later used again in [11] to regularize the different points tracked.

Interestingly enough, this model has a direct mechanical interpretation: the Elastic regularizer acts as a spring connecting two surface points where k in equation 1 is an homologous to the spring constant found in the Hook's law and the viscous regularizer as a damper where the weight b_{ij}^t in equation 2 is the proportional constant of a physical damper. This provides our formulation with a physical meaning, enhancing its interpretability.

B. Dynamic Deformation Graph

Our deformation model relates locally close points in order to constrain their deformations as they tend to locally behave in the same manner. However, this is not always true as there can be points that at some instant can be spatially close but belong to different surfaces that deform differently. Regularizing those points together would lead to incorrect deformation estimations. In order to control this, we propose to encode surface point relationships inside a graph structure that we call Dynamic Deformation Graph (DDG). Basically, DDG represents points as the graph nodes and the relationships between them as the graph edges, which carry the information of which points should be regularized together.

The DDG is first initialized according to the relative 3D distance between points but is periodically updated as the system observes the deformations of the scene (Fig. 3). This leads to a model that is able to understand the dynamics of different surfaces and integrates them correctly into our optimization backbone. The criteria used to prune a connection in the DDG is based in a simple geometric test: if the distance between two connected points grows beyond some streching threshold, points are considered to not behave in a similar fashion and therefor are disconnected in the DDG. Mathematically, this criteria is expressed as:

$$\text{prune edge } i,j \text{ in the DDG if}$$
$$\frac{d_{i,j,max} - d_{i,j,min}}{d_{i,j,min}} > th_{streching} \quad (4)$$

Following with the mechanical interpretation of our deformation model, this criteria can be interpreted as the spring and damper joining two points breaking if the points get too far apart.

IV. DEFORMABLE TRACKING

In this section we present the tracking module of NR-SLAM that integrates our deformation model embedded in the DDG into a pipeline that estimates in real time the camera pose and the deformations of the observed scene for each incoming frame.

A. Short term data association

Accurate inter-frame data association is crucial in order to obtain good accuracy and robustness, but medical images suffer from weak texture, local illumination changes and strong reflections. Direct methods like [22] have shown good performance in low texture environments getting point associations as a byproduct of the tracking. However, this is done by imposing a global rigid transformation to all points as the environment is assumed stationary, and by assuming illumination constancy or just a global illumination change. Both assumptions are far from true in endoscopy due to deformations and moving illumination.

Instead, we use a semi-direct approach that performs photometric data association with Shi-Tomasi features [23] using the modified multi-scale Lucas-Kanade algorithm proposed in [8]:

$$\arg\min_{\mathbf{d}_i, \alpha_i, \beta_i} \sum_{\mathbf{v} \in P(\mathbf{u}_i)} \left(I^0(\mathbf{v}) - \alpha_i I^t\left(\mathbf{v} + \mathbf{d}_i\right) - \beta_i \right)^2 \quad (5)$$

where $P(\mathbf{u}_i)$ is a small pixel patch centered at the keypoint $\mathbf{u}_i$; I^0 is the first frame, where the points are intialized, and I^t is the current frame at time t. These patches are updated every five images to account for big scale changes or rotations. Also a local illumination invariance is achieved by computing local gain α_i and bias β_i terms for each point. This algorithm has been proven to achieve excellent results when tracking image features in short time steps even in the presence of deformations or local illumination changes [8][11]. The key of its performance lies in using no global model: each point can move freely with respect the others.

In order to remove outlier tracks, we compute the *Structural Similarity Index* (SSIM) [24] between the reference x and tracked y pixel patches:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (6)$$

where μ_x and σ_x are the mean and covariance of the pixel patch, σ_{xy} is the crossed covariance between both patches and C_1 and C_2 are constant values to avoid inestability when means and covariances approaches to zero. This has been proven to be a good similarity metric for small pixel windows as it combines in a same metric a luminance, contrast and structure comparison.

B. Camera pose and map deformation tracking

The goal of the deformation tracking task is to jointly estimate the current camera pose $\mathbf{T}_{C^tW}$ and map point deformations $\boldsymbol{\delta}_i^t$ for the current input frame I^t. With that in mind, we combine our deformation model with a reprojection error obtained from the matches provided by our data association algorithm. The idea behind this is to confine possible solutions in a set that is consistent with what the camera is actually seeing. Therefor, we build our cost function with a reprojection term and a deformation term:

(a) Initial DDG.

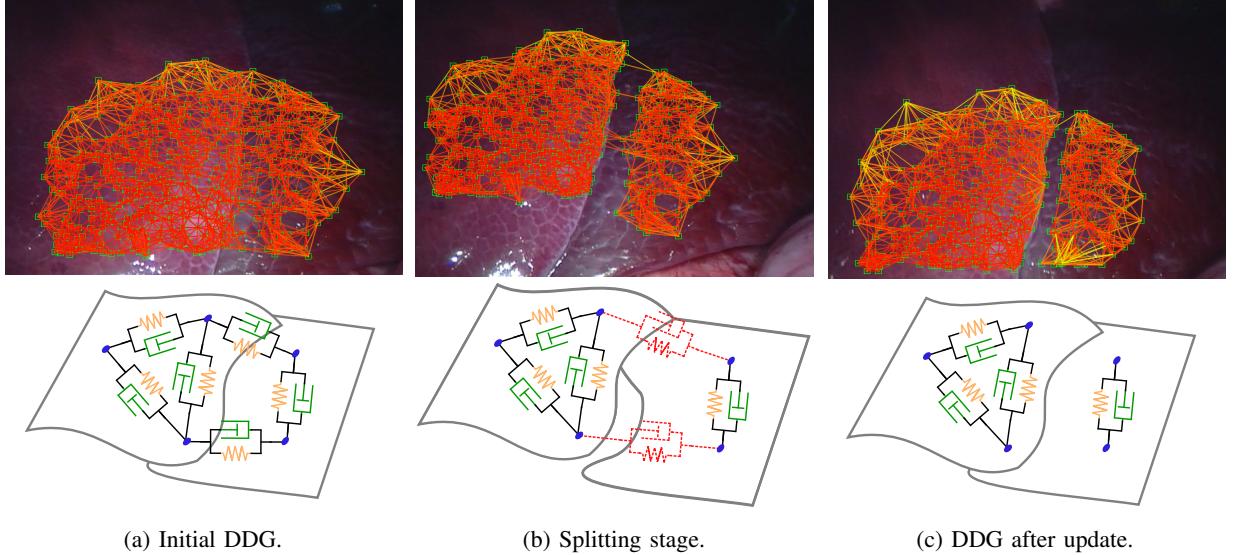(b) Splitting stage.

(c) DDG after update.

Figure 3: Visual representation of our DDG in a medical sequence where 2 surfaces move independently. At the start, the DDG is built by creating point connections according to their proximity in 3D (Fig. 3a). After some deformations, both surfaces move away up to a point in which some connections break according to the criterion presented in section III-B (Fig. 3b). After that, the DDG is updated and points belonging to different surfaces are not regularized together(Fig. 3c).

$$\mathcal{E}^t = \sum_{i \in \mathcal{P}} (E_{i,rep}^t + E_{i,def}^t) \qquad (7)$$

where $\mathcal{P}$ represents the set of map points tracked by our data association algorithm in the current frame. The reprojection error term can be expressed as:

$$E_{i,rep}^t = \rho(\|\mathbf{u}_i^t - \hat{\mathbf{u}}_i^t\|_{\Sigma_{rep}}^2) \qquad (8)$$

where ρ is the Hubber robust cost, Σ_{rep} is the uncertainty of the matched image features and $\mathbf{u}_i^t$ and $\hat{\mathbf{u}}_i^t$ are respectively the match of feature i in the current image I^t and its projection given by:

$$\hat{\mathbf{u}}_i^t = \Pi(\mathbf{T}_{C^tC^0}(\mathbf{x}_i^{t-1} + \boldsymbol{\delta}_i^t)) \qquad (9)$$

The accuracy of indirect methods is limited by the feature detector resolution (typically no less than 1 pixel). However, matches obtained with semi-direct methods provide subpixel accuracy boosting in this way the accuracy of our reprojection term while keeping its nice convergence basin.

The deformation term in eq. 7 encodes our deformation model. More precisely, points in $\mathcal{P}$ are regularized together according to our DDG following the equation below:

$$E_{i,def}^t = \sum_{j \in \mathcal{N}_{DDG}(i)} E_{ij,elas}^t + E_{ij,visc}^t \qquad (10)$$

where $\mathcal{N}_{DDG}(i)$ are the neighbours of point i in the DDG. Goal function in eq. 7 is minimized at frame rate using Weighted Non Linear Squares algorithms in order to find the optimal camera pose and deformations:

$$\mathbf{T}_{C^tW}, \boldsymbol{\delta}_i^t = \underset{T_{C^tW}, \boldsymbol{\delta}_i^t}{\arg\min} \ \mathcal{E}^t \qquad (11)$$

(a) Fully connected
$(D = \infty)$
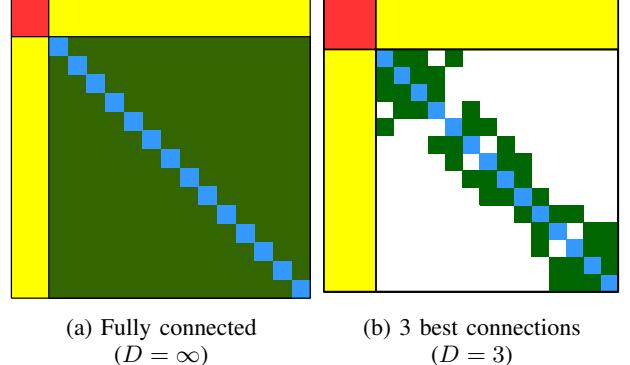
(b) 3 best connections
$(D = 3)$

Figure 4: Hessian of camera and deformation tracking with different values for the maximum degree D in the DDG. The visco-elastic regularizer densifies the point Hessian.

As this is a highly non-linear problem, it is really important to choose a good first guess for the solution ti ease the solver convergence. For that, in order to assert smooth camera trajectories, the seed for $\mathbf{T}_{C^tC^0}$ is computed as follows:

- First, a coarse estimation $\hat{\mathbf{T}}_{C^tC^0}$ is computed using a constant velocity model.
- Second, using the short term data association and $\hat{\mathbf{T}}_{C^tC^0}$, we run a rigid pose-only optimization to refine the seed.

The optimization problem stated in eq. (11) is fairly expensive due to several factors. First, we are estimating $6 + 3|\mathcal{P}|$ degrees of freedom instead of just 6 like in the classical rigid pose estimation problem. Secondly, the introduction of constraints in the form of regularizers between map points (eq. 10) induces off-diagonal blocks in the optimization Hessian, effectively densifying it and increasing the computational cost of the optimization (fig 4a). In an effort to keep a balance
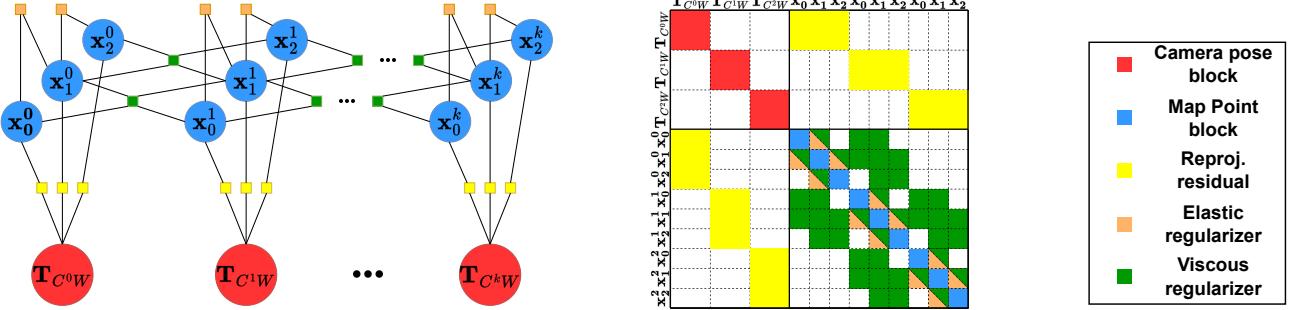
Figure 5: Factor graph and Hessian of our proposed Deformable Bundle Adjustment (DBA).

between accurate estimations and computational cost, we decide to set an upper limit D to the degree of nodes in the DDG, which limits the number of regularizers per map point to D, sparsifying the optimization Hessian (Fig. 4b). We select which points have to be regularized together by selecting the best connections in the DDG according to their weight b_{ij}^t (eq. 3).

C. Mid Term Data Association

Once the current camera pose and map deformation are estimated for the current frame, our algorithm tries to perform a mid-term data association step [25] to refine the initial solution by trying to reuse as much data as possible. In other words, map points that have not been used in the first estimation, either because they were lost by our data association algorithm or marked as outliers in previous estimations, are searched in the current image using a guided matching scheme: considering the current geometry, we project into the image candidate map points to be reused. These projections are used as seeds for a Lucas-Kanade tracker that tries to match them in the current image. Those map points successfully matched are integrated into the next tracking estimation, contributing to reducing the overall error.

The key here is the 3D positions used as seeds for the candidate points. There are two cases:

- Points updated thanks to the DDG: map points that are neighbours of the currently tracked features in the DDG are updated thanks to the visco-elastic regularizers.
- Other in-frustum map points: for the rest of the points, their last known position is used if it falls within the frustum of the current camera.

Having a map available allows us to compute these mid term data associations. This plays a crucial role in our system as it greatly helps to improve the reconstruction results by reusing as much information as possible from the environment already explored.

V. DEFORMABLE MAPPING

Mapping while exploring deformable environments is a crucial but hard task whose goals are to to create, extend and refine the map used by the deformable tracking part. In this section, we describe the basic functions performed by our proposed deformable mapping algorithm.

A. Deformable Bundle Adjustment

Inspired by state of the art SLAM algorithms, we perform a refinement of the map each time a keyframe is inserted by running a Bundle Adjustment. This operation has been shown to be too expensive to be performed at frame rate, and it is usually applied to a selected set of keyframes $\mathcal{K}$ and map points $\mathcal{P}$ to reduce the computational burden. $\mathcal{K}$ and $\mathcal{P}$ in rigid systems are selected using a covisibility criteria around the latest keyframe to maximize the information gained during the optimization [20]. However in this work, due to the nature of continuous deformations, $\mathcal{K}$ is built from the latest N_k keyframes in a sliding window fashion as we require them to be consecutive to correctly represent our deformation model.

This optimization, which we coin as Deformable Bundle Adjustment, aims to minimize the following objective function:

$$\mathcal{E}_{DBA} = \sum_{t \in \mathcal{K}} \sum_{i \in \mathcal{P}^t} (E_{i,rep}^t + E_{i,def}^t) \tag{12}$$

where $E_{i,rep}^t$ and $E_{i,def}^t$ are the same error terms as in eq. (7) and $\mathcal{P}^t$ is the set of points observed from keyframe t. This objective function is noticeably more expensive than a regular rigid Bundle Adjustment. First, the number of variables to estimate increases from $|\mathcal{K}|+|\mathcal{P}|$ to $|\mathcal{K}|\cdot|\mathcal{P}|$. Second, the cost of classical Bundle Adjustment is lineal in the number of points as one can apply the Schur trick (marginalizing map points using Schur complement) as the point-point section of the Hessian is block diagonal. In the deformable case, the introduction of regularizers between points adds off-diagonal blocks in the point-point section of the Hessian (Fig. 5). This densification destroys the benefit of the Schur trick, increasing the computational cost of the DBA with respect to its rigid counterpart.

B. Monocular Map Initialization

In order to track a camera and the deformations, first the mapping needs to create a map. In settings with stereo or RGB-D sensors, a map can be easily initialized from a single view. In rigid monocular SLAM this can be performed with Structure-from-Motion (SfM) algorithms to get a first map estimation up to an unknown scale factor.

However, in a deforming environment with a set of freely moving points, SfM has infinitely many solutions. Authors

in the literature have proposed different Non-Rigid Structure from Motion algorithms (NRSfM), that need to assert a set of strong conditions to constraint the solution like orthographic cameras, stationary cameras, isometric or area-preserving deformations, or close-to-planar surfaces. In most practical deformable SLAM scenarios, and in particular, in medical SLAM, these assumptions are invalid.

Instead, we propose a new map initialization method for deformable environments based on the CoD assumption: most image innovation comes from the camera motion rather than from deformations. With that in mind, our problem statement is to compute the relative camera position between two temporally close images $\mathbf{T}_{C^t,C^0}$ and the 3D positions of the observed map points in each of the images $\mathbf{x}_i^0$, $\mathbf{x}_i^t$.

Our proposed approach is a two step algorithm. First, using the CoD assumption, we compute an initial estimation of the map assuming using SfM techniques as if the scene was rigid. For that, we track a set of keypoints in the images and estimate from their projection rays an Essential Matrix at low parallax which is decomposed to obtain the relative pose of both cameras $\mathbf{T}_{C^t,C^0}$ that is used to perform a rigid triangulation of the matched keypoints. Second, we perform a Deformable Bundle Adjustment to refine the estimated geometry to adapt it to the possible deformations not considered in the first step. As our Deformable Bundle Adjustment makes use of the DDG we need to initialize it in a naive way using no geometrical information as no previous map is available. To cope with that, we cluster [26] feature tracks according to their optical flow shape to build a preliminary DDG connecting points that were clustered together.

The key to the performance of our proposed algorithm lies in the idea that under small deformations, rigid algorithms can get a reasonable first estimation of the geometry. For that, it is crucial to initialize the map as soon as possible to reduce the magnitude of any deformations observed. This implicitly means working under small parallax which can easily worsen the initial rigid estimation. We address this by combining the subpixel accuracy of our semi-direct data association algorithm with the weighted mid point triangulation algorithm proposed in [27] as it provides good 3D estimations under fairly small parallax ($\backsim$0.3 degrees). This triangulation method also has an interesting property: it is rotational invariant and is easily applied to any type of camera model either pinhole or fish-eye, as it uses projection rays, making it suitable for a wide range of applications.

Once a map has been successfully created, we need to initialize our DDG with the estimated geometry. This also requires setting a specific σ to compute the weights (eq. 3) between points. We make this process fully automatic by scaling the map to have a predefined mean depth (in colonoscopies we use 3cm) and σ is set as the depth standard deviation of the scaled map.

C. Map Point Triangulation

As the camera explores new regions of the scene, its pose is estimated by the deformable tracking method and it is necessary to add new points in the map to cover the new regions. In rigid environments, points can be triangulated from two views with enough parallax.

In deformable environments, there can be small time windows where the environment remains stationary enough for rigid triangulation to give consistent depth estimations. However, in strong deformation periods, deformation and camera motion estimations can be coupled together and rigid triangulation would be inaccurate. For that reason, we run in parallel two triangulation methods, one rigid and one deformable, and apply a model selection strategy to select the best one.

Both procedures start by tracking during a set of frames $\mathcal{F}$ a set of candidate features to be triangulated $\mathcal{C}$. Note that we perform triangulation at frame rate looking for a compromise between having enough parallax but not too much deformation that could degrade the results.

1) Rigid Triangulation: For each candidate track in $\mathcal{C}$ we take the first $\mathcal{F}_f$ and last frame $\mathcal{F}_l$ where it was observed and check that the mean map point deformation in all intermediate images is less than a threshold. If this rigidity check is satisfied, we apply the weighted mid point algorithm [27] with the first and last frame to triangulate the 3D point, for which we also check parallax and reprojection errors.

2) Deformable Triangulation: We propose a novel approach to triangulate new map points observed during deformable periods. This new algorithm relies on the fact that close points should deform together and so, have a similar 3D flow in the same period of time. With that in mind, for feature in $\mathcal{C}$ we try to adjust a 3D trajectory $\mathbf{X}_i = [\mathbf{x}_i^t, \mathbf{x}_i^{t+1}, ..., \mathbf{x}_i^{t+n}]$ so that:

- it has a low reprojection error in the frames of $\mathcal{F}$ in which the feature has been matched, and
- the 3D trajectory $\mathbf{X}_i$ is similar to those of some predefined neighbour map points $\mathcal{N}_i$. Since up to this point there is no geometry available, neighbours in $\mathcal{N}$ are selected by feature proximity in the last image.

We then perform the estimation of the 3D trajectory of the new landmark by solving the following minimization problem that models our assumptions:

$$\mathbf{X}_i = \arg\min_{X_i} \mathcal{E}_{i,tria} \qquad (13)$$

where:

$$\mathcal{E}_{i,tria} = \sum_{t \in \mathcal{F}} E_{i,rep}^t + E_{i,visc}^t$$
$$E_{i,visc}^t = \sum_{j \in \mathcal{N}_i} E_{ij,visc}^t \qquad (14)$$

Terms $E_{i,rep}^t$ and $E_{ij,visc}^t$ are the reprojection error and the Viscous regularizer presented in eqs. 8 and 2 respectively. As the problem is non-linear, the seed used for the 3D trajectory $\mathbf{X_i}$ may have a strong impact in the estimation output. For that, assuming that surfaces are locally smooth, we initialize each individual point $\mathbf{x}_i^t$ in $\mathbf{X}_i$ by unprojecting the tracked feature and setting its depths as the average depth of the neighbours $\mathcal{N}_i$. In order to accept a deformable triangulation as good, we check that the reprojection error $E_{i,rep}^t$ in each one of the frames $\mathcal{F}$ is lower than a given threshold. We also verify that

Table II: Comparison of monocular SLAM methods in simulated colonoscopies with different deformations.

Sequence	A (mm)	ω (rad/s)		ORB-SLAM3	SD-DefSLAM	NR-SLAM
simulated-0	0.0	0.0	RMSE (mm)	**3.52**	11.67	3.87
			# Fr.	84	73	84
simulated-1	2.5	2.5	RMSE (mm)	4.30	11.96	**2.26**
			# Fr.	84	27	84
simulated-2	2.5	5.0	RMSE (mm)	5.63	11.49	**2.72**
			# Fr.	84	78	84
simulated-3	5.0	2.5	RMSE (mm)	5.09	12.67	**2.48**
			# Fr.	84	27	84
simulated-4	5.0	5.0	RMSE (mm)	8.11	11.33	**3.95**
			# Fr.	84	29	84
simulated-5	10.0	2.5	RMSE (mm)	9.71	11.71	**3.80**
			# Fr.	84	84	84
simulated-6	10.0	5.0	RMSE (mm)	15.62	38.27	**5.82**
			# Fr.	56	71	84

the estimated trajectory $\mathbf{X}_i$ respects the viscous regularizer, that is, at least half of the viscous regularizers used $E_{ij,visc}^t$ are below a threshold error.

3) Model Selection: From $\mathcal{C}$ and $\mathcal{F}$, we apply both triangulation algorithms, obtaining a set of successful rigid triangulations $\mathcal{R}$ and a set of successful deformable triangulations $\mathcal{D}$. To decide which method got the best result we apply the following model selection method:

1) accept $\mathcal{R}$ if $|\mathcal{R}| > 1.5\,|\mathcal{D}|$
2) accept $\mathcal{D}$ if $|\mathcal{D}| > 1.5\,|\mathcal{R}|$
3) otherwise no clear winner is found and then, no triangulations are accepted.

Once inserted, the new features will be refined in the next DBA when a new keyframe is inserted in the map.

VI. EXPERIMENTS

We evaluate our proposed deformable SLAM system in different datasets to assess its capabilities. Our main focus is medical sequences, for that reason, we make use of the Hamlyn dataset [28] and the Endomapper dataset [1], two of the main datasets used in medical SLAM research. Specifically, we performed evaluations on:

1) **Endomapper simulation sequences** of colonoscopies that allow us to evaluate the camera pose estimation and the accuracy of reconstruction with different degrees of deformation.
2) **Hamlyn real sequences** acquired with a stereo endoscope, that allow us to evaluate reconstruction accuracy under different deformations and topologies.
3) **Endomapper real sequences** of colonoscopies, that allow qualitative performance evaluation on real medical scenarios.

Quantitative comparison with other state of the art SLAM algorithms for deformable scenarios is performed using the Hamlyn dataset. We compare with the reconstruction performance of deformable monocular methods in the literature, namely DefSLAM [21] and SD-DefSLAM [8], that are full SLAM methods, and DSDT [10] and MCPD [11], that are just deformable tracking methods. The main characteristics of these algorithms can be found in table I. In some cases, we also include in the comparison ORB-SLAM3 [25] as a rigid

monocular SLAM system. We refer the reader to the provided video to have a better visualization of the results.

A. Implementation details

The method was fully implemented from scratch in C++ and runs entirely on CPU. We use the OpenCV library [29] for computer vision and image processing tasks, Sophus library [30] for representing SE(3) objects and g2o library [31] to implement non-linear least squares optimizations. An open-source version of NR-SLAM is available for the benefit of the community.

B. Quantitative evaluation in simulated colonoscopies

For evaluating the performance of deformable SLAM methods, we provide in the Endomapper dataset [1] a set of simulated colonoscopies. These sequences were obtained with the VR-Caps simulator [32] using their colon model, which was obtained from a Computerized Tomography of a patient, adding photorealistic textures. The sequences simulate a colonoscope insertion manoeuvre with a forward motion, with synthetic deformations added to the scene. The deformations are modelled via a sine wave propagated along the colon model according to the following formula:

$$V_y^t = V_y^0 + A\sin(\omega t + V_x^0 + V_y^0 + V_z^0) \qquad (15)$$

where V_x^0, V_y^0 and V_z^0 are the coordinates of a surface point at rest. The strength of the deformations is controlled by the sine amplitude A and its frequency ω. We evaluate the reconstruction accuracy of NR-SLAM under different combinations of values of A and ω, enabling us to test our system under different deformation conditions, from smaller to more sudden and aggressive deformations.

Results are presented in table II. With a static scene (sequence-0), NR-SLAM is slightly worse than ORB-SLAM3. This suggests that the synthetic texture used is very good for ORB features and the deformability assumption, when false, plays against our method. With mild deformations (sequences 1 to 4) the accuracy of NR-SLAM actually increases, doubling that of rigid methods. With stronger deformations, that are unbearable for rigid methods, our method still maintains good accuracy.
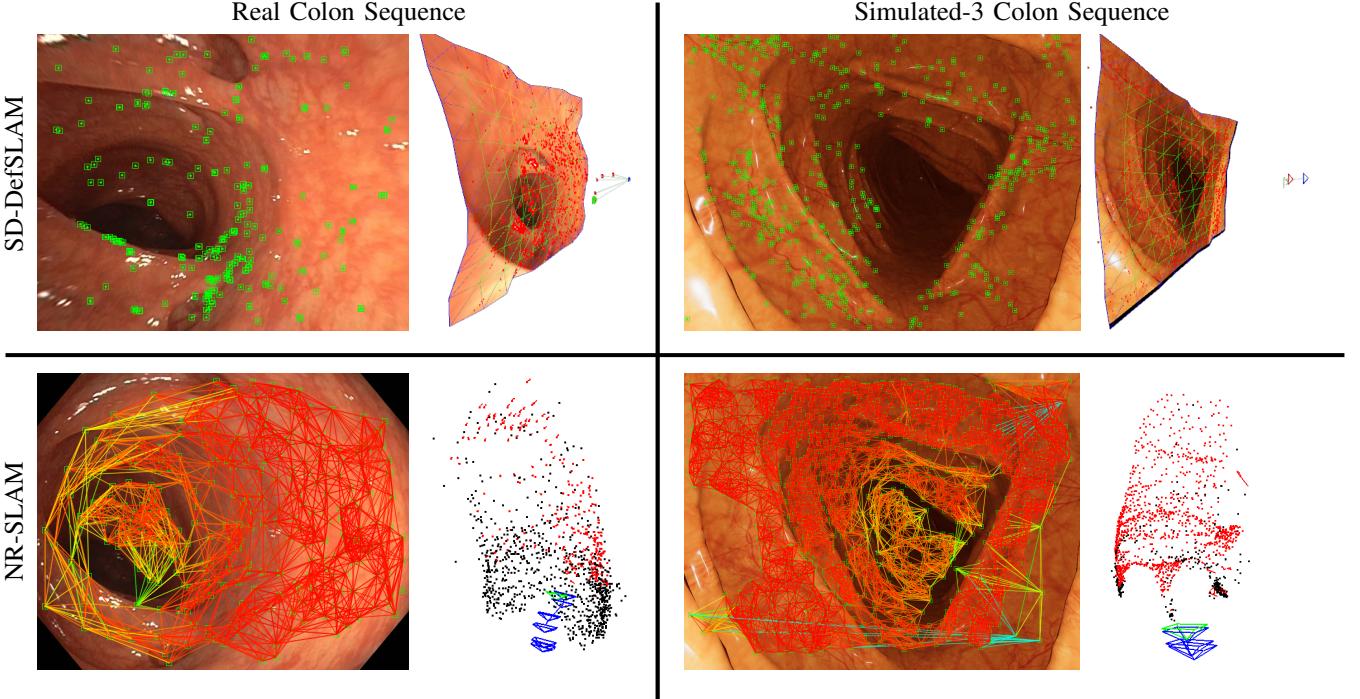
Figure 6: Comparison between SD-DefSLAM and NR-SLAM in a real and a simulated colonoscopy for Endomapper dataset. While SD-DefSLAM estimates an almost-planar shape, NR-SLAM successfully reconstructs the correct tubular shape.

Interestingly enough, SD-DefSLAM, that is the current state of the art in deformable SLAM for medical applications, loses track in most sequences, and obtains poor accuracy, being worst than rigid methods. The reason lies in its strong assumptions: (1) that the environment is a continuous surface with planar topology and (2) that deformations are isometric. The first one is clearly violated in colonoscopies (see Fig. 6) where discontinuities due to haustra and the tubular topology and shape seem to have a huge impact, causing poor accuracy even with a static scene. In comparison, the isometric assumption only seems to have a big impact under strong deformations.

C. Quantitative evaluation in real sequences

We perform a quantitave study of the reconstruction performance of our proposed system and other state of the art algorithms with the Hamlyn dataset [28]. This dataset is composed of several laparoscopic calibrated stereo sequences that can be used to get depth ground-truth. From all of them, we select the following sequences:

- **abdominal-sequence-1**: a quasi-rigid sequence of an abdominal cavity explored with a smooth camera motion.
- **abdominal-sequence-6**: a quasi-rigid sequence with a close look-up of another abdominal cavity.
- **organs-sequence-19**: a sequence imaging a deforming organ with some tools interfering in the field of view.
- **exploration-sequence-20**: a general exploration sequence of several deforming organs and tissues.
- **liver-sequence-21**: a sequence imaging two lobes of a liver, with significant relative motion due to breathing, as shown in figure 3.

We perform two comparisons with this dataset. First, an ablative study of the robustness and accuracy of camera tracking and deformation estimation, given an initial 3D reconstruction obtained with stereo. For this experiment, we compare the tracking part of ORB-SLAM3, DefSLAM, SD-DefSLAM and NR-SLAM, and DSDT and MPCD that are pure tracking algorithms, in tree sequences. We evaluate the number of successfully tracked frames and the reconstruction RMSE across all those frames, comparing with stereo ground-truth. Results are shown in table III. Clearly, the deformable tracking proposed in NR-SLAM obtains superior results in both metrics, setting the state of the art in monocular camera tracking and deformation estimation. This is due to both the quality of the data association and our deformation model. Interestingly enough, our deformable tracking algorithm is similar to the one presented in MCPD, but we get here much better results due to the addition of the DDG and the visco-elastic regularizer. Compared with SD-DefSLAM we get much better accuracy in sequences 20 and 21. Again, the reason lies in the strong assumptions made by SD-DefSLAM. In this case, surfaces are reasonably planar and continuous, but deformations are not isometric, which is well captured by our deformation model.

The second experiment is performed with full SLAM setups, comparing DefSLAM, SD-DefSLAM and NR-SLAM in three sequences. Again, results in table IV show that our proposed system outperforms other algorithms, getting significantly more tracked frames and smaller reconstruction errors in sequences 1 and 20. The exception is sequence 19 where SD-DefSLAM is able to get lower errors. However, this was an expected result as this sequence has several

Table III: Comparison of monocular tracking methods, with stereo initialization and no mapping.

		Rigid map	Deformable map			
		ORB-SLAM3	SD-DefSLAM	DSDT	MCPD	NR-SLAM
abdominal-sequence-6	RMSE (mm)	4.85	2.72	3.17	–	**2.37**
	# Fr.	128	286	300	–	**1236**
exploration-sequence-20	RMSE (mm)	1.37	4.68	2.90	1.48	**1.09**
	# Fr.	220	252	500	350	**609**
liver-sequence-21	RMSE (mm)	–	6.19	1.30	1.55	**0.59**
	# Fr.	–	323	300	300	**376**

Table IV: Comparison of full monocular deformable SLAM methods with initialization, tracking and mapping.

		DefSLAM	SD-DefSLAM	NR-SLAM
abdominal-sequence-1	RMSE (mm)	23.98	22.20	**12.26**
	# Fr.	858	958	**1030**
organs-sequence-19	RMSE (mm)	13.02	**6.63**	11.09
	# Fr.	1300	1300	1300
exploration-sequence-20	RMSE (mm)	17.02	12.56	**7.13**
	# Fr.	1579	1750	**1811**

surgical tools intrusions, and SD-DefSLAM uses deep learning techniques to segment and mask them. This makes it able to completely ignore surgical tools that worsen significantly the reconstruction accuracy of the other methods.

D. Qualitative evaluation in real sequences

We finally provide qualitative results in real in-vivo human colonoscopy sequences from the Endomapper dataset. These sequences only provide the RGB images of the procedures with no labels or ground-truth available. Nevertheless, it is an interesting dataset to qualitative test our proposed system as it contains a whole bunch of challenges typically found in medical imagery: tissue deformations, depth discontinuities, little to no texture, varying lighting conditions, specular reflections and fish-eye optics. Figure 7c shows how NR-SLAM, despite these challenges, is able to initialize and extend a deformable map represented with a DDG, while tracking the camera pose and scene deformation in a real colonoscopy, going beyond the capabilities of previous deformable SLAM systems.

VII. CONCLUSIONS

We have proposed NR-SLAM, the first monocular deformable SLAM system free from strong assumptions like planar topology, almost-planar shape, continuous surface and isometric or equi-areal deformations. It automatically creates, extends and refines a map that is used by our deformable tracking to estimate the camera pose and the map deformation for each frame. Our Visco-Elastic deformation model has a nice physical interpretation and is able to model the generic deformation appearing in medical environments. Our Dynamic Deformation Graph smoothly handles discontinuities from occlusion, like different colon haustra, or discontinuities that appear during mapping, like two organs moving apart.

Indeed, our deformable tracking method presents a general, robust and accurate solution for the reconstruction of deforming surfaces, achieving outstanding results. However, deformable mapping still has important open issues. General NRSfM is an undetermined problem, that needs priors to be solved. Previous methods use strong priors with a closed

mathematical formulation (isometry, orthographic cameras, ...) but our results show that they are problematic for many medical scenarios. We decided to use milder priors with a clear physical meaning in which the environment undergoes smooth deformations, both temporally and spatially, allowing us to cope with a wider range of scenes. Our experiments show good performance in deforming and almost planar environments (Fig. 7a) and, for the first time, in environments with depth discontinuities and mild deformations like colonoscopies (Fig. 7b and 7c). However, while our deformable mapping outperforms the state-of-the-art methods, it still struggles to create and extend a map in sequences with both, depth discontinuities and strong deformations, as in liver-sequence-21 (Fig. 3). This represents the main limitation of our system: in this type of scenarios it needs a given 3D seed to initialize the map or to add new points to it. Probably, this can be addressed in future work using neural networks that estimate depth from monocular images. Also, robustness can be significantly improved by adding surgical tools segmentation and place recognition techniques to recover from occlusions [8].

REFERENCES

[1] P. Azagra, C. Sostres, Á. Ferrandez, L. Riazuelo, C. Tomasini, O. L. Barbed, J. Morlana, D. Recasens, V. M. Batlle, J. J. Gómez-Rodríguez et al., "Endomapper dataset of complete calibrated endoscopy procedures," *arXiv preprint arXiv:2204.14240*, 2022.

[2] R. A. Newcombe, D. Fox, and S. 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.

[3] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic, "KillingFusion: Nonrigid 3D reconstruction without correspondences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1386–1395.

[4] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger, "VolumeDeform: Real-time volumetric non-rigid reconstruction," in *European conference on computer vision.* Springer, 2016, pp. 362–379.

[5] J. Song, J. Wang, L. Zhao, S. Huang, and G. Dissanayake, "Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 155–162, 2017.

[6] W. Gao and R. Tedrake, "SurfelWarp: Efficient non-volumetric single view dynamic reconstruction," *arXiv preprint arXiv:1904.13073*, 2019.

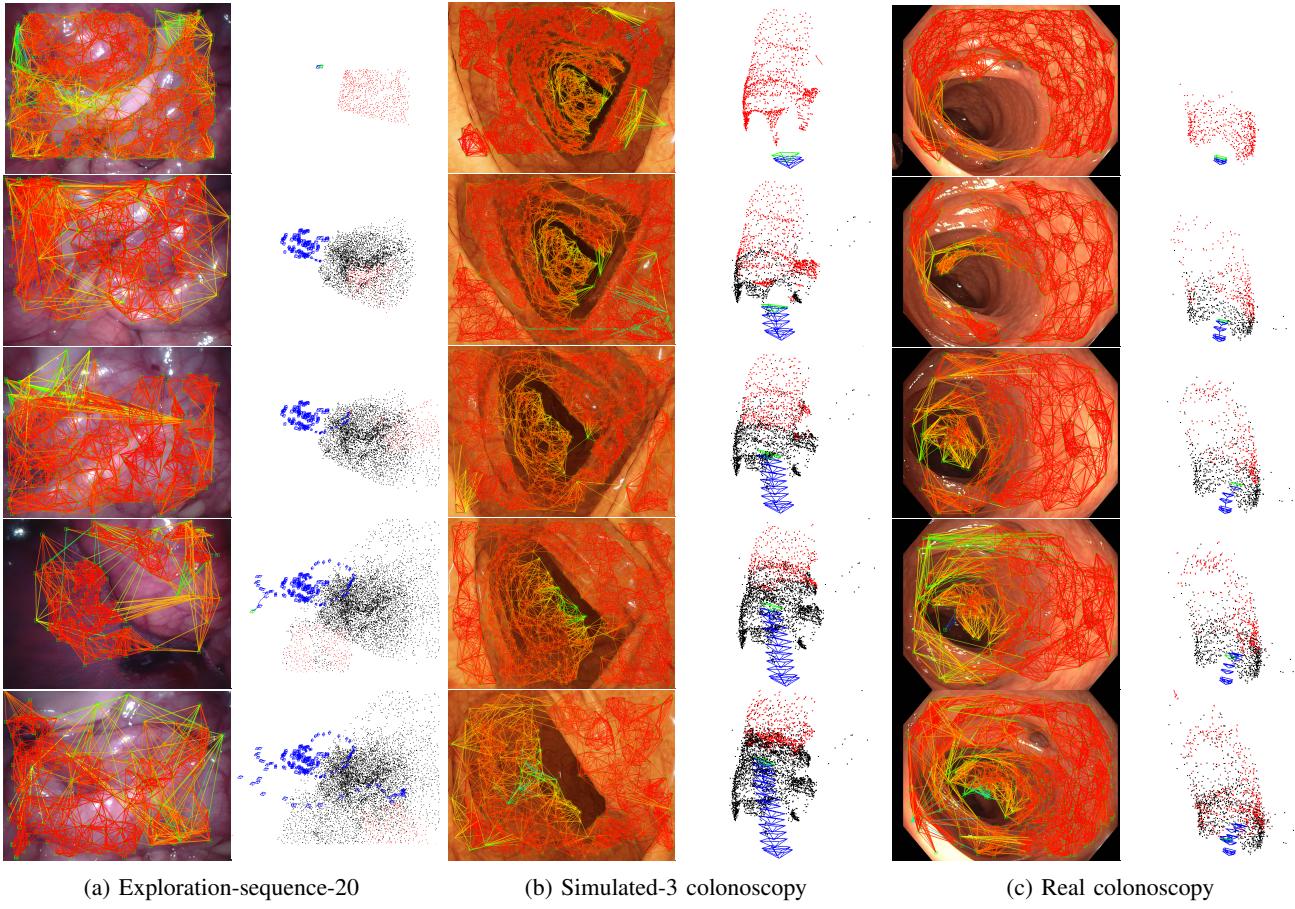(a) Exploration-sequence-20   (b) Simulated-3 colonoscopy   (c) Real colonoscopy

Figure 7: Example reconstructions obtained by NR-SLAM in Hamlyn and two Endomapper colonoscopies.

[7] J. Lamarca, S. Parashar, A. Bartoli, and J. M. M. Montiel, "DefSLAM: Tracking and mapping of deforming scenes from monocular sequences," *IEEE Transactions on Robotics*, vol. 37, no. 1, pp. 291–303, 2020.

[8] J. J. Gómez-Rodríguez, J. Lamarca, J. Morlana, J. D. Tardós, and J. M. M. Montiel, "SD-DefSLAM: Semi-direct monocular SLAM for deformable and intracorporeal scenes," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 5170–5177.

[9] R. Ma, R. Wang, Y. Zhang, S. Pizer, S. K. McGill, J. Rosenman, and J.-M. Frahm, "RNN-SLAM: Reconstructing the 3D colon to visualize missing regions during a colonoscopy," *Medical Image Analysis*, vol. 72, p. 102100, 2021.

[10] J. Lamarca, J. J. Gómez-Rodríguez, J. D. Tardós, and J. M. M. Montiel, "Direct and sparse deformable tracking," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11450–11457, 2022.

[11] J. J. Gómez-Rodríguez, J. M. M. Montiel, and J. D. Tardós, "Tracking monocular camera pose and deformation for SLAM inside the human body," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 5278–5285.

[12] R. W. Sumner, J. Schmid, and M. Pauly, "Embedded deformation for shape manipulation," *ACM Transactions on Graphics*, vol. 26, no. 3, p. article 80, July 2007.

[13] A. Bartoli, Y. Gérard, F. Chadebecq, T. Collins, and D. Pizarro, "Shape-from-template," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 10, pp. 2099–2118, 2015.

[14] A. Chhatkuli, D. Pizarro, A. Bartoli, and T. Collins, "A stable analytical framework for isometric shape-from-template by surface integration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 5, pp. 833–850, 2016.

[15] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3d shape from image streams," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2000, pp. 690–696.

[16] Y. Dai, H. Li, and M. He, "A simple prior-free method for non-rigid structure-from-motion factorization," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 101–122, 2014.

[17] A. Chhatkuli, D. Pizarro, T. Collins, and A. Bartoli, "Inextensible non-rigid shape-from-motion by second-order cone programming," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1719–1727.

[18] S. Parashar, D. Pizarro, and A. Bartoli, "Isometric non-rigid shape-from-motion with riemannian geometry solved in linear time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 2442–2454, 2017.

[19] A. Sengupta and A. Bartoli, "Convex relaxations for isometric and equiareal NRSfM," *arXiv preprint arXiv:2211.16005*, 2022.

[20] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[21] J. Lamarca and J. M. M. Montiel, "Camera tracking for SLAM in deformable maps," in *European Conference on Computer Vision (ECCV) Workshops*, 2018.

[22] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[23] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 593–600.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[25] C. Campos, R. Elvira, J. J. Gómez-Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[26] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *International Conference on Knowledge Discovery and Data Mining (KDD)*, vol. 96, no. 34, 1996, pp. 226–231.

[27] S. H. Lee and J. Civera, "Triangulation: why optimize?" in *30th British Machine Vision Conference (BMVC)*, 2019.

[28] P. Mountney, D. Stoyanov, and G.-Z. Yang, "Three-dimensional tissue deformation recovery and tracking," *IEEE Signal Processing Magazine*, vol. 27, pp. 14–24, 2010.

[29] G. Bradski, "The OpenCV library," *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, vol. 25, no. 11, pp. 120–123, 2000.

[30] H. Strasdat and S. Lovegrove, "Sophus," https://github.com/strasdat/Sophus, 2017, accessed: 2023-06-23.

[31] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g2o: A general framework for graph optimization," in *IEEE International Conference on Robotics and Automation*, 2011, pp. 3607–3613.

[32] K. İncetan, I. O. Celik, A. Obeid, G. I. Gokceler, K. B. Ozyoruk, Y. Almalioglu, R. J. Chen, F. Mahmood, H. Gilbert, N. J. Durr *et al.*, "VR-Caps: a virtual environment for capsule endoscopy," *Medical Image Analysis*, vol. 70, p. 101990, 2021.

Juan J. Gómez Rodríguez received a Bachelor's Degree in Informatics Engineering (mention in Computing) and Master's in Biomedical Engineering (mention in Information and Communication Technologies in Biomedical Engineering) from Universidad de Zaragoza, where he is currently working towards the PhD. degree with the I3A Robotics, Perception and Real-Time Group. His research interests are real-time visual SLAM for both rigid and deformable environments. He received an honorable mention to the King-Sun Fu Memorial IEEE Transactions on Robotics Best Paper Award in 2021, for the paper describing ORB-SLAM3.

J.M.M. Montiel (Arnedo, Spain, 1967) received the M.S. and PhD degrees in electrical engineering from Universidad de Zaragoza, Spain, in 1992 and 1996, respectively. He has been awarded several Spanish MEC grants to fund research with the University of Oxford, U.K., and Imperial College London, U.K.

He is currently a full professor with the Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, where he is in charge of perception and computer vision research grants and courses. His interests include real-time visual SLAM for rigid and non-rigid environments, and the transference of this technology to robotic and non- robotic application domains. He has received several awards, including the 2015 King-Sun Fu Memorial IEEE Transactions on Robotics Best Paper Award and an honorable mention to the same award in 2021. Since 2020 he coordinates the EU FET EndoMapper grant to bring visual SLAM to intracorporeal medical scenes.

Juan D. Tardós (Huesca, Spain, 1961) received the M.S. and Ph.D. degrees in electrical engineering from the University of Zaragoza, Spain, in 1985 and 1991, respectively. He is Full Professor with the Departamento de Informática e Ingeniería de Sistemas, University of Zaragoza, where he is in charge of courses in machine learning and SLAM. His research interests include SLAM, perception and mobile robotics. He received the King-Sun Fu Memorial IEEE Transactions on Robotics Best Paper Award in 2015 for the paper describing the monocular SLAM system ORB-SLAM and an honorable mention to the same award in 2021, for the paper describing ORB-SLAM3.