2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

SD-DefSLAM: Semi-Direct Monocular SLAM for Deformable and Intracorporeal Scenes

Juan J. Gómez-Rodríguez*, José Lamarca*, Javier Morlana, Juan D. Tardós and José M.M. Montiel

Abstract— **Conventional SLAM techniques strongly rely on scene rigidity to solve data association, ignoring dynamic parts of the scene. In this work we present Semi-Direct DefSLAM (SD-DefSLAM), a novel monocular deformable SLAM method able to map highly deforming environments, built on top of DefSLAM [1]. To robustly solve data association in challenging deforming scenes, SD-DefSLAM combines direct and indirect methods: an enhanced illumination-invariant Lucas-Kanade tracker for data association, geometric Bundle Adjustment for pose and deformable map estimation, and bag-of-words based on feature descriptors for camera relocalization. Dynamic objects are detected and segmented-out using a CNN trained for the specific application domain.**

We thoroughly evaluate our system in two public datasets. The mandala dataset is a SLAM benchmark with increasingly aggressive deformations. The Hamlyn dataset contains intracorporeal sequences that pose serious real-life challenges beyond deformation like weak texture, specular reflections, surgical tools and occlusions. Our results show that SD-DefSLAM outperforms DefSLAM in point tracking, reconstruction accuracy and scale drift thanks to the improvement in all the data association steps, being the first system able to robustly perform SLAM inside the human body.

I. Introduction

Simultaneous Localization and Mapping (SLAM) and Visual Odometry (VO) are fundamental blocks for many applications like autonomous robots or augmented reality. Existing methods can be classified as indirect or direct depending of the manner they perform data association. On the one hand, indirect methods estimate 3D geometry from a set of matched keypoints along covisible images, minimizing a geometric error. On the other hand, direct methods avoid extracting features, and work directly on pixel intensities to estimate the 3D geometry, optimizing a photometric error. Finally, semi-direct methods extract features and combine both types of errors.

However, regardless of that classification, all methods rely on a simple, yet important assumption: scene rigidity. This assumption greatly simplifies the SLAM and VO problem and perfectly models many of their application domains. Nevertheless, the increasing interest in Minimally Invasive Surgery (MIS) and medical robots has placed in the spotlight the rigidity assumption, as these kinds of applications work on highly deforming scenarios. That is why a new classification arises as rigid and non-rigid methods, the latter assuming that the 3D position of triangulated landmarks can vary over time.

* Both authors contributed equally to this work.

This work was supported by the EU-H2020 grant 863146: ENDOMAPPER, the Spanish government grants PGC2018-096367-B-I00, DPI2017-91104-EXP and the MINECO scholarship BES-2016-078678, and by Aragón government grant DGA_T45-17R.

The authors are with the Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, María de Luna 1, 50018 Zaragoza, Spain. E-mail: {jjgomez, jlamarca, jmorlana, tardos, josemari}@unizar.es.

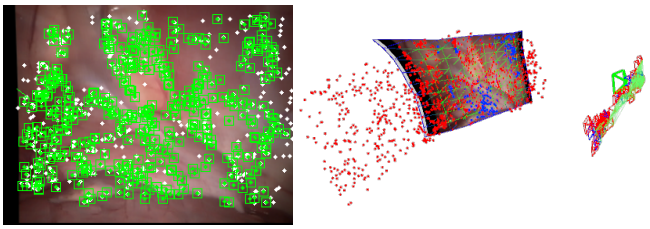

Fig. 1: SD-DefSLAM working on Dataset1 of Hamlyn. Left: in green, features tracked in the endoscopic image using photometric techniques. In white, projected points without a feature match. Right: camera motion and growing deformable map estimated by minimizing geometric error.

In this work, building on DefSLAM [1], we propose SD-DefSLAM, the first deformable semi-direct SLAM system, able to robustly process sequences under great deformations and weak texture, as it is the case of MIS videos. SD-DefSLAM is semi-direct as it extracts ORB features and uses an illumination-invariant Lukas-Kanade (LK) [2] [3] optical flow algorithm to perform data association, minimizing a photometric error, while the camera pose and deforming 3D geometry is estimated minimizing the geometric error (Fig. 1).

In non-rigid SLAM, dynamic objects are difficult to separate from the deforming background using conventional techniques. To achieve robustness, we mask-out moving objects with the help of a convolutional neural network (CNN) specifically trained to segment surgical tools. Finally, we include relocalization capabilities for which we perform long-term data association with ORB descriptors [4] and a bag of words [5], achieving robustness to camera occlusions.

II. Related Work

A. Rigid SLAM and VO

The first real-time SLAM systems followed the indirect approach. MonoSLAM [6] matches a set of sparse keypoints and recovers the scene geometry in an EKF-based framework. This work was later extended in [7] by using an inverse depth parametrization. Later PTAM [8] proposed a parallelization of the main tasks of a SLAM system to allow a Bundle Adjustment (BA) scheme to optimize the 3D geometry. ORB-SLAM [9] is currently the reference system among indirect methods by using the combination of FAST-ORB feature-descriptor [4] and BA to optimize the 3D information. In its successive versions [10, 11] it is extended to different type of sensors, ranging from stereo cameras to wide-lens to inertial sensors.

978-1-7281-9077-8/21/$31.00 ©2021 IEEE

As for direct methods, DSO [12] is the first fully direct VO algorithm that jointly optimizes structure and motion with photometric BA. This work is later extended in DSM [13] by building a direct SLAM algorithm that uses the same photometric model of DSO. While current direct methods are more robust in weakly textured areas, their accuracy degrades in presence of geometric distortions, and they assume photometric invariance, being only able to adapt to global illumination changes [12]. So, they are not applicable in endoscopic images where strong deformations and local illumination changes are prevalent.

Our work is more similar to SVO [14] that proposed a hybrid approach combining direct and indirect methods. SVO is a semi-direct VO method that extracts features in keyframes, uses photometric techniques to perform short-term data association, and ultimately optimizes the reprojection error in a BA.

The crucial novelty of our method is the use of per-feature illumination-invariant photometric data association, instead of the global image alignment used by DSO and SVO, that cannot handle deforming scenes. Our method also allows to obtain medium-term [11] photometric data associations, improving reconstruction accuracy.

B. Deformable SLAM and VO

Many deformable SLAM and VO systems were developed from rigid ones, aiming in many cases to process intracorporeal sequences, as it is a naturally deforming environment of high practical interest for which several datasets exist [15–18]. The first systems that processed this kind of images were [19] and [20], both making use of conventional feature-based SLAM and threshold strategies to differentiate between rigid and non rigid points. Later, ORBSLAM was tuned in [21] and [22] to be able to localize in MIS sequences. The seminal work DefSLAM [1] is the first indirect monocular SLAM system able to tackle with exploration in deformable scenarios. The system grows the map using a sequential Non-Rigid Structure-from-Motion (NRSfM) algorithm based on [23], and estimates at frame rate the deformation occurred and the camera pose by means of a Shape-from-Template (SfT) algorithm [24]. DefSLAM has been proved to work in some simple medical sequences, but the presence of typical challenges like poor texture, illumination changes and tools intrusion, makes it fail.

This evidences the need of more robust data-association methods to process highly deforming environments. In endoscopic sequences this is usually done by correlation matching in consecutive images [19], [20], as feature matching using descriptors such as ORB [4] or SIFT [25] usually do not perform well in low texture regions. AKAZE proposed in [26] is a feature designed to preserve the low texture gradient in the multiscale detector, performing especially well for intracorporeal images. However, it is too slow to be applied in a real-time SLAM algorithm. In [27], a deformable Lucas-Kanade [28] implementation is proposed for tracking tissue surfaces in non-exploratory sequences, including a term that controls the deformation. Deep learning techniques can also play an important role as shown in [29] in which they train a CNN to get dense descriptors in a sinus endoscopy dataset.

Finally, as deformable sequences pose a big challenge for SLAM and VO algorithms, it is essential a better understanding of the scene, identifying and removing dynamic objects that could degrade performance. DynaSLAM [30] uses CNNs to detect, remove and inpaint potentialy dynamic objects such as persons or cars. DOT [31] follows up the ideas from DynaSLAM to only mask-out objects that are actually moving. In the case of endoscopic images, the most typical dynamic objects are surgical tool. Segmentation of this kind of objects is of interest to the scientific community and several methods [32–34] have arisen as response.

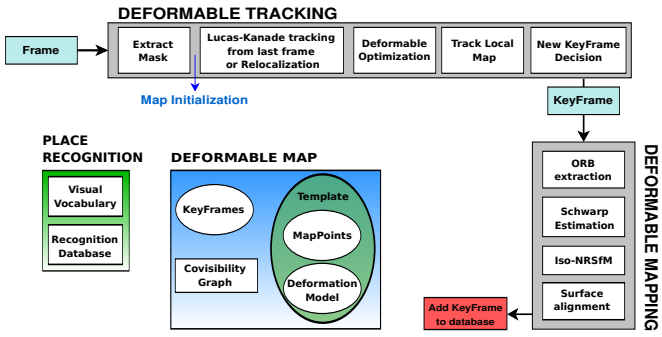

Fig. 2: SD-DefSLAM scheme with a tracking and a mapping thread running concurrently. The main novelties are in the tracking thread, that masks surgical tools using a CNN, achieves robustness with an illumination-invariant photometric method that tracks the previous frame and the local map, and includes bag-of-words relocalization and a new regularizer that smooths camera motion.

III. Semi-Direct DefSLAM

Our approach is called Semi-Direct Deformable SLAM (SD-DefSLAM) as it performs short-term and medium-term data association [11] using a photometric method (subsection III-A) while the deformable optimization backend (subsection III-B) optimizes a geometric error. A global overview of SD-DefSLAM is depicted in Fig. 2. It uses two threads, one for deformable mapping, that progressively builds a growing deformable map and other for deformable tracking, that estimates camera pose and map deformation for each frame processed. Although the main novelties with respect to DefSLAM are in the deformable tracking thread, for the reader convenience, we present here a brief summary of the whole system.

The map is formed by a set of *reference keyframes*, that have observed new parts of the scene as exploration progresses, with an associated surface template. Each template models the observed surface with a triangular mesh that represents its shape-at-rest, whose vertices are the 3D map points. The map also contains a set of *refining keyframes* that are used to refine the templates. Templates are created and refined by the deformable mapping thread at keyframe rate, and their deformation model is estimated by the deformable tracking thread at frame rate. Keyframes are added to a place recognition database [5] to enable relocalization after occlusions.

The deformation mapping thread estimates the surface observed in the reference keyframes and uses refining keyframes to improve this estimation incrementally. Templates are created to grow the map when exploring new places. The core of the deformation mapping is a Non-Rigid Structure-from-Motion (NRSfM) algorithm based on isometry and infinitesimal planarity [23]. It estimates the normal of the points of a keyframe. The points are

initialized assuming smoothness in the surface with respect to the rest of normals estimated and they are refined with each new observation. After estimating the normals, a shape-from-normals algorithm estimates a proportional shape of the surface that fits with those normals. Finally, it performs a SE(3) alignment to recover the correct scale with respect to the rest of the map. This new surface becomes the template for the deformation tracking.

The deformation tracking thread estimates the localization of the camera and the deformation of the 3D map surface at frame rate. The map surface is coded by means of its shape-at-rest and a deformation model. The input of the deformation tracking is the last pose of the camera, the last deformation of the template and the new frame. We use a LK tracker to get initial putative matches, that are computed independently for each point. With the putative matches we estimate an initial deformation of the mesh. This optimization is robust to outliers and give us a better estimation of the position of the points. With these new estimates we reinitialize the LK tracker and search for map points in the observed zone. This allows matches with larger baselines than with a standard LK tracker. In case of tracking lost, we have designed a relocalization module (subsection III-C) able to relocalize the system in this map. For our final application, we have incorporated a CNN that segments tools (subsection III-D) to remove matches in dynamic non-modeled objects.

A. Data Association

For data association, indirect methods rely on good texture to obtain distinctive features, a RANSAC step to enforce rigidity of the set of matchings found, and robust costs functions in BA to reduce the impact of the remaining outliers. In contrast, direct methods use global image alignment that can use pixels with lower texture but rely even more strongly in scene rigidity. In this section we present a photometric data association method that works reliably in low-textured areas, without relying neither in illumination constancy, not in scene rigidity. For this, we use an enhanced Lucas-Kanade (LK) algorithm to perform short-term data association among all the images in the sequence. Our LK algorithm allows us to track low textured surfaces with subpixel accuracy even though there have been local changes in lighting. Next, we describe the basic LK algorithm to better explain the improvements performed to increase accuracy and robustness.

1) Basic Lucas-Kanade algorithm: Let be I and J the reference and the current grayscaled images respectively, $\mathbf{u}=(x,y)^T$ a generic image point found in I and $P(\mathbf{u})$ a squared patch centered on $\mathbf{u}$ of size $(2\omega_x+1)\times(2\omega_y+1)$ pixels. The goal of LK algorithm is to find the optical flow vector $\mathbf{d}=(d_x,d_y)^t$ such us $I(P(\mathbf{u}))$ and $J(P(\mathbf{u}+\mathbf{d}))$ are similar. This is solved using Gauss-Newton gradient descent non-linear optimization:

$$\underset{\mathbf{d}}{\operatorname{argmin}} \sum_{\mathbf{x}\in P(\mathbf{u})} \left(I(\mathbf{x})-J(\mathbf{x}+\mathbf{d})\right)^2 \tag{1}$$

Note that the goal function depends directly on the gray values of both images and the size of the patch ω_x, ω_y.

2) Enhanced Lucas-Kanade algorithm: The basic LK optimization (Eq. 1) depends directly on the raw intensity values of I and J, which makes the LK algorithm very sensitive to illumination changes. While some direct methods address this issue with a global illumination compensation [12], we solve it in a more flexible way using local illumination compensation in a fashion similar to that shown in [3]. In other words, we compute a gain factor α and a bias value β per each tracked patch, which are added in the optimization:

$$\underset{\mathbf{d},\alpha,\beta}{\operatorname{argmin}} \sum_{\mathbf{x}\in P(\mathbf{u})} \left(I(\mathbf{x})-\alpha J(\mathbf{x}+\mathbf{d})-\beta\right)^2 \tag{2}$$

This is especially important when light changes do not occur uniformly across the image, as it happens in outdoor scenes in a cloud-and-clear day, in autonomous car sequences taken during the night, or crucially in endoscopic sequences where the light sources are attached to the endoscope, brightening the image in the areas that get approached, while other areas get darkened. In these cases, global illumination compensation would produce very poor results.

It is also important to keep in mind that the LK algorithm needs the initial guess for $\mathbf{d}$ to be close to the solution in order to converge. That means that if the point to be tracked suffers a big displacement in pixels (for example, due to camera motion or strong deformations) between images, LK may display poor convergence. This can be solved by taking the pyramidal approach proposed in [35] in which the algorithm estimates the optical flow along a pyramidal representation of I and J from the coarsest to the finest level.

Moreover, as we have geometrical estimates of the 3D scene surface and camera poses provided by the SLAM, we can compute an initial guess for $\mathbf{d}$ using that information to improve convergence. For that purpose, assuming local planarity around each tracked point, we can further compute an homography ($\mathbf{h}$) per point that synthesizes the shape of the patch in the new image, yielding the following error term:

$$\underset{\mathbf{d},\alpha,\beta}{\operatorname{argmin}} \sum_{\mathbf{x}\in P(\mathbf{u})} \left(I(\mathbf{x})-\alpha J(\mathbf{h}(\mathbf{x})+\mathbf{d})-\beta\right)^2 \tag{3}$$

Transformation defined by $\mathbf{h}$ compensates any rotation or scale change that the patch could have suffered, making our enhanced LK algorithm rotation and scale invariant. It is also essential to note that computations to synthesize the patch use bilinear interpolation to achieve subpixel accuracy. Now the algorithm guesses for $\mathbf{d}$, can be safely set to 0 because most of the flow is estimated from the available geometry.

Finally, even though LK algorithm converges, it is not guaranteed that it has converged to the correct solution. This can produce spurious feature tracks that negatively affect the overall robustness and accuracy of the algorithm. Most systems address this issue imposing scene rigidity, either in a RANSAC step or with global image alignment. In our case we detect and discard most outliers computing the *Structural Similarity Index* (SSIM) [36] between the reference and the tracked patches. The remaining outliers are successfully handled by a robust influence function in the deformable optimization.

B. Deformable Tracking

Even though our LK algorithm is able to track low textured surfaces using photometric error, due to scene deformation, the matches would be considered spurious if the camera was estimated with a rigid pose-only optimization. Instead, the

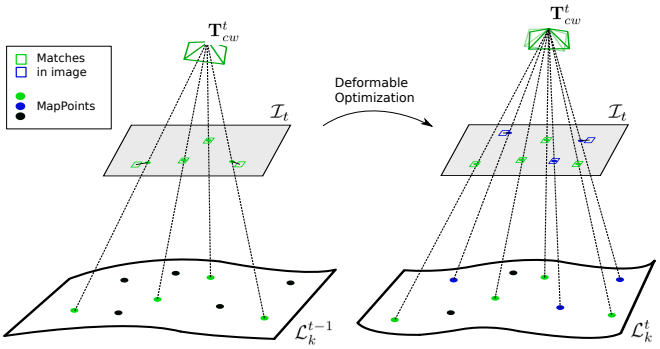

Fig. 3: Two-step data association and deformable tracking optimization.

tracking thread estimates simultaneously the camera pose and the surface deformation minimizing the geometric reprojection error. This dualism leads to the semi-direct name of our algorithm.

More precisely, our deformable tracking thread performs a two-step optimization (Fig. 3) designed to increase SLAM accuracy by reusing the map. For that purpose, as the camera performs exploration, we compute a local map around the current camera pose with covisible keyframes.

The first step aims to compute a first coarse estimation $\mathbf{T}_{cw}^t$ for the camera pose. It obtains putative matches for the points in the previous image (shown in green) using the LK tracker with no geometric information (eq. 2) and runs a first deformable pose optimization. With this early optimization, we also compute the local map $\mathcal{M}$ for the next step.

With the computed camera pose $\mathbf{T}_{cw}^t$ and local map $\mathcal{M}$ from the previous step, we reproject map points from the local map into the current image. Using the projections and the geometrical information from $\mathcal{M}$, we compute an homography $\mathbf{h}$ per projected point and we search its true image position by running our LK tracker with homographies (eq. 3). Finally, with the additional matches found (shown in blue), we run a second deformable pose optimization.

Both deformable optimizations estimate the local map $\mathcal{L}_k^t$ deformation at frame t, along with the camera pose $\mathbf{T}_{cw}^t$, using a modified version of the cost function proposed in [1]:

$$\underset{\mathcal{L}_k^t,\mathbf{T}_{cw}^t}{\operatorname{argmin}} \varphi_d(\mathcal{I}^t,\mathbf{T}_{cw}^t,\mathcal{L}_k^t)+\varphi_e(\mathcal{L}_k^t,\mathcal{T}_k)+\varphi_c(\mathbf{T}_{t,t-1}) \qquad (4)$$

where $\varphi_d(\mathcal{I}^t,\mathbf{T}_{cw}^t,\mathcal{L}_k^t)$ is the total squared reprojection error weighted with a robust Huber influence function, $\varphi_e(\mathcal{L}_k^t,\mathcal{T}_k)$ is the deformation energy of the template $\mathcal{T}_k$ that considers bending, stretching and temporal energy (see [1] for more details), and

$$\varphi_c(\mathbf{T}_{t,t-1})=\xi^T\mathbf{W}\xi \qquad (5)$$

is a new regularization term added to smooth camera motion in frames with low number of matches due to occlusions or sudden deformations, where $\xi=\log(\mathbf{T}_{t,t-1})^\vee\in\mathbb{R}^6$ encodes the translation and rotation between the current and previous frame in the Lie algebra, and $\mathbf{W}$ is the information matrix that controls the degree of smoothing performed.

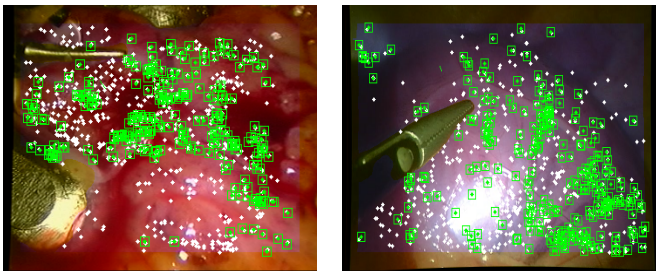

Fig. 4: Frames from Hamlin datasets 4 and 19 showing surgical tools, that are successfully detected and masked-out (yellow color) using semantic segmentation with a CNN.

C. Relocalization

The presence of deformations, really low textured areas, or complete occlusions can lead to system failure. In that context, it is of paramount importance to have a procedure that allows tracking recovery. As in ORB-SLAM, the detection of candidate keyframes for relocalization uses the bag-of-words (BoW) technique from [5], building a database with every keyframe in the sequence, converting them into BoW after extracting ORB descriptors. When the system gets lost, we convert the lost frame into BoW and query the recognition database, obtaining some keyframe candidates. For each keyframe, correspondences associated to map points are computed and then, we obtain an initial camera pose with PnP, performing RANSAC iterations. The main difference with the rigid case is that the inlier threshold has been increased to allow points with some deformation. If PnP is successful, we retrieve the template associated with the candidate keyframe and perform a deformable optimization, optimizing both the template and the camera pose. Tracking continues with this retrieved template. Although our method only works under mild deformations, as PnP is constrained by (weakened) rigidity, it is able to successfully solve the typical short-time occlusions appearing in endoscopies.

D. Moving Objects

In conventional SLAM, moving objects can be successfully detected as their motion is not consistent with the motion of the rest of the scene, except if they move too slowly. However, in a deformable scenario, separating object motion from scene deformation is far from trivial using just geometric information. Matches coming from moving objects lead to severe errors in scene deformation or even to total SLAM failure. We propose to solve this issue using semantic information with a CNN trained to identify and segment the typical moving objects in each application domain, masking the corresponding image regions to avoid matching features in them.

To segment surgical tools in medical scenes we use the CNN defined and trained in [37]. The network is directly integrated in the system and computes a mask for each incoming image. The mask is finally dilated to avoid keypoint detection in the borders of tools. In Fig. 4, we show examples of the masks obtained in two different sequences.

If the tool occludes large parts of the image, the camera pose estimation will become an ill-conditioned problem. For this reason, we constraint the camera motion with a smooth motion

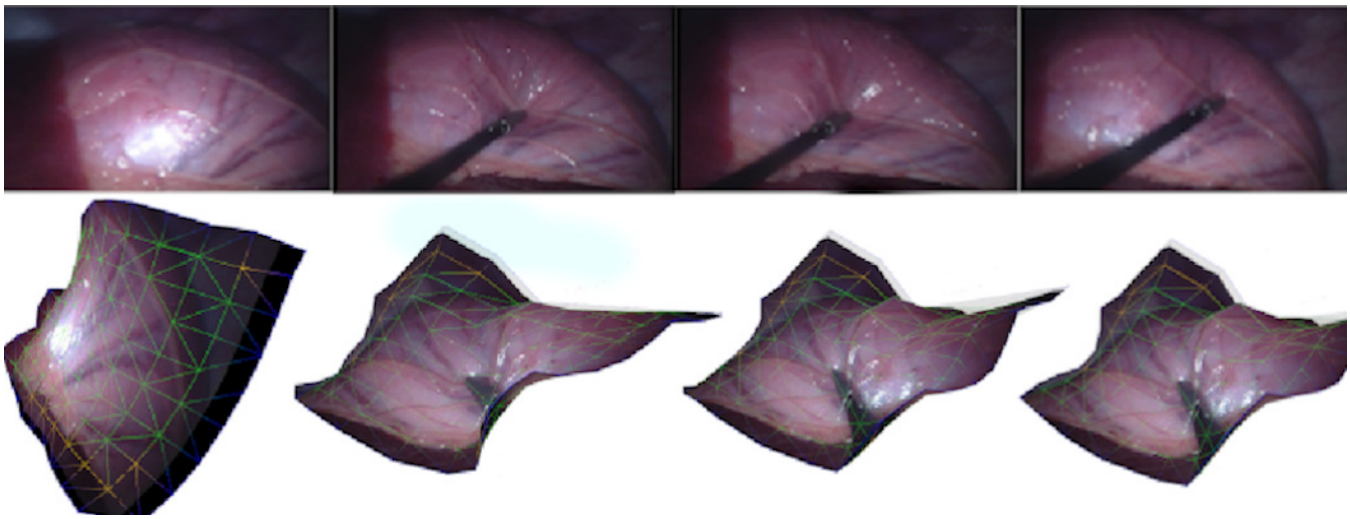

Fig. 5: Examples of the reconstructed surfaces in the *Sequence_organs*. Note how we reconstruct the deformation produced by medical tools. Top is the frame inserted, bottom the 3D reconstruction. From right to left: Frames #315, #1010,#1030,#1055

TABLE I: Comparison in Mandala Dataset.

	DefSLAM [1]		SD-DefSLAM	
	RMSE (mm)	Scale drift	RMSE (mm)	Scale drift
Mandala0	26.3	1.06	**23.1**	**1.03**
Mandala1	22.3	1.44	**21.3**	**1.32**
Mandala2	17.9	1.46	**16.1**	**1.41**
Mandala3	43.7	2.07	**41.8**	**1.26**
Mandala4	55.6	1.78	**48.1**	**1.27**

TABLE II: Comparison in Hamlyn Dataset

	DefSLAM		SD-DefSLAM	
	RMSE (mm)	Scale drift	RMSE (mm)	Scale drift
f5	5.00	1.01	**3.00**	0.99
f7	4.50	0.99	**4.35**	0.99
Seq_heart	3.84	2.00	**1.17**	**1.32**
Seq_abdominal	23.98	0.98	**22.2**	**1.01**
Seq_organs	13.02	1.27	**6.63**	**1.05**
Seq_exploration	17.02	2.60	**12.56**	**1.36**

prior. When the occlusion is complete, tracking is lost and the system relies on relocalization.

IV. Experiments

We have evaluated the proposed system and compared it with DefSLAM [1] in two datasets. The first one is the Mandala dataset from [1], created to evaluate deformable SLAM. The purpose of this dataset is to evaluate the performance of the system in a controlled environment with good texture and illumination conditions. Secondly, we further validated our system in several medical sequences of the Hamlyn dataset which pose a substantial challenge to SLAM algorithms. Our method is pure monocular, but in both cases, we selected datasets recorded with stereo cameras to extract a ground truth solution for the scene surface with the Libelas algorithm [38]. We analyze the 3D RMS error of the reconstruction, by means of the Euclidean distance between the ground truth and the scaled reconstruction of the system. We estimate the relative scale of the reconstruction for each frame, and report the scale drift observed along the trajectory. The scale drift is the quotient of the mean scale of the last 20 frames and the scale of the first 20 frames. We also provide a data association quality to compare the performance of the feature matching technique in DefSLAM with the new semi-direct technique that uses photometric information and gives subpixel accuracy.

A. Mandala dataset

The Mandala dataset consists of 5 sequences exploring a mandala kerchief that goes from a totally rigid situation (Mandala0) to a intensively deforming one (Mandala4). The kerchief is hanged and deformed creating waves that go through it. The intensity of the deformation is measured depending on the speed and amplitude of the waves.

Table I shows that SD-DefSLAM outperforms DefSLAM in all Mandala sequences, both in RMS reconstruction error and in scale drift. While in the most rigid sequence (Mandala0) the improvement is marginal, for those sequences with more aggressive deformations (Mandala3 and Mandala4), SD-DefSLAM achieves a significant improvement.

B. Medical scenes

We have evaluated our system in several laparoscopic scenes of the Hamlyn dataset. These sequences present a huge variety of scenarios, including phantom hearts with per-frame ground truth from computed tomography (Dataset11-f5 and Dataset12-f7 in Hamlyn [15, 16]), a non-exploratory heart sequence with tool intrusions (Dataset4 - Sequence_heart in Hamlyn [17]) and three exploratory sequences (Dataset1 - Sequence_abdominal, Dataset19 - Sequence_organs and Dataset20 - Sequence_exploration in Hamlyn [18]).

The improved data association enables our system to better compute the map deformation, improving the RMSE and scale drift, as shown in Table II. The addition of a CNN to mask out surgical tools in the Sequence_heart and Sequence_organs allows our system to robustly process the sequences with significant

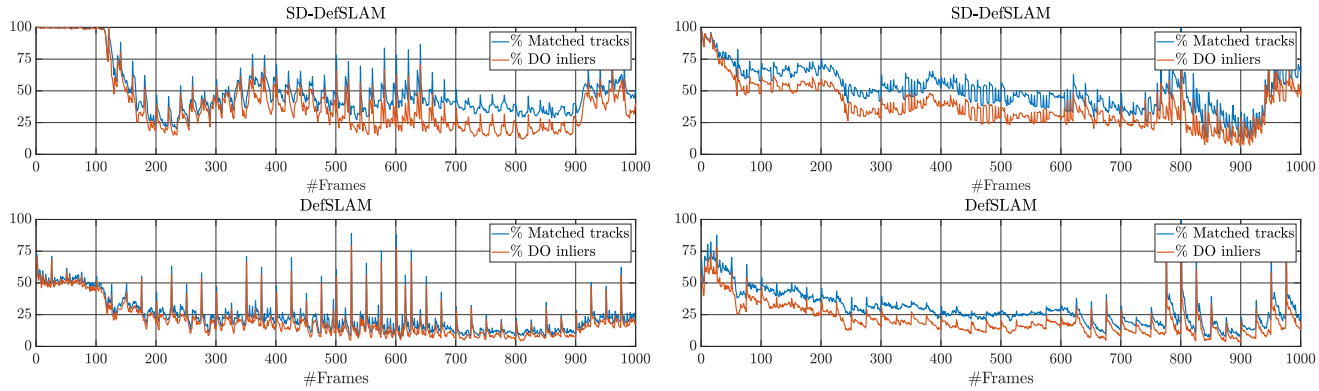

Fig. 6: Percentage of points in the local map that are tracked and that are considered inliers after deformable optimization in Mandala3 (left) and Hamlyn Dataset20 (right).

improvement in the performance. An example of the reconstructed surfaces under deformations is shown in Fig. 5.

C. Data association

The results in the last sections show how SD-DefSLAM outperforms DefSLAM in all the tested datasets. One of the keys for the improvement is the data association which is a fundamental part for both the tracking and the mapping. For the tracking, better association leads to better estimation of the deformations in the map. Concerning the mapping, longer tracks between keyframes speed up the convergence of NRSfM.

In this section, we analyse the proposed data association scheme. There are two key differences *wrt.* the original method. The most evident one is that the matching is performed photometrically, reaching subpixel accuracy. The other one is that each patch is initialized with the keyframes and actively tracked in the consecutive images, removing feature extraction from the matching stage. This is significant as FAST features have low repeatability between temporarily close images, impairing SLAM performance.

Figure 6 depicts a comparison between the SD-DefSLAM photometric data association (top) and the ORB matching of DefSLAM (bottom) in Mandala3 and Hamlyn Dataset20. In both cases, the percentage of matched map points (matched tracks) is shown in blue and the inliers after the deformation optimization (DO inliers) in orange. In both figures, our system gets more inliers than DefSLAM after the optimization, which intuitively shows the better matching efficiency.

In Mandala3 sequence, SD-DefSLAM doubles the percentage of correct matches obtained by DefSLAM. This greatly improves the overall robustness of the system at the same time that improves the accuracy. Dataset 20 poses a bigger challenge as the combination of low texture and image blurring penalizes both types of data association algorithms, but the new method is still clearly superior. This, together with the subpixel accuracy explains the more accurate reconstruction and smaller scale drift obtained (last row in Table II).

D. Relocalization

Besides the robustness of the system to tools or low texture, the camera still can get totally occluded or even the endoscope must leave the scene to clean the optics. Thanks to the relocalization module, we were able to relocalize the system after a tracking failure. In contrast with DefSLAM, which cannot manage tracking lost, we were able to process more frames in the proposed sequences, as shown in Fig. 7.

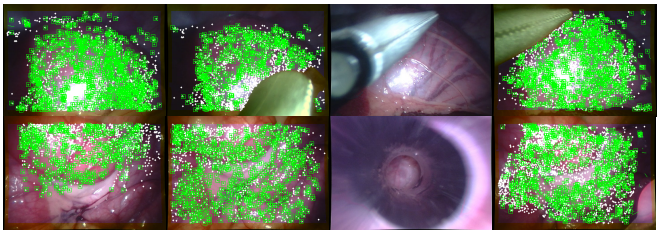

Fig. 7: First row: relocalization due to tool intrusion. The CNN is able to detect the tool correctly, but when it occludes most of the image, the system fails and performs relocalization. Second row: the endoscope is extracted from the scene to clean it, and the system relocalizes once it is introduced again in the body.

V. Conclusions

While rigid SLAM is mature, deformable environments pose serious challenges requiring to re-think all data association steps. We have shown that a semi-direct approach based on per-feature illumination-invariant photometric tracking greatly improves data association, reconstruction accuracy and scale drift. Its combination with CNN segmentation to detect moving objects, and relocalization capabilities to deal with occlusions, gives the first SLAM system able to robustly address the real-life challenges of medical sequences.

Our deformable model assumes isometric deformations. This is quite a restrictive assumption that is not always fulfilled as is the case of MIS sequences. This causes a worsening in the estimation of the deformation which in turn affects the quality of the data association. This can be addressed by exploring new deformation models that properly represent non-isometric deformations.

References

[1] J. Lamarca, S. Parashar, A. Bartoli, and J. M. M. Mon-

tiel, "DefSLAM: Tracking and mapping of deforming scenes from monocular sequences," *IEEE Transactions on Robotics*, vol. 37, no. 1, pp. 291–303, 2021.

[2] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Intl. Joint Conf. on Artificial Intelligence (IJCAI)*, 1981.

[3] S. Negahdaripour, "Revised definition of optical flow: Integration of radiometric and geometric cues for dynamic scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 9, pp. 961–979, 1998.

[4] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *IEEE Int. Conf. on Computer Vision*, 2011, pp. 2564–2571.

[5] D. Gálvez-López and J. D. Tardós, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. on Robotics*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012.

[6] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[7] J. Civera, A. J. Davison, and J. M. M. Montiel, "Inverse depth parametrization for monocular SLAM," *IEEE Trans. on Robotics*, vol. 24, no. 5, pp. 932–945, 2008.

[8] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *IEEE and ACM Int. Symp. on Mixed and Augmented Reality*, 2007, pp. 225–234.

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

[10] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[11] 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 multi-map SLAM," *arXiv preprint arXiv:2007.11898*, 2020.

[12] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, 2017.

[13] J. Zubizarreta, I. Aguinaga, and J. M. M. Montiel, "Direct sparse mapping," *IEEE Transactions on Robotics*, vol. 36, no. 4, pp. 1363–1370, Aug. 2020.

[14] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2014, pp. 15–22.

[15] D. Stoyanov, M. V. Scarzanella, P. Pratt, and G. Z. Yang, "Real-time stereo reconstruction in robotically assisted minimally invasive surgery," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2010, pp. 275–282.

[16] P. Pratt, D. Stoyanov, M. Visentini-Scarzanella, and G. Z. Yang, "Dynamic guidance for robotic surgery using image-constrained biomechanical models," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2010, pp. 77–85.

[17] D. Stoyanov, G. P. Mylonas, F. Deligianni, A. Darzi, and G. Z. Yang, "Soft-tissue motion tracking and structure estimation for robotic assisted mis procedures," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2005, pp. 139–146.

[18] 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.

[19] O. G. Grasa, J. Civera, and J. M. M. Montiel, "EKF monocular SLAM with relocalization for laparoscopic sequences," in *IEEE Int. Conf. on Robotics and Automation*, 2011, pp. 4816–4821.

[20] B. Lin, A. Johnson, X. Qian, J. Sanchez, and Y. Sun, "Simultaneous tracking, 3D reconstruction and deforming point detection for stereoscope guided surgery," in *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*, Springer, 2013, pp. 35–44.

[21] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler, J. Marescaux, and J. M. M. Montiel, "ORBSLAM-based endoscope tracking and 3D reconstruction," in *Int. Workshop on Computer-Assisted and Robotic Endoscopy*, Springer, 2016, pp. 72–83.

[22] N. Mahmoud, A. Hostettler, T. Collins, L. Soler, C. Doignon, and J. M. M. Montiel, "SLAM based quasi dense reconstruction for minimally invasive surgery scenes," *arXiv preprint arXiv:1705.09107*, 2017.

[23] S. Parashar, D. Pizarro, and A. Bartoli, "Isometric non-rigid shape-from-motion with riemannian geometry solved in linear time," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2442–2454, 2017.

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

[25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[26] P. F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, 2011.

[27] X. Du, N. Clancy, S. Arya, G. B. Hanna, J. Kelly, D. S. Elson, and D. Stoyanov, "Robust surface tracking combining features, intensity and illumination compensation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, no. 12, pp. 1915–1926, 2015.

[28] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[29] X. Liu, Y. Zheng, B. Killeen, M. Ishii, G. D. Hager, R. H. Taylor, and M. Unberath, "Extremely dense point correspondences using a learned feature descriptor," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4847–4856.

[30] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes,"

IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 4076–4083, 2018.

[31] I. Ballester, A. Fontan, J. Civera, K. H. Strobl, and R. Triebel, "DOT: Dynamic object tracking for visual SLAM," *arXiv preprint arXiv:2010.00052*, 2020.

[32] I. Laina, N. Rieke, C. Rupprecht, J. P. Vizcaíno, A. Eslami, F. Tombari, and N. Navab, "Concurrent segmentation and localization for tracking of surgical instruments," in *Int. Conf. on medical image computing and computer-assisted intervention*, Springer, 2017, pp. 664–672.

[33] T. Kurmann, P. M. Neila, X. Du, P. Fua, D. Stoyanov, S. Wolf, and R. Sznitman, "Simultaneous recognition and pose estimation of instruments in minimally invasive surgery," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2017, pp. 505–513.

[34] D. Pakhomov, V. Premachandran, M. Allan, M. Azizian, and N. Navab, "Deep residual learning for instrument segmentation in robotic surgery," in *International Workshop on Machine Learning in Medical Imaging*, Springer, 2019, pp. 566–573.

[35] J. Y. Bouguet, "Pyramidal implementation of the affine Lucas Kanade feature tracker. Description of the algorithm," Intel corporation, Tech. Rep., 2001.

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

[37] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *IEEE Int. Conf. on Machine Learning and Applications (ICMLA)*, 2018, pp. 624–628.

[38] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Asian Conference on Computer Vision (ACCV)*, 2010.