# EndoAbS Dataset: Endoscopic Abdominal Stereo Image Dataset for Benchmarking 3D Stereo Reconstruction Algorithms

Veronica Penza[⋆] [†1,2], Andrea S. Ciullo[⋆2], Sara Moccia[1,2], Leonardo S. Mattos[1], and Elena De Momi[2]

[1] *Department of Advanced Robotics, Istituto Italiano di Tecnologia , via Morego, 30, 16163 Genova, Italy*

[2] *Department of Electronics Information and Bioengineering, Politecnico di Milano, P.zza L. Da Vinci, 32, 20133 Milano, Italy*

**Abstract**

*Background* 3D reconstruction algorithms are of fundamental importance for Augmented Reality (AR) applications in computer-assisted surgery. However, few datasets of endoscopic stereo-images with associated 3D surface references are currently openly available, preventing the proper validation of such algorithms. This work presents a new and rich stereo endoscopic image dataset (*EndoAbS* dataset).

*Methods* The dataset includes: (i) Endoscopic stereo images of phantom abdominal organs; (ii) 3D organ surface Reference (RF) generated with a laser scanner; (iii) Camera calibration parameters. It is also provided a detailed description of the phantom generation and the camera-laser calibration method.

*Results* An estimation of the dataset creation overall error is reported (camera-laser calibration error $0.43mm$) and the performance of a 3D reconstruction algorithm is evaluated using *EndoAbS*, resulting in an accuracy error in accordance with state-of-the-art results ($< 2mm$).

⋆Authors equally contributed to this work.

† corresponding author - email: veronica.penza@iit.it

**Manuscript Information**

*Financial support:* This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

*Manuscript category:* Original Article.

*Word Count:* 4785 words

*Number of Figures: 11*

*Number of Tables: 7*

*Conclusions EndoAbS* dataset contributes to increase the number and variety of openly available surgical stereo-image datasets, including a highly accurate RF and different surgical conditions.

# 1 Introduction

In Minimally Invasive Surgery (MIS), the application of augmented reality systems is aimed at improving the outcome of surgery by intra-operatively enhancing the surgeon's perception and providing guidance inside the patient's body. Indeed, these systems provide the surgeon with additional useful information coming from a pre-operative planning, which fused into the intra-operative scenario can, for example, help in the localization of a tumor area, as described in [14], [2], [15] and [16]. However, during the surgery, the organs' geometry is constantly changing due to breathing, heart beating and tissue-instrument interaction, making the update of the registration of augmented reality features very challenging.

3D reconstruction algorithms can be integrated in such systems to retrieve the geometry of soft tissue surfaces intra-operatively, with the aim of measuring the surgical site deformation in real time [23]. These methods have the potential to replace the usage of intra-operative Computer Tomography (CT) or Magnetic Resonance Imaging (MRI), and overcome their drawbacks (such as non-real-time information, patient radiation exposure and high costs) by just exploiting only the images captured from a stereo-endoscope. Despite the performance of these algorithms is well established in different fields, such as domestic, industrial robots and game industry [21], their application to surgical endoscopic images has been proved to be challenging due to the peculiarities that a surgical scenario presents, such as homogeneous or periodic tissue texture, non-uniform illumination, presence of specular reflections for non-Lambertian tissue behaviour, blood and smoke caused by tissue cauterization. Thus, a proper evaluation on specific surgical endoscopic datasets is of special importance to assess their accuracy and robustness.

The evaluation is typically performed comparing the resulting point cloud against a 3D surface reference (in this paper referred as RF), assumed to correspond or, at least, to be close to the real solution [8]. Unfortunately, even if there are many stereo datasets representing static indoor scenes [21, 22], only few datasets providing surgical endoscopic images with an associated RF are publicly available (see Tab. 1). In [20], authors presented synthetic stereo-images and the corresponding RF, taken from a virtual model of liver by using a simulated stereo endoscope. In [24] and [19], it is proposed a stereo-image dataset of a moving heart phantom (Chamberlain Group, MA, USA), generated using the da Vinci® surgical system, providing a CT reference data[1]. More recently, a dataset of stereo-images of ex-vivo animal organs (liver, heart and kidneys) is presented in [11], providing a CT scanner-based RF and exploring different conditions, such as presence of blood and smoke, as well as different poses of the

[1] available at http://hamlyn.doc.ic.ac.uk/vision/

endoscope. These datasets [2] have been used for validating and benchmarking different 3D reconstruction algorithms, as summarized in Tab. 1.

Having in mind all these aspects, we can state that a surgical endoscopic dataset to be used for the evaluation of 3D reconstruction algorithms should present the following characteristics:

1. It should be made of stereo images, associated RF, camera calibration parameters and errors involved in the RF creation process that can affect the algorithms evaluation;
2. The images should present the main characteristics of real endoscopic surgical scenarios, mentioned before;
3. It should be publicly available in order to allow validation and benchmarking of image processing and computer vision algorithms.

[2] available at http://open-cas.com/

Table 1: Openly available surgical endoscopic datasets

| **Surgical Dataset** | | | | |
|---|---|---|---|---|
| Surgical scenario | Organ | RF | Characteristics | Refs |
| virtual phantom | liver | 3D model | 3 liver texture, endoscope-tissue of $5cm$, 360 endoscope rotation with 5 steps, zoom in and zoom out of the same liver spot (max zoom $40mm$ with $2mm$ step), tissue deformation | Hu et al., 2007 [7] Rohl et al., 2012 [20] Mountney and Yang, 2010 [13] |
| ex-vivo organs | porcine liver, kidney, heart, fatty tissue | CT scan | different illumination levels, smoke and blood presence, two endoscope-tissue distances ($5cm$ and $7cm$), two endoscope orientations angles (0 and 30 degrees) | Maier et al., 2014 [11] Lin et al., 2015 [9] |
| phantom organs | heart | CT scan | two views of a beating heart | Stoyanov et al., 2010 [24] Pratt et al., 2010 [19] Penza et al., 2015 [17] |

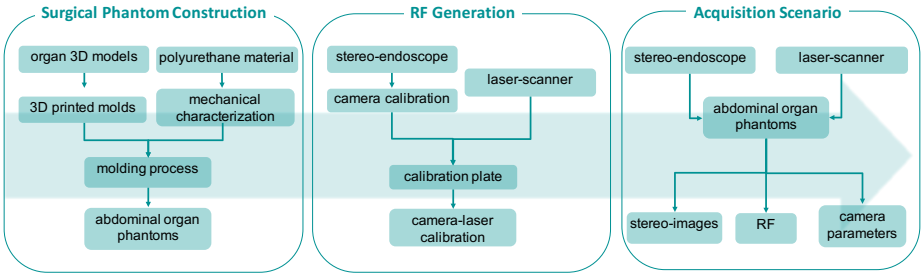



Figure 1: Workflow for the dataset generation. *Surgical Phantom Construction*: The phantom organs were designed using the 3D models of the liver, spleen and kidneys provided by 3DIRCADb and a mechanical characterization study was conducted to define the stiffness of the material used to create the phantoms. *RF Generation*: A camera-laser calibration method was developed to register the RF in the left camera reference system of the stereo-endoscope, in order to allow the algorithm validation. *Acquisition Scenario*: The surgical stereo-endoscopic dataset was generated, consisting in stereo-images, RF and camera parameters.

The ideal setup to obtain realistic images would be a real surgical scenario. However, measuring the RF during a surgical procedure is impractical due to the narrow access space to the operative field and the difficulties in performing a CT scan. For these reasons, synthetic data [7, 13, 20], phantoms [7, 19, 24] and ex-vivo organs [10, 11, 13] have been exploited to reproduce the surgical site. However, these methods present some issues: In the case of simulated data the conditions are too far away from the reality; In the case of ex-vivo organs, in order to preserve the shapes of the organs between RF scan and the images acquisition, the organs has to be kept in specific conditions as long as possible (in water and at low temperature), causing timing constrains during the experiments; In the case of organ phantoms, the main difficulties are related with the reproduction of the appearance and tissue mechanical properties (if tissue deformations are also simulated). In the latter two cases, another constraint is associated to the availability in research laboratories of CT scanner or laser scanner (used to generate the RF) due to their high cost.

Considering the increasing necessity of surgical stereo image datasets, the aim of this work is the generation of an Endoscopic Abdominal Stereo image dataset (*EndoAbS*) for 3D stereo-reconstruction algorithms validation, specifically focusing the attention on the evaluation of passive stereo reconstruction methods. *EndoAbS* dataset is composed of 120 stereo-images of phantoms of different abdominal organs, showing either flat organ surfaces (spleen), or more complex structures as vessels in liver and kidney. The different shape and texture of the organs, the variation of lighting conditions and the simulation of the presence of smoke, make the dataset useful to test the robustness of 3D stereo-reconstruction algorithms under different conditions. Each pair of images is

coupled with its RF, that was obtained using a high-resolution laser scanner. In order to encourage the generation of additional datasets, the paper provides a detailed description of the phantom generation process and, of the method used to refer the RF in the camera reference system (camera-laser calibration) and its accuracy performance. Moreover, in order to exemplify the usage of *EndoAbS* dataset, the performance of a 3D reconstruction algorithm, previously implemented by the authors, was evaluated using the proposed protocol evaluation.

With respect to the already openly available dataset, *EndoAbs* dataset is proposed to provide: (i) a higher numerosity of stereo images; (ii) a wider variety of tissue and organs' shape, ranging from smooth surface to more complex structure, as vessels; (iii) a high accurate RF acquired using a laser scanner; (iv) the description of an accurate markerless method for registering the RF with the reconstructed point cloud. This dataset and the camera-laser calibration code is openly available on-line for the benefit of the computer assisted surgery community[3]. A preliminary description of the *EndoAbs* dataset is presented in [3].

The paper is structured as follows: in Section 2, the workflow for the dataset generation is described, considering the abdominal phantom construction and the RF generation process with a description of the camera-laser calibration procedure. In Section 3, the experimental setup to validate the dataset generation errors is presented and results are shown in Section 4. The evaluation and results of a 3D stereo reconstruction algorithm are also presented, in order to assess the usability of the proposed dataset. Finally, conclusions and open issues are reported in Section 5.

# 2 Material and Methods

*EndoAbS* dataset was generated capturing the stereo images and the corresponding RF of a surgical scenario represented by phantom abdominal models. The images were captured using a stereo-endoscope made of 2 Ultra Mini CMOS analogical Color Cameras (MISUMI, Taiwan) with $640 \times 480$ pixels resolution, with a baseline of $6mm$, and two white LEDs. Two frame grabbers (GRABBY, TERRATEC, Alsdorf) were used to acquire the stereo images. The RF provided in the dataset is in the form of a point cloud and it represents the 3D surface of the surgical scenario as close as possible to the real values. It was generated using the laser scanner VIVID 910 (accuracy[4] of $x = \pm 0.22mm$, $y = \pm 0.16mm$, $z = \pm 0.07mm$ and a precision of $8\mu m$) and the software Polygon Editing Tool (KONICA MINOLTA).

The generation process of *EndoAbS* dataset, mainly involving the (i) construction of a phantom abdominal model, the (ii) RF generation and (iii) the acquisition scenario is described in detail in the following sections and it is shown in Fig. 1.

[3] http://nearlab.polimi.it/medical/dataset/

[4] Conditions: distance 0.6m, temperature 20°C, relative humidity 65%

Figure 2: On the left, spleen, kidneys and a liver detail are shown; on the right, the ribcage containing the organs is shown.

## 2.1 Surgical phantom construction

Liver, spleen and two kidneys were created through a moulding process as in [4], and a ribcage-like support was 3D printed to maintain the relative position between the organs, as shown in Fig. 2. The steps of the process are shown in Fig. 4 and described in the following sub-sections.

### 2.1.1 3D organ model and mold generation

The 3D models of the organs and ribcage were taken from 3D-IRCADb[5]. The 3D-IRCADb includes anonymized DICOM CT medical images (voxel size: $0.96mm \times 0.96mm \times 2.4mm$) with an associated manual segmentation performed by expert clinicians, and an organ surface model stored in VTK format, as shown in Fig. 4(a). 3D virtual negative molds were modelled using the software Blender 2.7.4 (Blender Foundation, Amsterdam), as shown in Fig. 4(b). The virtual molds were 3D printed in acrylonitrile butadiene styrene (ABS), using the Elite Dimension 3D printer (layer thickness: $0.25mm$), see Fig. 4(c).

### 2.1.2 Polyurethane organ phantom

We decided to recreate soft phantoms of abdominal organs with the aim of representing the surgical scenario as close as possible to the real one. This characteristic will also permit a future improvement of the dataset with tissue-instrument interaction images. To this end, a bi-component polyurethane elastomer (F-105 A/B 5 shore, from BJB Enterprise) was combined with a softening agent (SC-22, from BJB Enterprise) in order to modify the elastomer stiffness and match approximately the real tissue characteristics. We considered different stiffness values for liver tissue reported in the literature: $1.3kPa$ [25], 0.90 to $1.730kPa$ [28], $2.0kPa$ [12]. However, since the measured viscoelastic properties can vary depending on experimental conditions and on the used testing method [12], we decided to perform a compressive mechanical test comparing the results obtained from a cylindrical sample ($height = 15mm$, $diameter = 28.2mm$) of porcine liver against samples of polyurethane made with different percentage of

[5] http://www.ircad.fr/research/3dircadb/

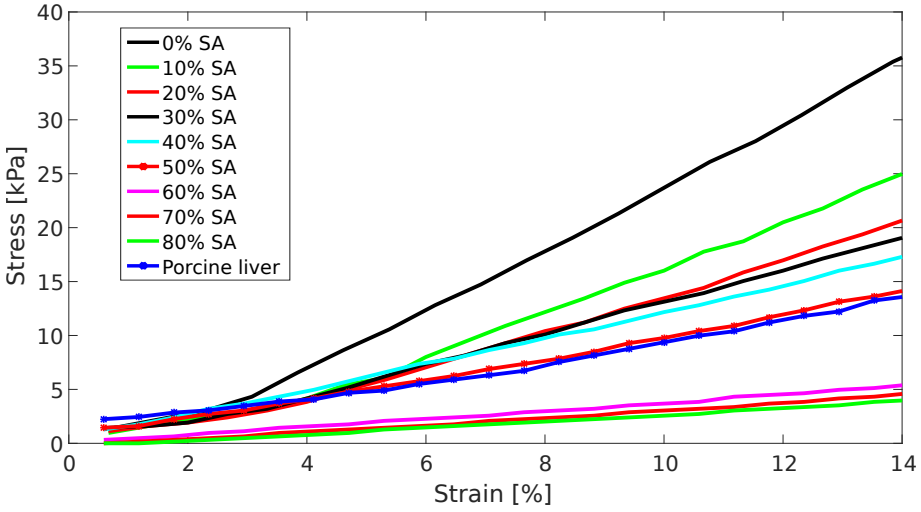


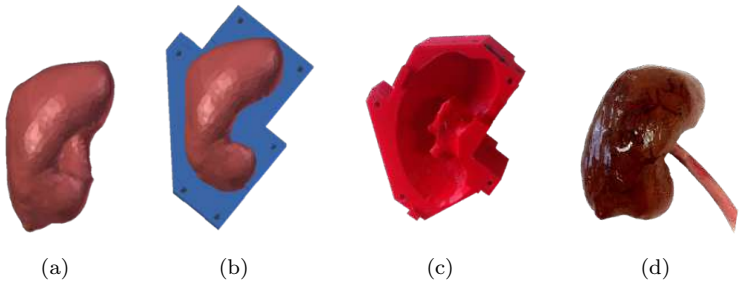

Figure 3: Stress-strain curves for each phantom sample with different percentage of softening agent and the liver sample used as template.

softening agent (from 0% to 80% with steps of 10%). The compressive mechanical test was done with a testing machine (EASYDUR DYNO), compressing the samples until $2.5mm$, with discrete steps of $0.1mm$. Each trial was performed from a starting configuration in which the piston was in contact with the sample, introducing a pre-strain of $0.1mm$ on the samples. Consequently, the stress-strain curves (see Fig.3) and the Young's modulus for each sample was computed, allowing to find the right percentage of softening agent.

Furthermore, the organs were painted with acrylic colors to simulate the tissue superficial texture, with the aid of a sponge, and small vessels using acrylic markers with fine tip, as shown in Fig. 4(d). In the liver and kidney phantoms, plastic tubular structures were attached on the surface and painted to represent main vessels, as it is shown in Fig. 2. A transparent ultrasound gel was laid on the surface of the organs to reproduce the typical wet surface, and thus the specular highlights in the images.

## 2.2 RF generation

In order to compare the reconstructed point cloud with the RF, they both have to be in the same reference system. For this reason, a camera-laser calibration method for estimating the geometrical transformation between the laser and the left camera of the stereo endoscope was developed. We chose the left camera since it is standardly used as the reference system in 3D reconstruction algorithms.



(a) (b) (c) (d)

Figure 4: Example of moulding process for the creation of kidney phantom: (a) 3D virtual model from 3D-IRCADb CT database; (b) 3D virtual negative molds; (c) 3D printed negative mould; (d) polyurethane kidney phantom.

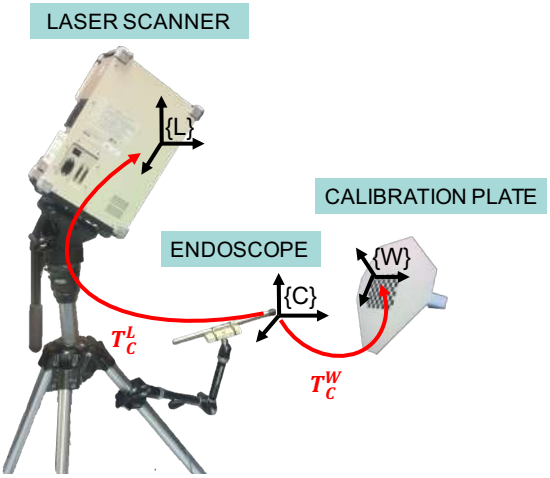



Figure 5: Camera reference system $\{\mathbf{C}\}$, Laser reference system $\{\mathbf{L}\}$ and chessboard reference system $\{\mathbf{W}\}$ are the reference systems involved in the camera-laser calibration. $\mathbf{T_C^W}$ is the transformation from $\{\mathbf{C}\}$ to $\{\mathbf{W}\}$; $\mathbf{T_C^L}$ is the unknown transformation from $\{\mathbf{C}\}$ to $\{\mathbf{L}\}$.

### 2.2.1 Camera-Laser calibration

The camera-laser calibration method consists in computing the rigid transformation between the same set of points measured in the laser scanner and in the left camera reference systems, $\{\mathbf{L}\}$ and $\{\mathbf{C}\}$ respectively. For the sake of clarity, the setup, the reference systems and the geometrical transformation involved in this method are summarized in Fig. 5.

In order to perform this calibration, it is necessary to use a custom target which corners can be identified both by the laser, as 3D geometrical features, and

the camera, as 2D visual information. To this end, an asymmetrical octagonal calibration plate was designed, which vertices $\mathbf{p}_{vert}$ were used as the set of points for the calibration process, as shown in Fig. 6.

In order to improve the manual selection of the vertices in $\{\mathbf{C}\}$ ($\mathbf{p}^{\mathbf{c}}_{vert}$), a standard square chessboard ($7 \times 11$, square size: $2.5mm$) was placed on the calibration plate. Knowing the relative location of the plate vertices $\mathbf{p}_{vert}$ with respect to the chessboard, it is possible to computer the position of $\mathbf{p}^{\mathbf{c}}_{vert}$, exploiting the relative transformation of the chessboard reference system $\{\mathbf{W}\}$ and the camera reference system $\{\mathbf{C}\}$, obtained from the extrinsic calibration. The same vertices were identified in $\{\mathbf{L}\}$ ($\mathbf{p}^{\mathbf{L}}_{vert}$) as the intersection of the calibration plate edges estimated on the point cloud measured with the laser scanner, as shown in Fig. 6(b). A detailed description of $\mathbf{p}^{\mathbf{L}}_{vert}$ and $\mathbf{p}^{\mathbf{C}}_{vert}$ estimation process is reported in the following paragraphs:

**Vertices estimation in $\{\mathbf{C}\}$.** $\mathbf{p}^{\mathbf{C}}_{vert}$ were computed as stated in the following equation:

$$\mathbf{p}^{\mathbf{C}}_{vert} = \mathbf{T}^{\mathbf{W}}_{\mathbf{C}} * \mathbf{p}^{\mathbf{W}}_{vert} \tag{1}$$

The vertices $\mathbf{p}^{\mathbf{W}}_{vert}$ were geometrically identified in $\{\mathbf{W}\}$ knowing the vertices distances from the origin of the chessboard reference system, and $\mathbf{T}^{\mathbf{W}}_{\mathbf{C}}$ was computed using the Stereo Camera Calibrator Toolbox of Matlab 2015b (The MathWorks, Inc.) [5, 29].

**Vertices estimation in $\{\mathbf{L}\}$.** The pipeline for $\mathbf{p}^{\mathbf{L}}_{vert}$ identification is:

- The points belonging to the calibration plate were manually selected from the laser scan point cloud (removing non-informative points belonging to the background);

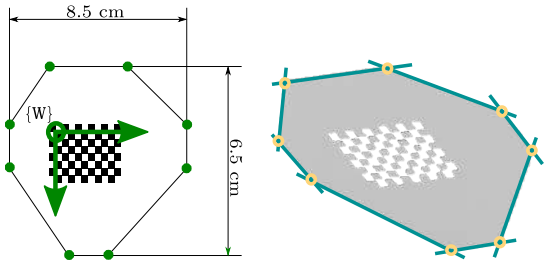



Figure 6: (a) Vertex estimation in $\{\mathbf{C}\}$: view of the calibration plate. The vertex points (green dots) are at known distances from $\{\mathbf{W}\}$ origin. (b) Vertex estimation in $\{\mathbf{L}\}$: view of the calibration plate point cloud. The vertex coordinates in $\{\mathbf{L}\}$ (yellow circle) were calculated as the intersection of each pair of estimated lines (green lines).

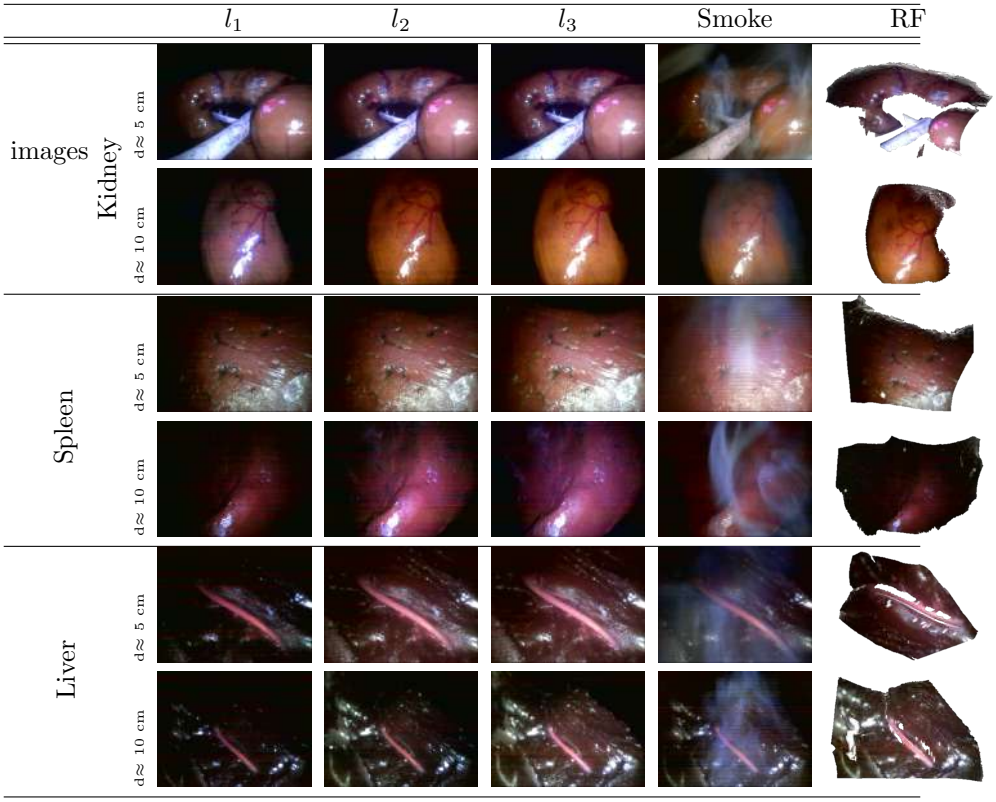



Figure 7: Example of endoscopic stereo dataset images. All the different conditions are represented (distances, levels of light and smoke) for one pose of the spleen, kidney and liver.

- Noise reduction was carried out estimating the calibration plate plane, according to Maximum Likelihood Estimation SAmple Consensus (MLESAC) [26] and projecting on the estimated plane all the points distant less than a threshold (comparable with the accuracy of the laser scanner);
- The edges of the calibration plate were semi-automatically identified: (1) the calibration plate contour was identified searching the minimum and maximum value of the coordinates $x$ and $y$ for each row and column of the discretized point cloud; (2) manually selecting the points belonging to each edge and estimating the corresponding line;
- The $\mathbf{p}^{\mathbf{L}}_{vert}$ were computed as the intersection of each pair of lines, as in Fig. 6(b).

The scan of the calibration plate and the image acquisition were done consecutively, to avoid interference between the laser and the camera.

Once $\mathbf{p}^{\mathbf{L}}_{vert}$ and $\mathbf{p}^{\mathbf{C}}_{vert}$ were identified, $\mathbf{T}^{\mathbf{L}}_{\mathbf{C}}$ was estimated solving the equation (2) with the Singular Value Decomposition (SVD) method:

$$\mathbf{p}^{\mathbf{C}}_{vert} = \mathbf{T}^{\mathbf{L}}_{\mathbf{C}} * \mathbf{p}^{\mathbf{L}}_{vert} \tag{2}$$

The mathematical solution was guaranteed by using more than three non-collinear points [1, 6], namely the eight vertices of the calibration plate. The camera-laser calibration procedure was implemented in Matlab 2015b (The MathWorks, Inc.).

#### 2.2.2 RF 2D map

To facilitate the comparison between the RF transformed in $\{\mathbf{C}\}$ and the 3D reconstructed point cloud, the RF was stored into a 2D map. Each $(u, v)$ cell of the map contains the 3D coordinates $(x, y, z)$ of the point projected on the image plane using the left camera intrinsic parameters. The projection does not take into account the stereo camera rectification.

### 2.3 Acquisition scenario

For the acquisition of *EndoAbS* dataset, the laser scanner and the stereo endoscope were positioned having approximately the same field of view (see Fig. 5). The stereo images and the scans were separately captured, minimising as much as possible the time interval between the acquisition to avoid any changes in the phantoms' pose. All the acquisitions were performed with only the endoscopic light turned on (the external lights were switched off) to mimic the internal abdomen illumination during a surgical procedure. In order to test the robustness of 3D stereo-reconstruction algorithms under different conditions, the images were created introducing: (i) Presence of smoke, created immersing dry-ice in hot water; (ii) 3 different endoscopic light levels ($l_1$, $l_2$, $l_3$), varying the light intensity; (iii) Two phantom-endoscope distances ($dist_{min} \approx 5cm$ and $dist_{max} \approx 10cm$). Sample images of the dataset are shown in Fig. 7.

Thus, the obtained dataset is made of:

- 120 stereo-images of the surgical scenario (PNG format), organised as indicated in Tab. 2;
- RF in form of a point cloud (TXT format);
- Intrinsic parameters for both cameras, and stereo calibration extrinsic parameters (TXT format);
- Description of the errors involved: laser accuracy, mean re-projection error of the camera calibration and camera-laser calibration error.

Table 2: Description of the *EndoAbS* Dataset structure, including acquisition conditions for each organ.

| *Organ* | $d_{min}$ | $d_{max}$ | Total |
|---|---|---|---|
| Spleen | 4 poses | 3 poses | 7 poses |
| Kidney | 4 poses | 4 poses | 8 poses |
| Liver | 2 poses | 3 poses | 5 poses |
| Total | 10 poses | 10 poses | **20 poses** |

Each pose comprises six images: three images with a different level of illumination ($l_1$, $l_2$, $l_3$) and three images with smoke.

# 3 Experimental Evaluation

## 3.1 Dataset generation error evaluation

The errors involved in the dataset generation are introduced by (i) characteristics of the camera and the laser scanner; (ii) camera calibration; (iii) strategy for vertices identification in $\{\mathbf{C}\}$ and $\{\mathbf{L}\}$. The error resulting from the evaluation of the camera-laser calibration procedure is assumed to be the overall estimation of the error. In order to measure this error, 10 *validation sets* consisting of images and laser scans were acquired with the experimental setup shown in Fig. 5. In each set, 9 calibration plate orientations (Fig. 8) were exploited, varying approximately $\pm 30°$along vertical and horizontal direction. These sets were used to compute $\mathbf{T}_{\mathbf{C}}^{\mathbf{L}}$.

In addition, a *test set* composed of 27 image-scan pairs was acquired to evaluate the camera-laser calibration error $\epsilon$, defined as the median Euclidean distance between $\mathbf{p}_{vert}^{\mathbf{C}}$ and $\mathbf{p}_{vert}^{\mathbf{L}}$ projected in $\{\mathbf{C}\}$ with the computed $\mathbf{T}_{\mathbf{C}}^{\mathbf{L}}$.

The median was considered since the error population was not normally distributed (Kolmogorov-Smirnov test $p_{value} < 0.05$).

A statistical analysis was conducted to verify if there is a correlation between: (i) $\epsilon$ and the number of image-scan pairs used in the camera-laser calibration procedure, (ii) $\epsilon$ and the orientation of the calibration plate with respect to the camera-laser configuration.

- *Number of image-scan pairs*: The correlation between $\epsilon$ and the number of the used image-scan pairs was estimated computing $\mathbf{T}_{\mathbf{C}}^{\mathbf{L}}$ varying the number of image-scan pair from 1 to 10 and computing $\epsilon$ applying the obtained $\mathbf{T}_{\mathbf{C}}^{\mathbf{L}}$ to the *test set*. The statistical correlation was evaluated through the Pearson Product-Moment Correlation Coefficient ($p_{value} < 0.05$).

- *Orientation of the calibration plate*: The statistical dependence between $\epsilon$ and the orientation of the calibration plate was evaluated computing $\mathbf{T}_{\mathbf{C}}^{\mathbf{L}}$ for 9 different calibration plate orientations of the *validation set* (Fig. 8). Kruskal-Wallis test was performed ($p_{value} < 0.05$) to assess the presence of statistical difference among the different orientations of the calibration plate.

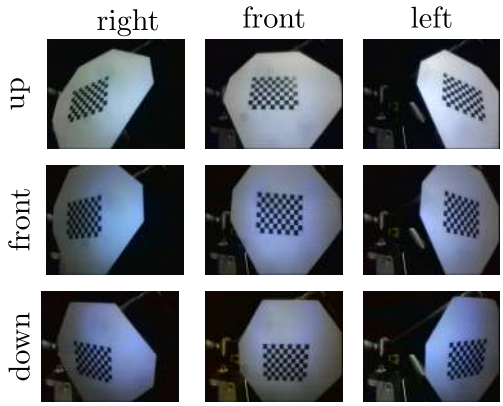



Figure 8: Example of calibration plate orientation with respect to the image plane.

A comparison of our method with respect to a state-of-the-art calibration method [27] was conducted, using 18 image-scan pairs, as suggested in the paper. The calibration error $\epsilon$ was evaluated for 10 different trials of calibration for both methods, and Kruskal-Wallis test was performed ($p_{value} < 0.05$) to assess the presence of statistical difference between the two methods. Since the state-of-the-art algorithm requires image-scan pairs of the calibration plate at different orientations, the calibration plate was positioned farther away from the laser scanner and the camera, in order to be visible by both of them, however compromising the calibration accuracy. For this reason, we also evaluated our method using one single image-scan pair oriented towards the laser scanner and the camera, thus shortening the distance of the calibration plate from them.

## 3.2 Dataset realism evaluation

In order to investigate the realism of the endoscopic images as regards to real clinical images, evaluations of surgeons were collected as part of a questionnaire with scores in 5-point Likert-type scale. The users involved were 9 medical doctors with 1 to 30 years of experience in general, urology and cancer surgery. Their field of expertise ranges from open surgery $(11,1\%)$ to robotic minimally-invasive surgery $(22,2\%)$ and laparoscopic surgery $(66,7\%)$.

The questionnaire was made of 24 images, a sample of images taken from *EndoAbS* dataset and representative of the different level of lights, distances, presence of smoke and different organs. The order of the images was randomized to provide a global overview of the image realism. For each image the users had to indicate an answer to the question *How much are the characteristics represented in the image (tissue, illumination, specular highlight, smoke, distances from the tissues) similar to a real scenario?* along a line divided in 5 intervals

(the score 1 means very dissimilar, while 5 implies very similar).

The $score_{mean}$ was considered as the average of the scores assigned to the images.

### 3.3 3D reconstruction evaluation

*EndoAbS* dataset was used for the evaluation of a 3D reconstruction algorithm previously developed by the authors [18]. For the validation process, we decided to follow the terminology and methodology of the protocol for reference-based validation studies proposed in [8], in order to allow for comparability of the results. The protocol starts requiring the definition of the specification of the validation objective, including the clinical context and objective (C) that in our case can be identified as dense and accurate ($accuracy < 2mm$) 3D reconstruction for abdominal MIS surgery. Following the terminology of the protocol, the evaluated method $M$ [18] is referred as $F_M$, and $Ref$ stands for the RF, also called Ground Truth. The dataset used for the validation is referred as $D_I$, i.e. in this case the *EndoAbS* dataset, where $D_I^M$ are the 120 stereo images and $R_{Ref}$ the associated RF already transformed in the camera reference system. $E_{Ref}$ is the error committed in the generation of the RF associated to the images and it is described in Sec. 2.3. The hardware used for the acquisition of $D_I^M$ and $D_I^{Ref}$ is also described in Sec. 2.3. The parameters $P_I$ used by $F_M$ are described in Tab. 3.

Table 3: Parameters $P_I$ used by $F_M$ (3D reconstruction algorithm)

| $F_M$ **Parameters** | value |
|---|---|
| Census window | 9x9 |
| Census Block Size | 11x11 |
| Threshold Spurious remover | 10 |
| Threshold LRConsistency Check | 4 |
| LO-RANSAC max iteration | 100 |
| number of super pixel | 70 |
| disparity range | 150-250 |

As validation criterion $VC$, we are proposing n evaluation protocol made of the following metrics:

**Accuracy.** The 3D reconstruction *accuracy* was evaluated as the median of the Euclidean distances between the reconstructed point cloud $R_M$ and $R_{Ref}$, which is the discrepancy obtained through the comparison function $O_D = F_C(R_M, R_{Ref})$ defined by the protocol. Since the point cloud and the ground truth are stored in a 2D map, the error can be calculated for each pixel of the image. Note that the 2D map of the RF is expressed in the non-rectified left camera image plane. Therefore, a rectification of the 2D map of the RF was necessary in order to perform a pixel to pixel comparison. Only the pixels of the left grayscale image with an intensity value greater than 16 (hereafter

called region of interest) were considered in the evaluation, eliminating the areas where the organ phantom is not present, following the same criteria used in [18]. The region of interest (ROI) was chosen on the image with the highest level of illumination, since the low level of illumination could present pixels with low intensity even if they belong to the organ surface. The same ROI was used for the evaluation of the images with other levels of illumination and presence of smoke.

**Percentage of reconstructed points.** It was computed as the ratio between the number of reconstructed points with respect to the number of RF points, both identified in the region of interest.

**Robustness.** The algorithm was applied to the entire dataset, considering $l_1$, $l_2$ and $l_3$ different illumination levels, $dist_{min}$ and $dist_{max}$ between the endoscope and the organs, and *smoke* presence. A non-parametric test (Kruskal-Wallis $p_{value} < 0.05$) was performed to test if the *accuracy* was statistically different varying (i) $l_1$, $l_2$ and $l_3$, (ii) $dist_{min}$ and $dist_{max}$ ($\approx$ 5cm and $\approx$ 10cm) considering only $l_3$ and (iii) *smoke* presence against $l_3$ for $dist_{min}$ and $dist_{max}$.

# 4 Results

## 4.1 Phantom surgical scenario

An abdominal surgical scenario was recreated with phantoms of liver, spleen and kidneys. Superficial vessels were painted and big vessels were added to increase the realism of the organs. The stress-strain curves, obtained from the compressive mechanical test, revealed that using the 50% of softening agent the Young's modulus of the polyurethane material (0.97 kPa) is comparable with the liver one (see Sec. 2.1). A cost analysis of the moulding process is reported in Tab. 4.

Table 4: Abdominal Phantom Costs

| organ | molds [€] | polyurethane [€] | total [€] |
|---|---|---|---|
| Spleen | 80 | 30 | 110 |
| kidney | 40 | 10 | 50 |
| Liver | 170 | 90 | 260 |
| | | TOTAL [€] | 420 |

## 4.2 Dataset error

Regarding the camera-laser calibration evaluation, no statistical correlation between the calibration error $\epsilon$ and the number of image-scan pairs used for the calibration was found. Moreover, no statistical difference for the calibration errors $\epsilon$ varying the calibration plate orientation was found.

Statistical difference was found comparing the calibration error of the presented method with respect to the method proposed by [27]. There are two main causes for this difference. The first one is attributed to the filtering of the laser scanner data during the plane estimation in the proposed method, as described in Sec. 2.2.1. And the second one is related with the fact that [27] uses plane-to-plane distance minimization instead of point-to-point distance minimization, not taking into account the translation along the plane directions and rotation around the plane axis. Numerical results are summarized in Tab. 5. The evaluation of the camera-laser calibration using only one image-scan pair showed an error equal to $0.43mm$ ($Q_1 = 0.41mm$ - $Q_3 = 0.43mm$).

A description of the specifications of the instruments used for the generation of the dataset and of the errors measured in the process is reported in Tab. 6.

Table 5: Camera-Laser Calibration Errors

| | $\epsilon$[mm] | $Q_1$[mm] | $Q_3$[mm] |
|---|---|---|---|
| State of the Art [27] | 1.94 | 1.83 | 2.62 |
| *Our method* | 1.43 | 1.02 | 1.78 |
| *Our method** | 0.43 | 0.41 | 0.43 |

* These results come from an evaluation of the proposed camera-laser calibration method using only one image-scan pair, as explained in Sec. 3

Table 6: *EndoAbS* Dataset generation errors

| | |
|---|---|
| Laser scanner accuracy* | $x = \pm 0.22mm$<br>$y = \pm 0.16mm$<br>$z = \pm 0.07mm$ |
| Mean Reprojection Error (left camera) | 0.250 pixels |
| Mean Reprojection Error (righ camera) | 0.235 pixels |
| Camera-laser calibration error | $0.43mm$ |

* Conditions: distance $0.6m$, temperature $20^{\circ}$C, relative humidity 65% or less

## 4.3 Dataset qualitative evaluation

In Fig. 9 is presented a box plot summarizing the score assigned to the 24 sample images included in the questionnaire. The $score_{mean}$ is 2.7 ($\pm 0.50$). Note that the image on the x-axis are presented in the same order as they appeared in the questionnaire.

## 4.4 3D reconstruction evaluation

In Tab. 7, the *accuracy* and the *percentage of reconstructed points* are reported. In case of $dist_{min}$, *accuracy* was not statistically different for 3 levels of illumination ($p_{value} = 0.38$), and with or without the presence of smoke

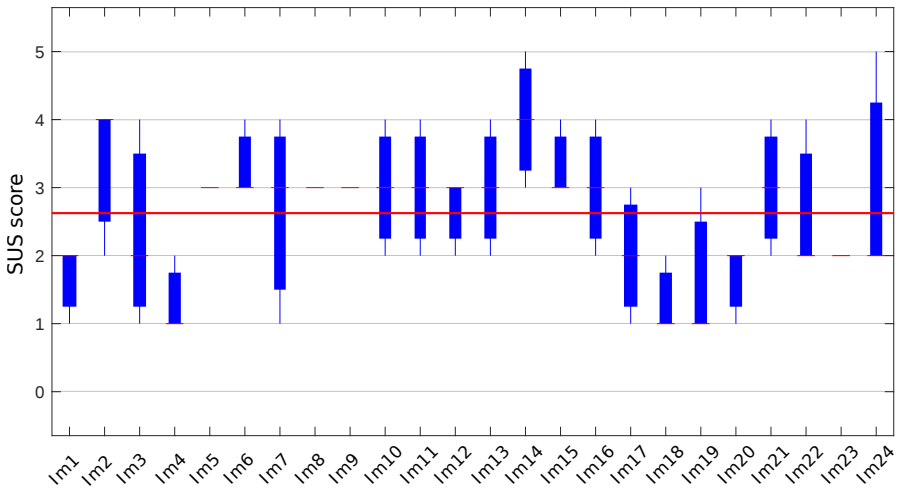



Figure 9: Boxplot showing the SUS score (from one to five considering the System Usability Scale questionnaire) assigned to 24 selected images of *EndoAbS* dataset by the surgeons answering to a questionnaire, in order to evaluate the realism of the characteristics represented by the images.

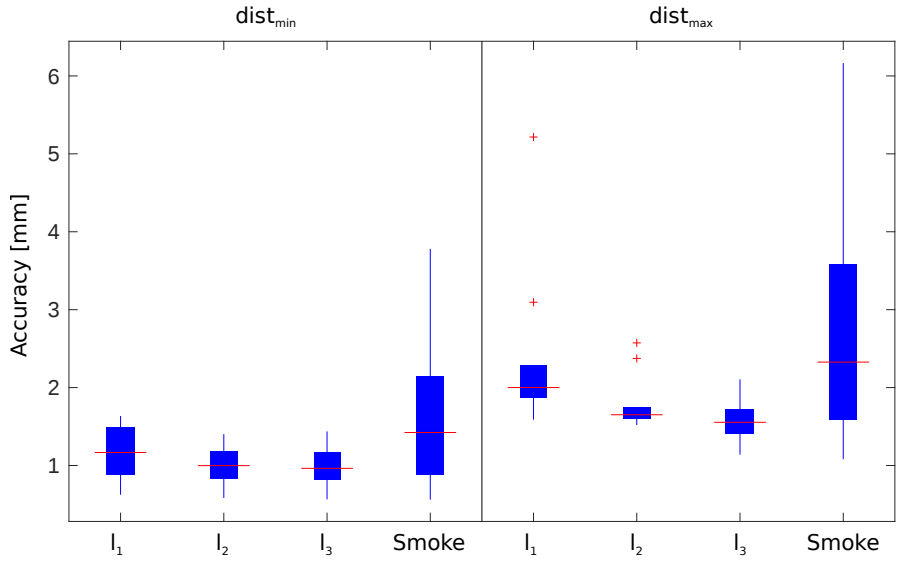



Figure 10: Boxplot showing the accuracy of the 3D reconstruction algorithm varying the level of illumination ($l_1$, $l_2$, $l_3$) and the presence of smoke for $dist_{min}$ (left) and $dist_{max}$ (right).

Table 7: 3D stereo-reconstruction algorithm performance in terms of accuracy (mean and standard deviation) and percentage of reconstructed points.

| | $dist_{min}$ | | | |
|---|---|---|---|---|
| | $l_1$ | $l_3$ | $l_3$ | $s$ |
| *accuracy* [mm] | 1.16 | 1.01 | 1.00 | 2.62 |
| *std* [mm] | 0.34 | 0.25 | 0.27 | 4.17 |
| *points* [% ] | 98.99 | 93.25 | 93.25 | 89.52 |
| | $dist_{max}$ | | | |
| *accuracy* [mm] | 2.40 | 1.80 | 1.55 | 3.61 |
| *std* [mm] | 1.07 | 0.37 | 0.30 | 3.81 |
| *points* [% ] | 76.64 | 87.56 | 93.04 | 97.69 |

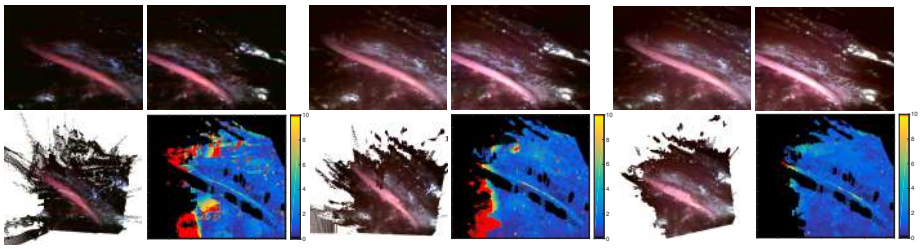


Figure 11: Example of results of the 3D reconstruction algorithm varying the level of light from the left ($l_1$) to the right ($l_3$). For each pair of stereo images, the bottom row shows the error map (right) and the reconstructed point cloud (left). The color bars represent the error in $mm$.

($p_{value} = 0.17$). In case of $dist_{max}$ there was significance difference between $l_1$ and $l_3$ ($p_{value} = 0.0052$), and between $l_3$ and *smoke* ($p_{value} = 0.049$). The same test performed between $dist_{min}$ and $dist_{max}$ for the illumination level $l_3$ showed that there was statistical difference with $p_{value} = 0.0025$. These results are shown in Fig. 10. An example of the errors in the point cloud 3D reconstruction is shown in Fig. 11 and Fig. 12.

# 5 Discussion and Conclusion

This paper describes the creation of *EndoAbS* dataset for the quantitative evaluation of 3D reconstruction algorithms based on stereo-images. The dataset consists in 120 endoscopic stereo-images and the associated RF. The main contribution of this work is to increase the number and variety of openly available surgical stereo-image datasets, which are essential to test and benchmark the accuracy and robustness of 3D stereo-reconstruction algorithms under realistic conditions. In this paper, we also provide an analysis of the errors involved in the dataset creation process, particularly the camera-laser calibration error,

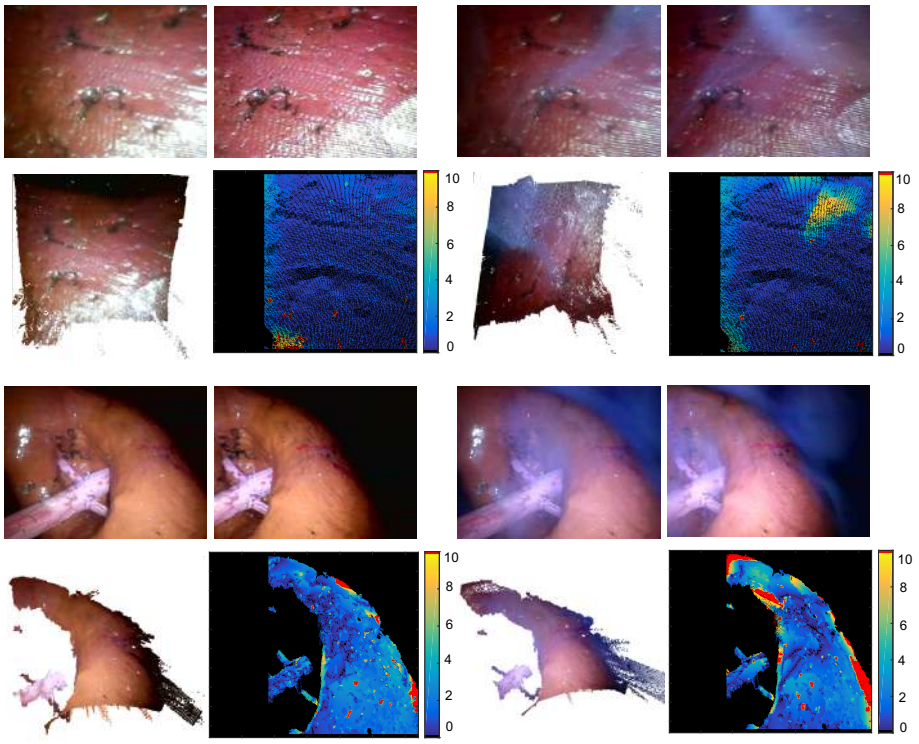



Figure 12: Example of results of the 3D reconstruction algorithm without and with presence of *smoke*. For each pair of stereo images, the bottom row shows the error map (right) and the reconstructed point cloud (left). The color bars represent the error in *mm*.

which represents an overall estimation of the dataset error and it is an important parameter for proper assessment of reconstruction algorithms. In addition, a detailed description of the phantom development and the methods used was provided to facilitate future expansion of *EndoAbS* dataset or the development of additional datasets adapted to other specific needs.

A surgical scenario made of phantoms was specifically fabricated for the creation of *EndoAbS*. This provided a positive tradeoff between the quality of the RF that can be obtained and the clinical realism of the data. Indeed, phantom does not suffer of changes in shape in the short time and can be re-used many times without deteriorating, as opposed to ex-vivo organs. Liver, kidney and spleen were created with a moulding process, and a compressive mechanical test was conducted to give approximately to the phantom the same stiffness of real tissues.

Moreover, in order to make the models as realistic as possible, the organ surfaces were painted emulating tissue texture and superficial tiny vessels. Big

vessels were also reproduced to allow the evaluation of 3D reconstruction algorithms in case of more complicated structures and at different depths. Nevertheless, not all of these realistic properties were exploited for the generation of images in the dataset yet. The acquisition of images of tissue deformations and of their interaction with surgical instruments will be part of a future expansion of *EndoAbS*.

Regarding the assessment of the realism of the images, results obtained from questionnaires demonstrate that surgeons consider them of satisfactory realism. As expected, the average rating score was not high since surgeons can easily distinguish between real images and those in our dataset. Nevertheless, the quality of the images was deemed satisfactory for the scope of this contribution.

The corresponding RF of each stereo-image pair was generated using a laser scanner, and a calibration algorithm was designed to register the RF in the left camera reference system. The benefits of the proposed calibration approach with respect to state-of-the-art methods were demonstrated by the highly accurate calibration achieved with a single scan of the calibration plate (median calibration error $0.43mm$). When using other methods, e.g. [27], a comparable level of accuracy can be achieved with 15 to 20 image-scan pairs. This factor accelerates and facilitates the calibration process, since it is difficult to find the right workspace in which the calibration plate is seen by both measuring systems.

The evaluation of a 3D reconstruction algorithm using the dataset has demonstrated its applicability. The computed accuracy errors for the evaluated algorithm are in accordance with the ones previously reported in [18]. A deeper analysis of the algorithm has confirmed that the results are more accurate if the endoscope is closer to the tissue. In this case, the algorithm performs well even under varying lighting conditions or the presence of smoke. In case of higher distances from tissue, the accuracy is more affected by the illumination level or the presence of smoke. Note that when the endoscope-tissue distance increases, the illumination and the disparity resolution decreases, directly affecting the algorithm performance. This could explain the difference in accuracy and percentage of reconstructed points between $dist_{min}$ and $dist_{max}$.

During this work, the usage of a custom-made endoscope and light was motivated by the unavailability of a standard commercial equipment, due to their high cost. Such endoscope does provide lower resolution images compared to modern clinical devices. Moreover, the images were captured with the endoscope in a fixed position, thus, they do not reproduce the shivering behaviour due to the manipulation of the endoscope by the clinician, better simulating the condition of robotic surgery, where the camera is moved using a robotic arm.

As part of future work, *EndoAbS* dataset will be expanded to include images with presence of blood, instrument occlusion, and dynamic changes. This will include tissue motions caused by heart beating and breathing, and deformations due to the contact with surgical instruments. Adding dynamic information to the dataset, would also give the opportunity to use it within a simulator environment like SOFA (https://www.sofa-framework.org).

# References


[1] Arun KS, Huang TS, Blostein SD (1987) Least-squares fitting of two 3-d point sets. IEEE Transactions on pattern analysis and machine intelligence (5):698–700

[2] Bernhardt S, Nicolau SA, Soler L, Doignon C (2017) The status of augmented reality in laparoscopic surgery as of 2016. Medical image analysis 37:66–90

[3] Ciullo AS, Penza V, Mattos L, De Momi E (2016) Development of a surgical stereo endoscopic image dataset for validating 3d stereo reconstruction algorithms. 6th Joint Workshop on New Technologies for Computer/Robot Assisted Surgery

[4] Condino S, Carbone M, Ferrari V, Faggioni L, Peri A, Ferrari M, Mosca F (2011) How to build patient-specific synthetic abdominal anatomies. an innovative approach from physical toward hybrid surgical simulators. The International Journal of Medical Robotics and Computer Assisted Surgery 7(2):202–213

[5] Heikkila J, Silvén O (1997) A four-step camera calibration procedure with implicit image correction. In: Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, IEEE, pp 1106–1112

[6] Horn BK (1987) Closed-form solution of absolute orientation using unit quaternions. JOSA A 4(4):629–642

[7] Hu M, Penney G, Edwards P, Figl M, Hawkes DJ (2007) 3d reconstruction of internal organ surfaces for minimal invasive surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 68–77

[8] Jannin P, Grova C, Maurer Jr CR (2006) Model for defining and reporting reference-based validation protocols in medical image processing. International Journal of Computer Assisted Radiology and Surgery 1(2):63–73

[9] Lin J, Clancy NT, Stoyanov D, Elson DS (2015) Tissue surface reconstruction aided by local normal information using a self-calibrated endoscopic structured light system. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 405–412

[10] Maier-Hein L, Mountney P, Bartoli A, Elhawary H, Elson D, Groch A, Kolb A, Rodrigues M, Sorger J, Speidel S, Stoyanov D (2013) Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. Medical image analysis 17(8):974–996

[11] Maier-Hein L, Groch A, Bartoli A, Bodenstedt S, Boissonnat G, Chang PL, Clancy N, Elson DS, Haase S, Heim E, Hornegger J, Jannin P, Kenngott H, Kilgus T, Muller-Stich B, Oladokun D, Rhl S, dos Santos TR, Schlemmer HP, Seitel A, Speidel S, Wagner M, Stoyanov D (2014) Comparative validation of single-shot optical techniques for laparoscopic 3-d surface reconstruction. IEEE transactions on medical imaging 33(10):1913–1930

[12] Mattei G, Tirella A, Gallone G, Ahluwalia A (2014) Viscoelastic characterisation of pig liver in unconfined compression. Journal of biomechanics 47(11):2641–2646

[13] Mountney P, Yang GZ (2010) Motion compensated slam for image guided surgery. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010, Springer, pp 496–504

[14] Nicolau S, Soler L, Mutter D, Marescaux J (2011) Augmented reality in laparoscopic surgical oncology. Surgical oncology 20(3):189–201

[15] Okamoto T, Onda S, Yanaga K, Suzuki N, Hattori A (2015) Clinical application of navigation surgery using augmented reality in the abdominal field. Surgery today 45(4):397–406

[16] Penza V, Ortiz J, De Momi E, Forgione A, Mattos L (2014) Virtual assistive system for robotic single incision laparoscopic surgery. In: 4th Joint workshop on computer/robot assisted surgery, pp 52–55

[17] Penza V, Bacchini S, Ciullo AS, De Momi E, Forgione A, Mattos LS (2015) Label-based optimization of dense disparity estimation for robotic single incision abdominal surgery. In: Proceedings of The Hamlyn Symposium on Medical Robotics

[18] Penza V, Ortiz J, Mattos LS, Forgione A, De Momi E (2015) Dense soft tissue 3d reconstruction refined with super-pixel segmentation for robotic abdominal surgery. International journal of computer assisted radiology and surgery pp 1–10

[19] Pratt P, Stoyanov D, Visentini-Scarzanella M, Yang GZ (2010) Dynamic guidance for robotic surgery using image-constrained biomechanical models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 77–85

[20] Röhl S, Bodenstedt S, Suwelack S, Kenngott H, Müller-Stich BP, Dillmann R, Speidel S (2012) Dense gpu-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration. Medical physics 39(3):1632–1645

[21] Scharstein D, Szeliski R (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision 47(1-3):7–42

[22] Scharstein D, Hirschmüller H, Kitajima Y, Krathwohl G, Nešić N, Wang X, Westling P (2014) High-resolution stereo datasets with subpixel-accurate ground truth. In: German Conference on Pattern Recognition, Springer, pp 31–42

[23] Stoyanov D (2012) Surgical vision. Annals of biomedical engineering 40(2):332–345

[24] Stoyanov D, Scarzanella MV, Pratt P, Yang GZ (2010) Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 275–282

[25] Tirella A, Mattei G, Ahluwalia A (2014) Strain rate viscoelastic analysis of soft and highly hydrated biomaterials. Journal of Biomedical Materials Research Part A 102(10):3352–3360

[26] Torr PH, Zisserman A (2000) Mlesac: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding 78(1):138–156

[27] Unnikrishnan R, Hebert M (2005) Fast extrinsic calibration of a laser rangefinder to a camera. Carnegie Mellon University

[28] Yeh WC, Li PC, Jeng YM, Hsu HC, Kuo PL, Li ML, Yang PM, Lee PH (2002) Elastic modulus measurements of human liver and correlation with pathology. Ultrasound in medicine & biology 28(4):467–474

[29] Zhang Z (2000) A flexible new technique for camera calibration. IEEE Transactions on pattern analysis and machine intelligence 22(11):1330–1334