


ORIGINAL ARTICLE

WILEY

The International Journal of Medical Robotics and Computer Assisted Surgery



EndoAbS dataset: Endoscopic abdominal stereo image dataset for benchmarking 3D stereo reconstruction algorithms

Veronica Penza^{1,2}  | Andrea S. Ciullo² | Sara Moccia^{1,2} | Leonardo S. Mattos¹ | Elena De Momi²

¹Department of Advanced Robotics, Istituto Italiano di Tecnologia, 16163 Genova, Italy

²Department of Electronics Information and Bioengineering, Politecnico di Milano, 20133 Milano, Italy

Correspondence

Veronica Penza, Department of Advanced Robotics, Istituto Italiano di Tecnologia, via Morego, 30, 16163 Genova, Italy.
Email: veronica.penza@iit.it

Handling Editor: Loughlin Clive

Abstract

Background: 3D reconstruction algorithms are of fundamental importance for augmented reality 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 dataset of endoscopic stereo images (*EndoAbS* dataset).

Methods: The dataset includes (i) endoscopic stereo images of phantom abdominal organs, (ii) a 3D organ surface reference (RF) generated with a laser scanner and (iii) camera calibration parameters. A detailed description of the generation of the phantom and the camera–laser calibration method is also provided.

Results: An estimation of the overall error in creation of the dataset is reported (camera–laser calibration error 0.43 mm) 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 (< 2 mm).

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

KEYWORDS

camera–laser calibration, robotic surgery, soft abdominal organ phantoms, stereo reconstruction, surgical image dataset

1 | INTRODUCTION

In minimally invasive surgery (MIS), the application of augmented reality systems is aimed at improving the outcome of surgery by intraoperatively enhancing the surgeon's perception and providing guidance inside the patient's body. Indeed, these systems provide the surgeon with additional useful information from preoperative planning, which when fused into the intraoperative scenario can, for example, help in the localization of a tumour area, as described by several authors.^{1–4} However, during the surgery, the organs' geometry is constantly changing due to breathing, heartbeat 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 intraoperatively, with the

aim of measuring the deformation of the surgical site in real time.⁵ These methods have the potential to replace the usage of intraoperative computed tomography (CT) or magnetic resonance imaging (MRI), and overcome their drawbacks (such as non-real-time information exposure of the patient to radiation and high costs) by just exploiting only the images captured from a stereo endoscope. Although the performance of these algorithms is well established in different fields, such as domestic, industrial robots and the gaming industry,⁶ their application to surgical endoscopic images has proved to be challenging due to the peculiarities that a surgical scenario presents, such as homogeneous or periodic tissue texture, non-uniform illumination, the 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 by comparing the resulting point cloud against a 3D surface reference (in this paper referred as RF),

Veronica Penza and Andrea S. Ciullo equally contributed to this work.

**TABLE 1** Openly available surgical endoscopic datasets

| Surgical scenario | Organ | RF | Characteristics | References |
|-------------------|---|----------|---|---|
| Virtual phantom | liver | 3D model | 3 liver texture, endoscope-tissue of 5 cm, 360° endoscope rotation with 5° steps, zoom in and zoom out of the same liver spot (max zoom 40mm with 2mm steps), tissue deformation | Hu et al. ⁹ Rohl et al. ¹⁰ Mountney and Yang ¹¹ |
| Ex vivo organs | porcine liver, kidney, heart, fatty tissue | CT scan | different illumination levels, presence of smoke and blood, two endoscope-tissue distances (5cm and 7 cm), two endoscope orientations angles (0° and 30°) | Maier et al. ¹² Lin et al. ¹³ |
| Phantom organs | heart | CT scan | two views of a beating heart | Stoyanov et al. ¹⁴ Pratt et al. ¹⁵ Penza et al. ¹⁶ |

assumed to correspond, or at least to be close, to the real solution.⁷ Unfortunately, even if there are many stereo datasets representing static indoor scenes,^{6,8} only a few datasets providing surgical endoscopic images with an associated RF are publicly available (see Table 1). Röhl et al.¹⁰ presented synthetic stereo images and the corresponding RF, taken from a virtual model of the liver by using a simulated stereo endoscope. Stoyanov et al.¹⁴ and Pratt et al.¹⁵ proposed a stereo-image dataset of a moving heart phantom (Chamberlain Group, MA, USA), generated using the da Vinci[®] surgical system, providing CT reference data*. More recently, a dataset of stereo-images of *ex vivo* animal organs (liver, heart and kidneys) has been presented by Maier-Hein et al.,¹² providing a CT scanner-based RF and exploring different conditions, such as the presence of blood and smoke, as well as different poses of the endoscope. These datasets[†] have been used for validating and benchmarking different 3D reconstruction algorithms, as summarized in Table 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 consist of stereo images, associated RF, camera calibration parameters and errors involved in the RF creation process that can affect the evaluation of the algorithms.
2. The images should present the main characteristics of real endoscopic surgical scenarios, as mentioned previously.
3. It should be publicly available in order to allow validation and benchmarking of image processing and computer vision algorithms.

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,⁹⁻¹¹ phantoms^{9,14,15} and *ex vivo* organs^{11,12,17} 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 reality; in the case of *ex vivo* organs, in order to preserve the shapes of the organs between the RF scan and the acquisition of images, the organs have to be kept in specific conditions for as long as possible (in water and at low temperature), causing timing constraints during the experiments; in the case of organ phantoms, the main difficulties relate to the reproduction of the appearance and the mechanical properties of tissue (if tissue deformations are also simulated). In the latter two cases, another constraint is associated with the availability in research laboratories of a CT scanner or laser scanner (used to generate the RF) due to their high cost.

Considering the increasing need for surgical stereo image datasets, the aim of this work is the generation of an endoscopic abdominal stereo image dataset (*EndoAbS*) for validation of 3D stereo reconstruction algorithms, specifically focusing attention on the evaluation of passive stereo reconstruction methods. The *EndoAbS* dataset is composed of 120 stereo images of phantoms of different abdominal organs, showing either flat organ surfaces (spleen) or more complex structures such as vessels in the 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 for testing the robustness of 3D stereo reconstruction algorithms under different conditions. Each pair of images is coupled with its RF, which 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 to the RF in the camera reference system (camera-laser calibration) and its accuracy in use. Moreover, in order to exemplify the usage of the *EndoAbS* dataset, the performance of a 3D reconstruction algorithm, previously implemented by the authors, was evaluated using the proposed protocol.

* Available at <http://hamlyn.doc.ic.ac.uk/vision/>

† Available at <http://open-cas.com/>

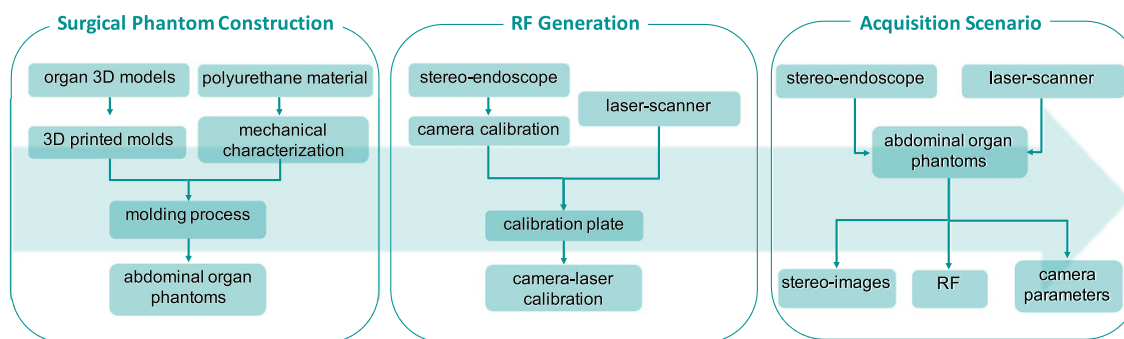


FIGURE 1 Workflow for generation of the dataset. *Surgical phantom construction*: the phantom organs were designed using the 3D models of the liver, spleen and kidneys provided by 3DIRCADb, and a mechanical characterization study was conducted to define the stiffness of the material used to create the 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 validation of the algorithm. *Acquisition scenario*: the surgical stereo-endoscopic dataset was generated, consisting of stereo images, RF and camera parameters



FIGURE 2 Left: details of spleen, kidneys and a liver. Right: the ribcage containing the organs

With respect to the already openly available dataset, the *EndoAbs* dataset is proposed to provide (i) a higher number of stereo images; (ii) a wider variety of shapes of tissue and organs, ranging from a smooth surface to more complex structures such as vessels; (iii) a highly 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 are openly available online for the benefit of the computer-assisted surgery community[‡]. A preliminary description of the *EndoAbs* dataset is presented by Ciullo et al.¹⁸

The paper is structured as follows: in section 2, the workflow for the generation of the dataset is described, considering the construction of the abdominal phantom and the RF generation process with a description of the camera–laser calibration procedure. The experimental setup to validate the errors in dataset generation is presented in section 3 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

The *EndoAbs* dataset was generated by 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 a resolution of 640×480 pixels, a baseline of 6 mm 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 real values of the 3D surface of the surgical scenario as closely as possible. It was generated using the laser scanner Vivid 910 (accuracy[§] of $x = \pm 0.22$ mm, $y = \pm 0.16$ mm, $z = \pm 0.07$ mm and a precision of $8\mu\text{m}$) and the software Polygon Editing Tool (Konica Minolta).

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

2.1 | Construction of the surgical phantom

Liver, spleen and two kidneys were created through a moulding process as in Condino et al.,¹⁹ and a ribcage-like support was 3D printed to maintain the relative position between the organs, as shown in Figure 2. The steps of the process are shown in Figure 3 and described in the following subsections.

2.1.1 | 3D organ model and mould generation

The 3D models of the organs and ribcage were taken from 3D-IRCADb[¶]. 3D-IRCADb includes anonymized DICOM CT medical

[‡]<http://hearlab.polimi.it/medical/dataset/>

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

[¶]<http://www.ircadb.fr/research/3dircadb/>

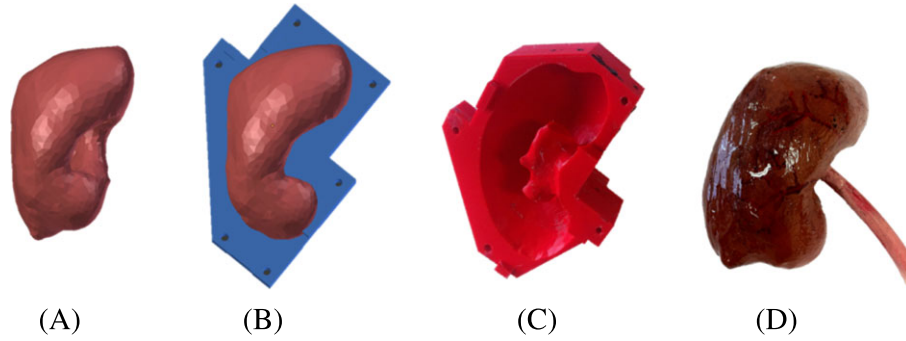


FIGURE 3 Example of the moulding process for the creation of a kidney phantom: (a) 3D virtual model from the 3D-IRCADb CT database; (b) 3D virtual negative moulds; (c) 3D printed negative mould; (d) polyurethane kidney phantom

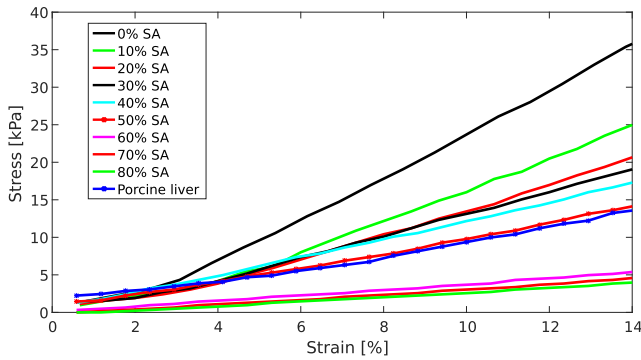


FIGURE 4 Stress-strain curves for each phantom sample with different percentages of softening agent (SA) and the liver sample used as a template

images (voxel size: $0.96\text{mm} \times 0.96\text{mm} \times 2.4\text{mm}$) with an associated manual segmentation performed by expert clinicians, and an organ surface model stored in VTK format, as shown in Figure 3(a). 3D virtual negative moulds were modelled using the software Blender 2.7.4 (Blender Foundation, Amsterdam), as shown in Figure 3(b). The virtual moulds were 3D printed in acrylonitrile butadiene styrene (ABS), using the Elite Dimension 3D printer (layer thickness: 0.25 mm), see Figure 3(c).

2.1.2 | Polyurethane organ phantom

We decided to recreate soft phantoms of abdominal organs with the aim of representing the real surgical scenario as closely as possible. This characteristic will also permit a future improvement of the dataset with images of tissue-instrument interaction. 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 approximately match the real tissue characteristics. We considered different stiffness values for liver tissue reported in the literature: 1.3 kPa,²⁰ 0.90 to 1.730 kPa,²¹ 2.0 kPa.²² However, since the measured viscoelastic properties can vary depending on experimental conditions and on the testing method used,²² we decided to perform a compressive mechanical test comparing the results obtained from a cylindrical sample (height 15 mm, diameter 28.2 mm) of porcine liver against samples of polyurethane made with different percentages of softening agent (from 0% to 80% in steps of 10%). The compressive mechanical test was done with a testing machine (Easydur Dyno), compressing the samples to a height of

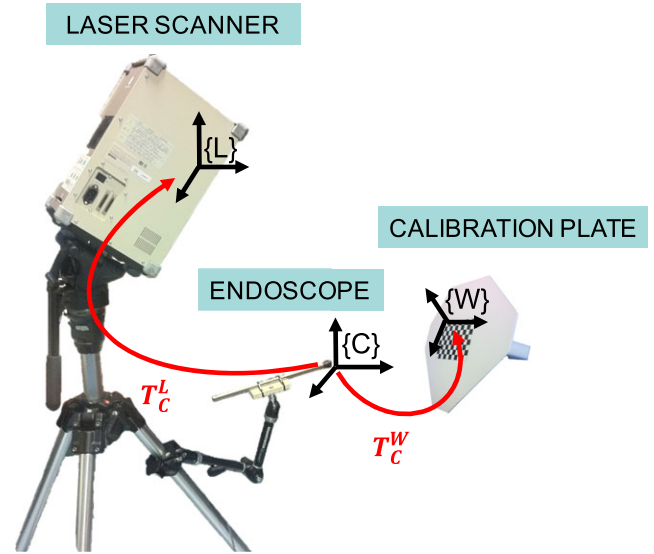


FIGURE 5 The camera reference system $\{C\}$, laser reference system $\{L\}$ and chessboard reference system $\{W\}$ are the reference systems involved in the camera-laser calibration. T_C^W is the transformation from $\{C\}$ to $\{W\}$; T_C^L is the unknown transformation from $\{C\}$ to $\{L\}$

2.5 mm in discrete steps of 0.1 mm. Each trial was performed from a starting configuration in which the piston was in contact with the sample, introducing a pre-strain of 0.1 mm to the samples. Consequently, the stress-strain curves (see Figure 4) and the Young's modulus for each sample were computed, allowing the right percentage of softening agent to be found.

Furthermore, the organs were painted with acrylic colours to simulate the superficial texture of the tissue with the aid of a sponge, and small vessels using acrylic markers with a fine tip, as shown in Figure 3(d). In the liver and kidney phantoms, plastic tubular structures were attached on the surface and painted to represent the main vessels, as shown in Figure 2. A transparent ultrasound gel was placed 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 because it is standard practice to use it as the reference system in 3D reconstruction algorithms.

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 reference systems of the laser scanner and left camera, $\{L\}$ and $\{C\}$ respectively. For the sake of clarity, the setup, the reference systems and the geometrical transformation involved in this method are summarized in Figure 5.

In order to perform this calibration, it is necessary to use a custom target whose corners can be identified by both the laser, as 3D geometrical features, and the camera, as 2D visual information. To this end, an asymmetrical octagonal calibration plate was designed, whose vertices p_{vert}^L were used as the set of points for the calibration process, as shown in Figure 6.

In order to improve the manual selection of vertices in $\{C\}$ (p_{vert}^C), a standard square chessboard (7×10 , square size: 2.5 mm) was placed on the calibration plate. Knowing the location of the plate vertices p_{vert}^L relative to the chessboard, it is possible to compute the position of p_{vert}^C , exploiting the relative transformation of the chessboard reference system $\{W\}$ and the camera reference system $\{C\}$, obtained from the extrinsic calibration. The same vertices were identified in $\{L\}$ (p_{vert}^L) as the intersection of the calibration plate edges estimated on the point cloud measured with the laser scanner, as shown in Figure 6(b). A detailed description of the process of estimating p_{vert}^L and p_{vert}^C is reported in the following paragraphs.

Vertex estimation in $\{C\}$. p_{vert}^C were computed as stated in the following equation:

$$p_{\text{vert}}^C = T_C^W * p_{\text{vert}}^W. \quad (1)$$

The vertices p_{vert}^W were geometrically identified in $\{W\}$ knowing their distances from the origin of the chessboard reference system,

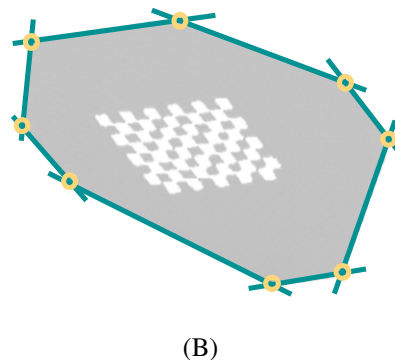
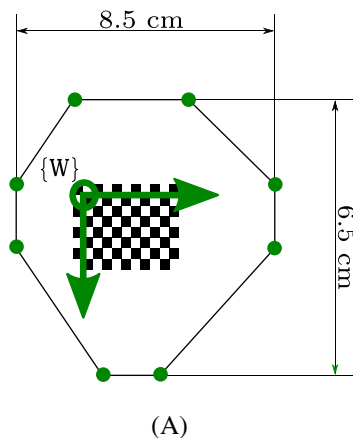


FIGURE 6 (a) Vertex estimation in $\{C\}$: view of the calibration plate. The vertex points (green dots) are at known distances from the origin of $\{W\}$. (b) Vertex estimation in $\{L\}$: view of the point cloud of the calibration plate. The vertex coordinates in $\{L\}$ (yellow circles) were calculated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

and T_C^W was computed using the Stereo Camera Calibrator Toolbox of Matlab 2015b (The MathWorks, Inc.).^{23,24}

Vertex estimation in $\{L\}$. The pipeline for identification of p_{vert}^L is as follows.

- The points belonging to the calibration plate were manually selected from the point cloud of the laser scan (removing uninformative points belonging to the background).
- Noise reduction was carried out by estimating the plane of the calibration plate according to the Maximum Likelihood Estimation Sample Consensus (MLESAC),²⁵ and projecting on the estimated plane all the points nearer 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 by searching for the minimum and maximum values of the coordinates x and y for each row and column of the discretized point cloud; (2) the points belonging to each edge were manually selected and the corresponding line was estimated.
- The p_{vert}^L were computed as the intersection of each pair of lines, as in Figure 6(b).

The scan of the calibration plate and the image acquisition were performed consecutively in order to avoid interference between the laser and the camera.

Once p_{vert}^L and p_{vert}^C were identified, T_C^L was estimated by solving equation (2) with the singular value decomposition (SVD) method:

$$p_{\text{vert}}^C = T_C^L * p_{\text{vert}}^L. \quad (2)$$

The mathematical solution was guaranteed by using more than three non-collinear points,^{26,27} namely the 8 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 $\{C\}$ and the 3D reconstructed point cloud, the RF was stored as a 2D map. Each cell (u, v) of the map contains the 3D coordinates (x, y, z) of the point projected on the image plane using the intrinsic parameters of the

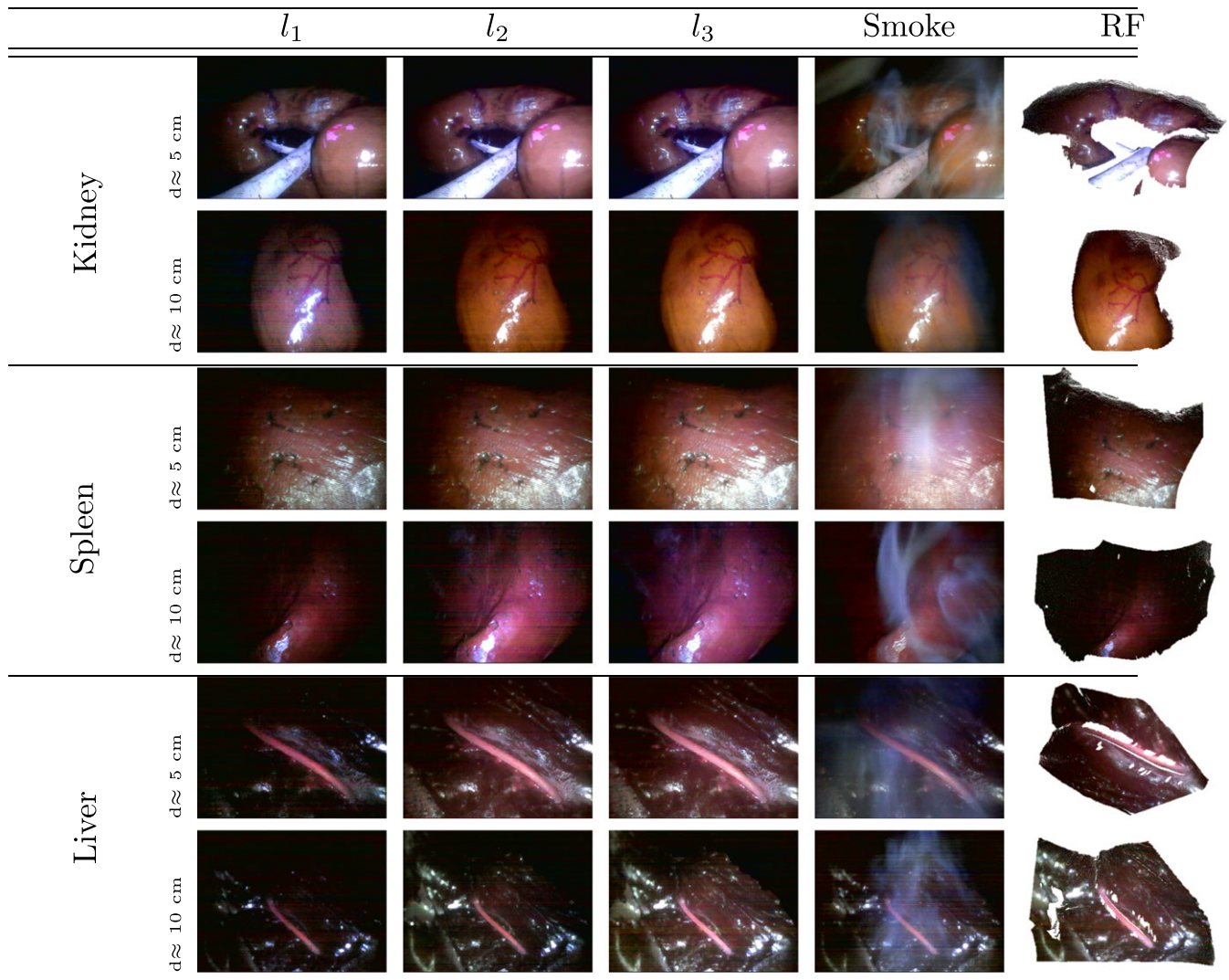


FIGURE 7 Example of endoscopic stereo images in the dataset. All the different conditions are represented (distances, levels of light and smoke) for one pose of the spleen, kidney and liver

left camera. The projection does not take into account the stereo camera rectification.

2.3 | Acquisition scenario

For the acquisition of the *EndoAbS* dataset, the laser scanner and the stereo endoscope were positioned so as to have approximately the same field of view (see Figure 5). The stereo images and the scans were captured separately, minimising the time interval between acquisitions as far as possible 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 illumination in the internal abdomen during a surgical procedure. In order to test the robustness of 3D stereo reconstruction algorithms under different conditions, the images were created with the introduction of (i) the presence of smoke, created by immersing dry-ice in hot water; (ii) 3 different endoscopic light levels (l_1 , l_2 , l_3) to vary the light intensity; (iii) 2 different phantom–endoscope distances ($\text{dist}_{\min} \approx 5$ cm and $\text{dist}_{\max} \approx 10$ cm). Sample images of the dataset are shown in Figure 7.

Thus, the obtained dataset consists of:

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

TABLE 2 Description of the structure of the *EndoAbS* dataset, 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 6 images: 3 images with different levels of illumination (l_1 , l_2 , l_3) and 3 with smoke.

3 | EXPERIMENTAL EVALUATION

3.1 | Evaluation of errors in dataset generation

The errors involved in the dataset generation are introduced by (i) characteristics of the camera and the laser scanner; (ii) camera calibration; (iii) the strategy for identification of vertices in $\{C\}$ and $\{L\}$. 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 Figure 5. In each set, 9 orientations of the calibration plate (Figure 8) were exploited, varying approximately $\pm 30^\circ$ along vertical and horizontal directions. These sets were used to compute T_C^L .

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

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

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

Number of image-scan pairs. The correlation between ϵ and the number of image-scan pairs used was estimated by computing T_C^L and varying the number of image-scan pairs from 1 to 10, then computing ϵ by applying the obtained T_C^L to the *test set*. The statistical correlation was evaluated through the Pearson product-moment correlation coefficient ($p_{\text{value}} < 0.05$).

Orientation of the calibration plate. The statistical relationship between ϵ and the orientation of the calibration plate was evaluated

by computing T_C^L for 9 different orientations of the calibration plate of the *validation set* (Figure 8). The Kruskal–Wallis test was performed ($p_{\text{value}} < 0.05$) to assess the presence of a statistical difference among the different orientations of the calibration plate.

A comparison of our method with a state-of-the-art calibration method²⁸ was conducted, using 18 image-scan pairs, as suggested in the paper. The calibration error ϵ was evaluated for 10 different trials of calibration for both methods, and the Kruskal–Wallis test was performed ($p_{\text{value}} < 0.05$) to assess the presence of a 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, but this compromised 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 their distance from the calibration plate.

3.2 | Evaluation of the realism of the dataset

In order to investigate the realism of the endoscopic images as regards real clinical images, evaluations of surgeons were collected as part of a questionnaire with scores on a 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 fields of expertise range from open surgery (11.1%) to robotic minimally-invasive surgery (22.2%) and laparoscopic surgery (66.7%).

The questionnaire consisted of 24 images, a sample that were taken from the *EndoAbs* dataset and were representative of the different levels of light, distances, the presence of smoke and different organs. The order of the images was randomized to provide a global overview of the

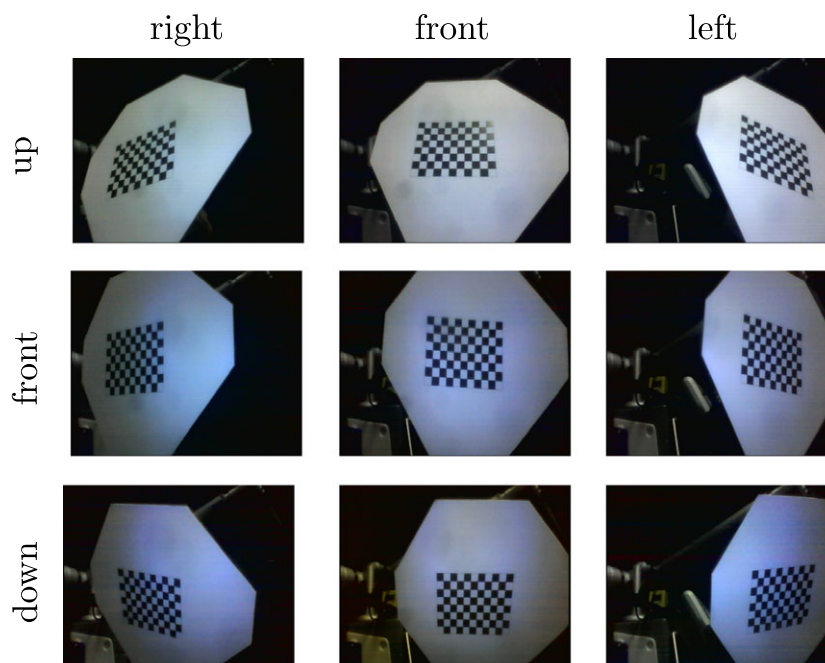


FIGURE 8 Example of orientation of the calibration plate with respect to the image plane



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 into 5 intervals (a score of 1 means very dissimilar, while 5 implies very similar). $\text{score}_{\text{mean}}$ was considered to be the average of the scores assigned to the images.

3.3 | Evaluation of 3D reconstruction

The *EndoAbs* dataset was used to evaluate a 3D reconstruction algorithm previously developed by the authors.²⁹ For the validation process, we decided to follow the terminology and methodology of the protocol for reference-based validation studies proposed by Jannin et al.,⁷ in order to allow for comparability of the results. The protocol starts by requiring the definition of the specification of the validation objective, including the clinical context and objective (C), which in our case can be identified as dense and accurate ($\text{accuracy} < 2 \text{ mm}$) 3D reconstruction for abdominal MIS surgery. Following the terminology of the protocol, the evaluated method M^{29} is referred as F_M , and Ref stands for the RF, also called the 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 incurred in the generation of the RF associated to the images and it is described in section 2.3. The hardware used for the acquisition of D_I^M and D_I^{Ref} is also described in section 2.3. The parameters P_I used by F_M are described in Table 3.

As a validation criterion, we propose n evaluation protocols consisting 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 $O_D = F_C(R_M, R_{\text{Ref}})$ 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 image plane of the unrectified left camera. 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 the region of interest, ROI) were considered in the evaluation, eliminating the areas where the organ phantom is not present, following the same criteria used by Penza et al.²⁹ The ROI was chosen on the image with the highest level of

illumination, since a 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 the presence of smoke.

Percentage of reconstructed points. This was computed as the ratio of the number of reconstructed points to the number of RF points, both identified in the region of interest.

Robustness. The algorithm was applied to the entire dataset, considering different illumination levels I_1 , I_2 and I_3 , distances dist_{min} and dist_{max} between the endoscope and the organs, and the presence of smoke. A non-parametric test (Kruskal–Wallis $p_{\text{value}} < 0.05$) was performed to test if the *accuracy* was statistically different when varying (i) I_1 , I_2 and I_3 , (ii) dist_{min} and dist_{max} ($\approx 5 \text{ cm}$ and $\approx 10 \text{ cm}$) considering only I_3 , and (iii) the presence of smoke against I_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 when using 50% of softening agent the Young's modulus of the polyurethane material (0.97 kPa) is comparable with that of the liver (see section 2.1). A cost analysis of the moulding process is reported in Table 4.

4.2 | Dataset error

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

A statistical difference was found when comparing the calibration error of the presented method and that of the method proposed by Unnikrishnan and Hebert.²⁸ There are two main causes of this difference. The first is attributed to the filtering of the laser scanner data during the plane estimation in the proposed method, as described in section 2.2.1. The second is related to the fact that Unnikrishnan and Hebert²⁸ use 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 Table 5. The evaluation of

TABLE 3 Parameters P_I used by F_M (3D reconstruction algorithm)

| F_M Parameters | value |
|--|----------------|
| Census window | 9×9 |
| Census block size | 11×11 |
| Threshold spurious remover | 10 |
| Threshold Left-Right Consistency Check | 4 |
| LO-RANSAC max iteration | 100 |
| Number of super pixels | 70 |
| Disparity range | 150–250 |

TABLE 4 Costs of the abdominal phantom

| Organ | Moulds € | Polyurethane € | Total € |
|---------|----------|----------------|---------|
| Spleen | 80 | 30 | 110 |
| Kidney | 40 | 10 | 50 |
| Liver | 170 | 90 | 260 |
| Total € | | | 420 |

TABLE 5 Camera–laser calibration errors

| | ϵ (mm) | Q_1 (mm) | Q_3 (mm) |
|--------------------------------|-----------------|------------|------------|
| State of the art ²⁸ | 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 section 3

TABLE 6 Errors in generation of the *EndoAbS* dataset

| | |
|--|---|
| Laser scanner accuracy* | $x = \pm 0.22$ mm $y = \pm 0.16$ mm $z = \pm 0.07$ mm |
| Mean reprojection error (left camera) | 0.250 pixels |
| Mean reprojection error (right camera) | 0.235 pixels |
| Camera–laser calibration error | 0.43mm |

* Conditions: distance 0.6 m, temperature 20°C, relative humidity 65% or less

the camera–laser calibration using only one image–scan pair showed an error equal to 0.43mm ($Q_1 = 0.41$ mm, $Q_3 = 0.43$ mm).

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 Table 6.

4.3 | Qualitative evaluation of the dataset

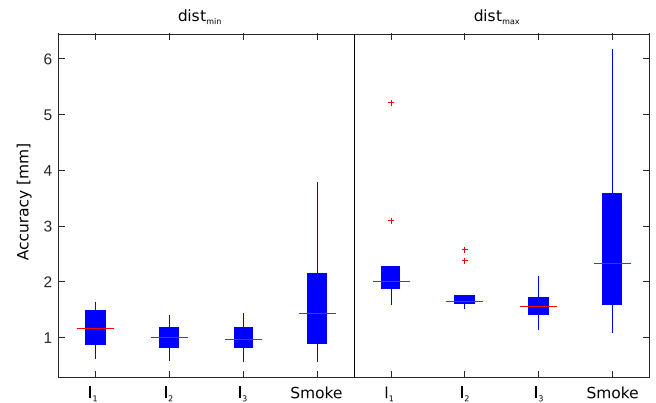
Figure 9 presents a box plot summarizing the score assigned to the 24 sample images included in the questionnaire. $\text{score}_{\text{mean}}$ is 2.7 (± 0.50). Note that the images on the x-axis are presented in the same order as they appeared in the questionnaire.

4.4 | Evaluation of the 3D reconstruction

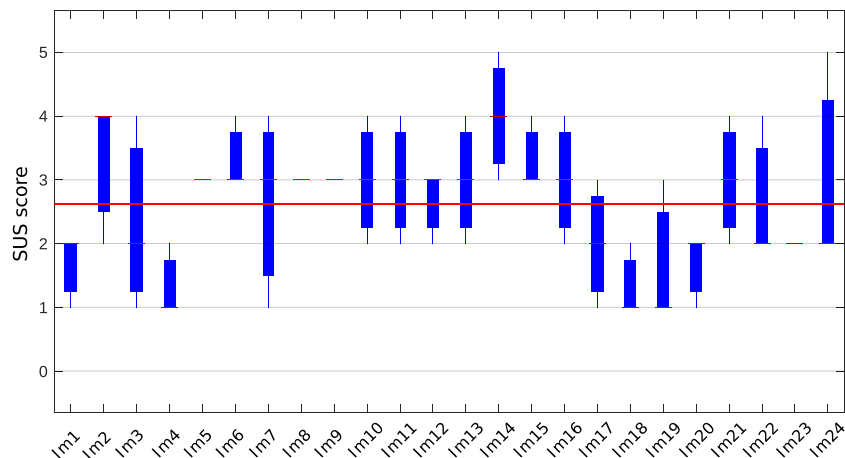
In Table 7, the accuracy and the percentage of reconstructed points are reported. In the case of dist_{min} , the accuracy was not statistically different for 3 levels of illumination ($p_{\text{value}} = 0.38$), and with or without the presence of smoke ($p_{\text{value}} = 0.17$). In the case of dist_{max}

TABLE 7 Performance of the 3D stereo reconstruction algorithm in terms of accuracy (mean and standard deviation) and percentage of reconstructed points

| | dist_{min} | | | |
|---------------|----------------------------|-------|-------|-------|
| | I_1 | I_3 | I_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} | | | |
| | I_1 | I_3 | I_3 | s |
| 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 10** Boxplot showing the accuracy of the 3D reconstruction algorithm for varying levels of illumination (I_1, I_2, I_3) and the presence of smoke for dist_{min} (left) and dist_{max} (right)

there was a significant difference between I_1 and I_3 ($p_{\text{value}} = 0.0052$), and between I_3 and the presence of smoke ($p_{\text{value}} = 0.049$). The same test performed between dist_{min} and dist_{max} for the illumination level I_3 showed that there was a statistical difference with $p_{\text{value}} = 0.0025$. These results are shown in Figure 10. An example of the errors in the 3D reconstruction of the point cloud is shown in Figure 11 and Figure 12.

**FIGURE 9** Boxplot showing the score (from 1 to 5 considering the System Usability Scale questionnaire) assigned to 24 selected images of the *EndoAbS* dataset by the surgeons answering a questionnaire in order to evaluate the realism of the characteristics represented by the images

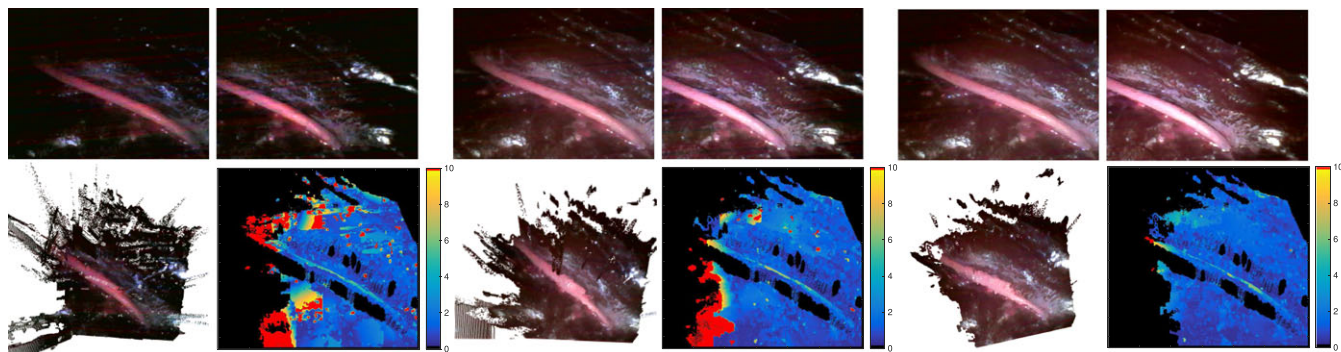


FIGURE 11 Example of results of the 3D reconstruction algorithm for varying levels of light from the left (I_1) to the right (I_3). For each pair of stereo images, the bottom row shows the error map (right) and the reconstructed point cloud (left). The colour bars represent the error in mm

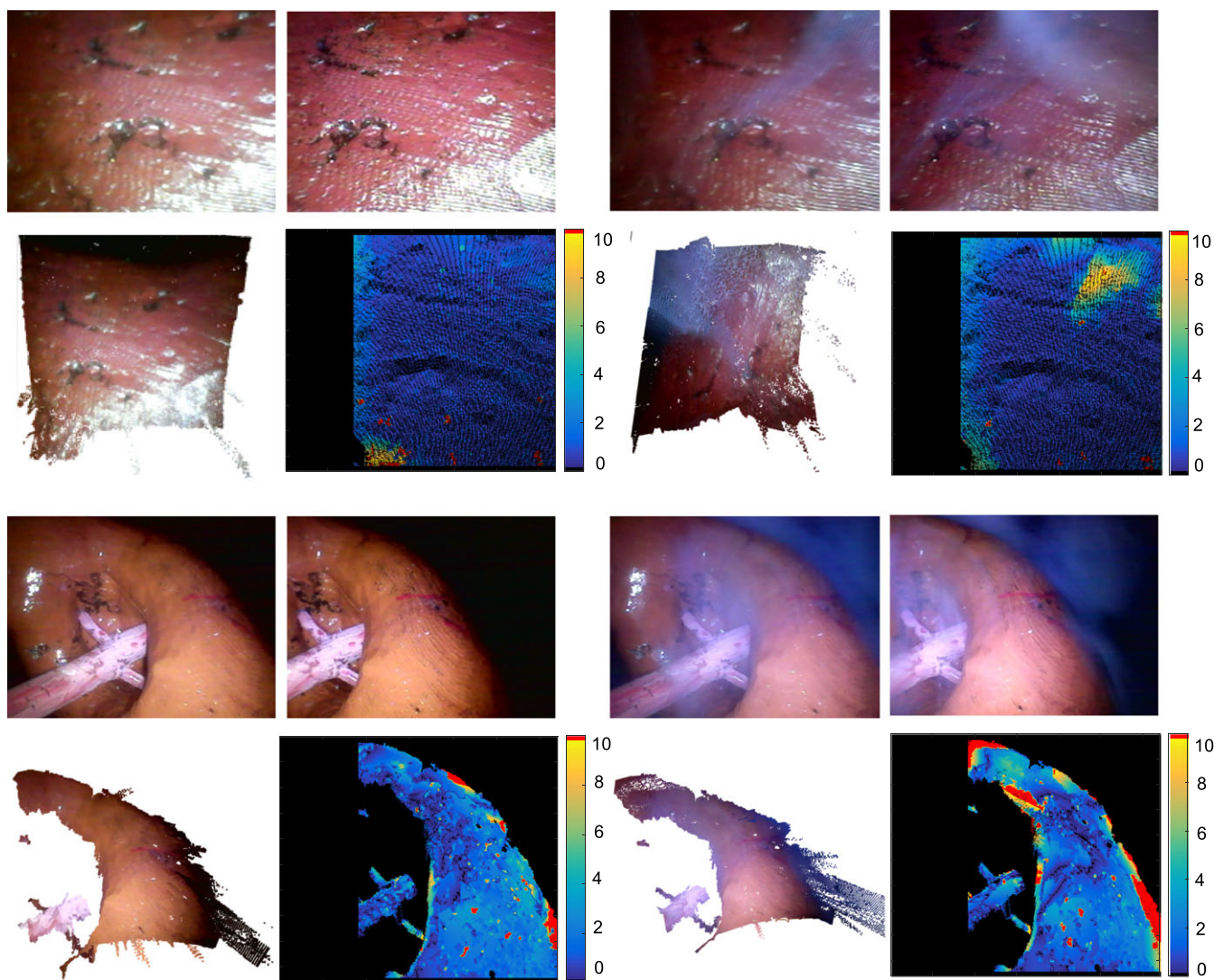


FIGURE 12 Example of results of the 3D reconstruction algorithm without and with the presence of smoke. For each pair of stereo images, the bottom row shows the error map (right) and the reconstructed point cloud (left). The colour bars represent the error in mm

5 | DISCUSSION AND CONCLUSION

This paper describes the creation of the *EndoAbS* dataset for the quantitative evaluation of 3D reconstruction algorithms based on stereo images. The dataset consists of 120 endoscopic stereo images and the associated RF. The main contribution of this work is to increase

the number and variety of openly available datasets of surgical stereo images, which are essential for testing and benchmarking 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 process of creating the dataset, particularly the camera–laser calibration error, which represents an overall estimation



of the dataset error and is an important parameter for proper assessment of reconstruction algorithms. In addition, a detailed description of the development of phantoms and the methods used was provided to facilitate future expansion of the *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, a phantom does not suffer changes in shape in the short term and can be reused 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 the phantom approximately the same stiffness as real tissues.

Moreover, in order to make the models as realistic as possible, the organ surfaces were painted to emulate tissue texture and superficial tiny vessels. Big vessels were also reproduced to allow the evaluation of 3D reconstruction algorithms in the case of more complicated structures and at different depths. Nevertheless, not all of these realistic properties have yet been exploited for the generation of images in the dataset. 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 to show 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 reference system of the left camera. 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 0.43 mm). When using other methods, e.g. that of Unnikrishnan and Hebert,²⁸ 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 those previously reported by Penza et al.²⁹ 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 in the presence of smoke. In the case of greater distances from the 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 decrease, directly affecting the performance of the algorithm. 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 standard commercial equipment, due to their high cost. Such an endoscope does provide images of

lower resolution than modern clinical devices. Moreover, the images were captured with the endoscope in a fixed position, and thus they do not reproduce the quivering behaviour due to the manipulation of the endoscope by the clinician, better simulating the conditions of robotic surgery, where the camera is moved using a robotic arm.

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

FINANCIAL SUPPORT

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

CONFLICT OF INTEREST

The authors declare that there are no known conflicts of interest associated with this publication.

ORCID

Veronica Penza  <http://orcid.org/0000-0003-1096-560X>

REFERENCES

- Nicolau S, Soler L, Mutter D, Marescaux J. Augmented reality in laparoscopic surgical oncology. *Surg Oncol*. 2011;20(3):189-201.
- Bernhardt S, Nicolau SA, Soler L, Doignon C. The status of augmented reality in laparoscopic surgery as of 2016. *Med Image Anal*. 2017;37:66-90.
- Okamoto T, Onda S, Yanaga K, Suzuki N, Hattori A. Clinical application of navigation surgery using augmented reality in the abdominal field. *Surg Today*. 2015;45(4):397-406.
- Penza V, Ortiz J, De Momi E, Forgione A, Mattos L. Virtual assistive system for robotic single incision laparoscopic surgery. In: 4th Joint Workshop on Computer/Robot Assisted Surgery; 2014; Genoa, Italy. 52-55.
- Stoyanov D. Surgical vision. *Ann Biomed Eng*. 2012;40(2):332-345.
- Scharstein D, Szeliski R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int J Comput Vis*. 2002;47(1-3):7-42.
- Jannin P, Grova C, Maurer CR Jr. Model for defining and reporting reference-based validation protocols in medical image processing. *Int J Comput Assist Radiol Surg*. 2006;1(2):63-73.
- Scharstein D, Hirschmüller H, Kitajima Y, et al. High-resolution stereo datasets with subpixel-accurate ground truth. In: German conference on pattern recognition; 2014; Münster, Germany. 31-42.
- Hu M, Penney G, Edwards P, Figl M, Hawkes DJ. 3D reconstruction of internal organ surfaces for minimal invasive surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention; 2007; Brisbane, Australia. 68-77.
- Röhl S, Bodenstedt S, Suwelack S, et al. Dense gpu-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration. *Med Phys*. 2012;39(3):1632-1645.
- Mountney P, Yang G-Z. Motion Compensated Slam for Image Guided Surgery. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2010*. Berlin, Heidelberg: Springer; 2010: 496-504.



12. Maier-Hein L, Groch A, Bartoli A, other. Comparative validation of single-shot optical techniques for laparoscopic 3-d surface reconstruction. *IEEE Trans Med Imaging*. 2014;33(10):1913-1930.
13. Lin J, Clancy NT, Stoyanov D, Elson DS. 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; 2015; Cham. 405-412.
14. Stoyanov D, Scarzanella MV, Pratt P, Yang G-Z. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention; 2010; Beijing, China. 275-282.
15. Pratt P, Stoyanov D, Visentini-Scarzanella M, Yang G-Z. Dynamic guidance for robotic surgery using image-constrained biomechanical models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention; 2010; Berlin, Heidelberg: 77-85.
16. Penza V, Bacchini S, Ciullo AS, De Momi E, Forgione A, Mattos LS. Label-based optimization of dense disparity estimation for robotic single incision abdominal surgery. In: Proceedings of the Hamlyn Symposium on Medical Robotics; 2015; London, UK. 79-80.
17. Maier-Hein L, Mountney P, Bartoli A, et al. Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. *Med Image Anal*. 2013;17(8):974-996.
18. Ciullo AS, Penza V, Mattos L, De Momi E. Development of a surgical stereo endoscopic image dataset for validating 3D stereo reconstruction algorithms. In: 6th Joint Workshop on New Technologies for Computer/Robot Assisted Surgery; 2016; Pisa, Italy.
19. Condino S, Carbone M, Ferrari V, et al. How to build patient-specific synthetic abdominal anatomies. an innovative approach from physical toward hybrid surgical simulators. *Int J Med Rob Comput Assist Surg*. 2011;7(2):202-213.
20. Tirella A, Mattei G, Ahluwalia A. Strain rate viscoelastic analysis of soft and highly hydrated biomaterials. *J Biomed Mater Res Part A*. 2014;102(10):3352-3360.
21. Yeh W-C, Li P-C, Jeng Y-M, et al. Elastic modulus measurements of human liver and correlation with pathology. *Ultrasound Med Biol*. 2002;28(4):467-474.
22. Mattei G, Tirella A, Gallone G, Ahluwalia A. Viscoelastic characterisation of pig liver in unconfined compression. *J Biomech*. 2014;47(11):2641-2646.
23. Zhang Z. A flexible new technique for camera calibration. *IEEE Trans Pattern Anal Mach Intell*. 2000;22(11):1330-1334.
24. Heikkila J, Silvén O. A four-step camera calibration procedure with implicit image correction. In: Proceedings., 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 1997; San Juan, Puerto Rico. 1102-1112.
25. Torr PH, Zisserman A. MLESAC: A new robust estimator with application to estimating image geometry. *Comput Vis Image Underst*. 2000;78(1):138-156.
26. Horn BK. Closed-form solution of absolute orientation using unit quaternions. *JOSA A*. 1987;4(4):629-642.
27. Arun KS, Huang TS, Blostein SD. Least-squares fitting of two 3-D point sets. *IEEE Trans Pattern Anal Mach Intell*. 1987;5:698-700.
28. Unnikrishnan R, Hebert M. *Fast Extrinsic Calibration of a Laser Rangefinder to a Camera*. Pittsburgh, PA: Carnegie Mellon University; 2005.
29. Penza V, Ortiz J, Mattos LS, Forgione A, De Momi E. Dense soft tissue 3D reconstruction refined with super-pixel segmentation for robotic abdominal surgery. *Int J Comput Assist Radiol Surg*. 2015;11:1-10.

How to cite this article: Penza V, Ciullo AS, Moccia S, Mattos LS, De Momi E. EndoAbS Dataset: Endoscopic abdominal stereo image dataset for benchmarking 3D stereo reconstruction algorithms. *Int J Med Robotics Comput Assist Surg*. 2018;e1926. <https://doi.org/10.1002/rcs.1926>