Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

SERV-CT: A disparity dataset from cone-beam CT for validation of endoscopic 3D reconstruction

P.J. Eddie Edwards [a,*], Dimitris Psychogyios [a], Stefanie Speidel [b], Lena Maier-Hein [c], Danail Stoyanov [a]

[a] Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London (UCL), Charles Bell House, 43-45 Foley Street, London W1W 7TS, UK
[b] Division of Translational Surgical Oncology, National Center for Tumor Diseases (NCT) Dresden, Dresden, 01307, Germany
[c] Division of Medical and Biological Informatics, German Cancer Research Center (DKFZ), Heidelberg, Germany

ARTICLE INFO

ABSTRACT

In computer vision, reference datasets from simulation and real outdoor scenes have been highly successful in promoting algorithmic development in stereo reconstruction. Endoscopic stereo reconstruction for surgical scenes gives rise to specific problems, including the lack of clear corner features, highly specular surface properties and the presence of blood and smoke. These issues present difficulties for both stereo reconstruction itself and also for standardised dataset production. Previous datasets have been produced using computed tomography (CT) or structured light reconstruction on phantom or *ex vivo* models. We present a stereo-endoscopic reconstruction validation dataset based on cone-beam CT (SERV-CT). Two *ex vivo* small porcine full torso cadavers were placed within the view of the endoscope with both the endoscope and target anatomy visible in the CT scan. Subsequent orientation of the endoscope was manually aligned to match the stereoscopic view and benchmark disparities, depths and occlusions are calculated. The requirement of a CT scan limited the number of stereo pairs to 8 from each *ex vivo* sample. For the second sample an RGB surface was acquired to aid alignment of smooth, featureless surfaces. Repeated manual alignments showed an RMS disparity accuracy of around 2 pixels and a depth accuracy of about 2 mm. A simplified reference dataset is provided consisting of endoscope image pairs with corresponding calibration, disparities, depths and occlusions covering the majority of the endoscopic image and a range of tissue types, including smooth specular surfaces, as well as significant variation of depth. We assessed the performance of various stereo algorithms from online available repositories. There is a significant variation between algorithms, highlighting some of the challenges of surgical endoscopic images. The SERV-CT dataset provides an easy to use stereoscopic validation for surgical applications with smooth reference disparities and depths covering the majority of the endoscopic image. This complements existing resources well and we hope will aid the development of surgical endoscopic anatomical reconstruction algorithms.

1. Introduction

In minimally invasive surgery (MIS), endoscopic visualisation facilitates procedures performed through small incisions that have the potential advantages of lower blood loss and infection rates than open surgery as well as better cosmetic outcome for the patient. Despite the potential advantages of MIS, working within the limited endoscopic field-of-view (FoV) can make surgical tasks more demanding, which may lead to complications and adds significantly to the learning curve for such procedures.

Techniques such as augmented reality (AR) can be used to enhance preoperative surgical visualization and provide detailed, multi-modal anatomical information. This is a significant research field in itself, with AR abdominal laparoscopic applications having been proposed in various fields, including surgery of the liver (Hansen et al., 2010; Thompson et al., 2018; Quero et al., 2019), uterus (Bourdel et al., 2017), kidney and prostate (van Oosterom et al., 2018; Bertolo et al., 2020; Hughes-Hallett et al., 2014). The application-specific literature in all these areas as well as more

---

* Corresponding author.
E-mail address: eddie.edwards@ucl.ac.uk (P.J.E. Edwards).

general reviews of AR surgery (Bernhardt et al., 2017; Edwards et al., 2021) all conclude that the technology shows great promise, but also note that alignment accuracy, workflow integration and perceptual issues have limited clinical uptake.

Vision has the potential to provide both the 3D shape of the surgical site and also the relative location of the camera within the 3D anatomy, especially when stereoscopic devices are used in robotic MIS (Mountney et al., 2010; Stoyanov, 2012). A range of optical reconstruction approaches have been explored for endoscopy with computational stereo being by far the most popular due to the clinical availability of stereo endoscopes (Maier-Hein et al., 2013; Röhl et al., 2012; Chang et al., 2013). Despite recent major advances in computational stereo algorithms (Zhou and Jagadeesan, 2019), especially with deep learning models, in the surgical setting robust 3D reconstruction remains difficult due to various challenges including specular reflections and dynamic occlusions from smoke, blood and surgical tools.

While stereo endoscopy is routinely used in robotic MIS, the majority of endoscopic surgical procedures are performed using monocular cameras. With monocular endoscopes, 3D reconstruction can be approached as a non-rigid structure-from-motion (SfM) or simultaneous localisation and mapping (SLAM) problem (Grasa et al., 2013). This is typically more challenging than stereo because non-rigid effects and singularities need to be accounted for as the camera moves within the deformable surgical site. Alternative vision cues such as shading have also been explored historically but with limited success until recent promising results from monocular single image 3D using deep learning models (Mahmoud et al., 2018; Bae et al., 2020). Such approaches have been applied in general abdominal surgery (Lin et al., 2016), sinus surgery (Liu et al., 2018), bronchoscopy (Visentini-Scarzanella et al., 2017), and colonoscopy (Mahmood and Durr, 2018; Rau et al., 2019). While monocular reconstruction has been shown to be feasible and extremely promising, the accuracy and robustness of the surfaces produced requires improvement and further development is needed.

Whatever the chosen method for 3D reconstruction, evaluation of the reconstruction accuracy by establishing appropriate benchmark datasets has been a major hurdle impeding development of the field. Standardised datasets providing accurate references have been instrumental in the rapidly advancing development of 3D reconstruction algorithms in computer vision (Scharstein and Szeliski, 2003; Scharstein et al., 2014; Menze and Geiger, 2015). In surgical applications, however, it has proved difficult to produce accurate standardised datasets in a form that facilitate easy and widespread use and adoption. Assessment of reconstruction accuracy requires 3D reference information and during in vivo surgical procedures this is not currently available. Phantom models made from synthetic materials have been used as surrogate environments with a corresponding gold standard from CT[1] (Stoyanov et al., 2010). The CT model is registered to the stereo-endoscopic view using fiducials and dynamic CT is used to provide low frequency estimates of the phantom motion. More recently, the EndoAbS dataset was reported with stereo-endoscopic images of phantoms with gold standard depth provided by a laser rangefinder (Penza et al., 2018)[2]. Many challenging images are presented, including low light levels and smoke, and the dataset concentrates on the robustness of algorithms to these conditions. One of the main problems with phantom environments is their limited representation of the visual complexity of real *in vivo* images though some exciting progress in phantom fabrication and design,

such as the work by Ghazi et al. (2017), may overcome this in the future.

Stereo-endoscopic datasets have also been created using *ex vivo* animal environments. The first contains endoscopic images of samples from different porcine organs (liver, kidney, heart) captured from various angles and distances including examples with smoke and artificial blood (Maier-Hein et al., 2014). Gold standard 3D reconstruction is provided within a masked region for each stereo pair from CT scans registered using markers visible in both the CT scan and the endoscopic images. Analysis tools are also provided within the MITK[3] framework. The area of analysis is restricted to fairly small regions near the centre of the images leading to a comparatively narrow range of depths and tissue types within one sample.

Most recently, as part of the Endoscopic Vision (EndoVis) series of challenges, the 2019 SCARED challenge[4] provides per-pixel depth ground truth annotations as a gold standard to be used for training of learning based methods or to evaluate reconstruction algorithms (Allan et al., 2021). The HD endoscope images from *ex vivo* samples are brightly lit and the coverage of scene is excellent. Since reconstruction occurs from data captured directly from the endoscope's camera there is no requirement for registration. The SCARED dataset is an important contribution to surgical stereo reconstruction covering a range of tissues with frames including surgical tools and blood. There are some limitations, however. Briefly those are, the format of the provided dataset, occasional outliers and artefacts in the ground truth and inconsistent calibration for some frames, all of which are mentioned in the challenge paper and the reader can refer to section 5.1 for more details.

To overcome some of these limitations, we examine the feasibility of producing such a dataset using a cone-beam CT scan encompassing both the endoscope and the viewed anatomy. The contributions of this work are as follows:

1. The release of an *ex vivo* stereo endoscopic dataset and evaluation toolkit, aiming to assist the development of reconstruction algorithms in the medical field.
2. The detailed description of the data acquisition and processing protocol used and the release of the manual alignment software used for registration.
3. The evaluation of several stereo correspondence algorithms using our verification set.

2. The SERV-CT reference dataset construction

The working hypothesis of this research is that a CT scan containing both the endoscope and the viewed anatomical surface can be used to provide a sufficiently accurate reference for stereo reconstruction validation. To establish whether this is feasible, multiple stereo-endoscopic views of two *ex vivo* porcine full torsos were taken. A schematic of the overall process is shown in Fig. 1 and the range of views can be seen in Fig. 2(b). As a secondary, aim we examine whether a textured RGB surface model can facilitate registration, particularly where there are few visible geometrical features to use for alignment.

2.1. Equipment

As our anatomical model, we used two fresh *ex vivo* full torso porcine cadavers including thorax and abdomen. In this study, we have focused on the abdomen as this is where most endoscopic surgery occurs. The outer tissue layers were removed by a clinical
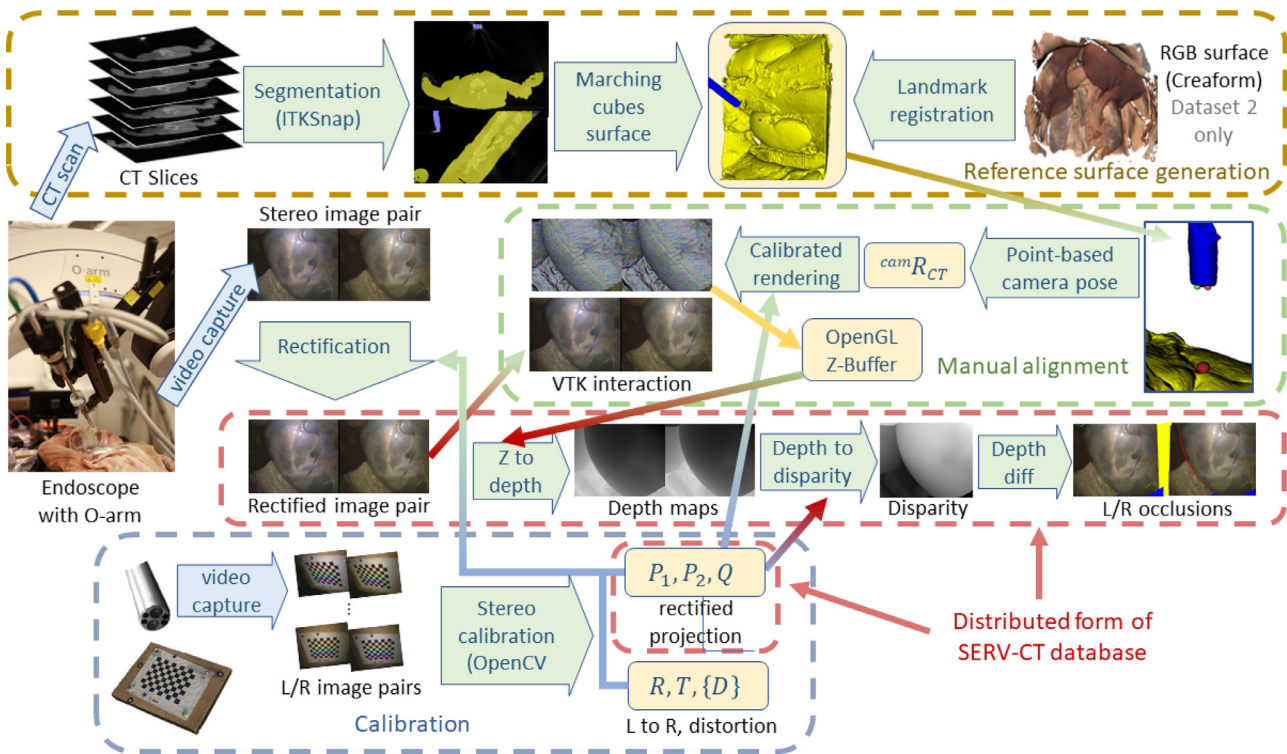

Fig. 1. A schematic flowchart showing the process of dataset generation, processing and alignment. Corresponding cone-beam CT and stereo endoscope images are taken (middle left). The top row (brown box) shows the process of anatomical surface and endoscope segmentation detailed in section 2.3. The output is the segmented volume, which forms the input to the interactive manual alignment (see section 2.6 for details). The resulting OpenGL Z-buffer of the aligned rendering can generate both depth and disparity maps by the process described in section 2.6.2. The blue box (bottom row) shows the standard OpenCV stereo camera calibration process (see section 2.2). This generates left/right projection matrices, the Q matrix that relates disparity to depth, rigid left-to-right transformation and distortion coefficients. These feed into the rectification process, rectified manual alignment and disparity-to-depth calculation as shown. In the red boxes we have the final form of the released dataset, including rectified stereo image pairs, calibration parameters, depth and disparity maps as well as occlusion areas for both left and right images.
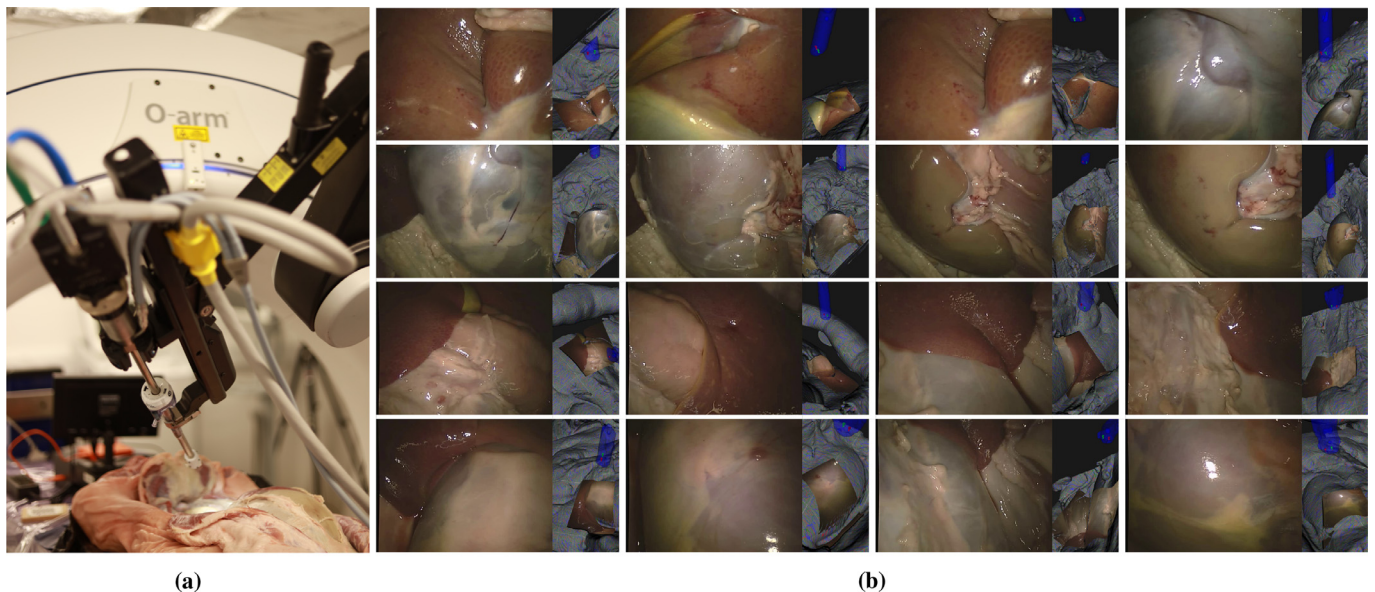

Fig. 2. The equipment setup is shown, with the endoscope attached to the da Vinci[TM] surgical robot placed within the O-arm[TM] interventional scanner (a). The left images of all the views from the dataset alongside renderings of each endoscope view (b). A range of tissue types is evident. The top row shows features with interesting variation of depth (image 1 and 3 are chosen to be similar to assess repeatability). The second row shows the kidney at different depths, with and without renal fascia. The third row shows the liver and surrounding tissues from the second *ex vivo* sample. The bottom row includes some smooth, featureless and highly specular regions from the kidney of sample 2. There is considerable variation of depth in most of these images (see Fig. 3).
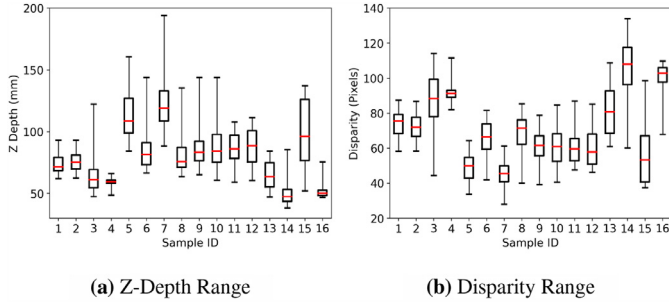
(a) Z-Depth Range   **(b)** Disparity Range

Fig. 3. Z-Depth and disparity ranges calculated from the reference standard for each sample in our dataset showing wide variation of depths provided by the sample images.

colleague in a manner that mimics surgical intervention to reveal the inner organs.

CT scans were provided using the O-arm^{TM} Surgical Imaging System (Medtronic Inc., Dublin, Ireland). This is a cone-beam CT interventional scanner. For simplicity, we will refer to the images from this device as CT scans throughout the paper. For the second dataset, in order to facilitate alignment in very smooth anatomical regions with few geometric features, the sample torso was also scanned with a Creaform Go SCAN 20 hand-held scanner (Creaform Inc., Lévis, Canada) to provide a textured, triangulated RGB surface.

Endoscope images are collected using a first generation da Vinci^{TM} surgical system (Intuitive Surgical, Inc., Sunnyvale, CA, USA), which can also be used to manipulate the endoscope position. This robotic surgical setup is not ideally designed to fit within the confines of the O-arm^{TM} and positioning requires considerable angulation of both the robot setup joints and the O-arm^{TM} itself. The setup can be seen in Fig. 2(a), showing the robotic endoscope within the O-arm^{TM} and the *ex vivo* model. Images were gathered in two separate experiments using the straight and 30° endoscopes supplied with the da Vinci^{TM} system.

2.2. Endoscope calibration

Endoscope calibration follows a standard chessboard OpenCV[5] stereo calibration protocol. Images of a chessboard calibration object are taken from multiple viewpoints. The corners detected, enable intrinsic calibration of the two cameras of the stereo endoscope. The same chessboard pattern viewed in corresponding left and right views can also provide the transformation from the left to the right camera. We used 18 images pairs in Expt. 1 and 14 image pairs in Expt. 2 covering the endoscope's filed of view and a range of depths.

Once the left and right images have been rectified, the projection for each eye is given by two matrices:

$$P_1 = \begin{bmatrix} f & 0 & C_x^1 & 0 \\ 0 & f & C_y^1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, P_2 = \begin{bmatrix} f & 0 & C_x^2 & T_x f \\ 0 & f & C_y^2 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (1)$$

Their elements combine to give a matrix, Q, that relates the disparity, δ, at a pixel (u, v) to a to 3D location (x, y, z):

$$k \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = Q \begin{bmatrix} u \\ v \\ \delta \\ 1 \end{bmatrix} \quad (2)$$

---

5 https://opencv.org/.

where

$$Q = \begin{bmatrix} 1 & 0 & 0 & -C_x^1 \\ 0 & 1 & 0 & -C_y^1 \\ 0 & 0 & 0 & f \\ 1 & 0 & \frac{-1}{T_x} & C_x^1 - C_x^2 \end{bmatrix} \quad (3)$$

In addition to multiple views being used for camera calibration, several CT scans depicting the endoscope and the chessboard are taken. Six fixed spherical ceramic coloured markers can be seen in both the CT scan and the endoscope views (see Fig. 6(a)). These can be used to validate our alignment method and provide accuracy measurement of the CT alignment process.

The calibration process is depicted boxed in blue in the schematic flowchart (Fig. 1). The output feeds into the image rectification, rectified rendering for manual alignment and depth-to-disparity calculations. For the simplified form of the released dataset, only the three matrices, P_1, P_2 and Q are provided alongside the rectified left and right images with corresponding reference disparities, depths and occlusions.

2.3. CT segmentation of endoscope and anatomy

In this section we describe the process of anatomy and endoscope surface generation from CT depicted in the top row of the flowchart shown in Fig. 1. As can be seen in Fig. 4(a), both the endoscope and the anatomical surface are visible in the CT scan. The ITKSnap software (version 3.8.0) is used for its convenience and simple user interface (Yushkevich et al., 2006). Automated segmentation with seeding and a single threshold is used initially. The endoscope is mostly at the fully saturated CT value so a very high threshold is used (2885). For the anatomical tissue-air interface, a very low value works well (-650 to -750). Partly due to the artifacts from the presence of the endoscope, but also because of thin membranes and air filled pockets, a single threshold does not capture all the anatomical surface. Some of the anatomical detail must be hand segmented (see Fig. 4(c)). ITKSnap provides tools for this purpose, including pencils and adaptive brushes. Hand segmentation is limited to a few small regions, with most of the anatomical surface having clear contrast in the CT image.

2.4. Identification of CT endoscope position and orientation

To relate the coordinates of the endoscope to the CT, we first mark a rough position for the endoscope cameras. By rotating the
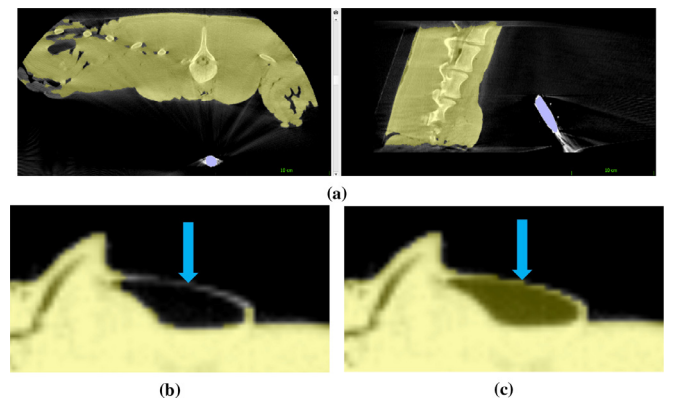

(a)

(b)   **(c)**

Fig. 4. O-arm^{TM} CT scan showing our anatomical full torso model (yellow) and the endoscope (blue) visible in the same scan (a). Streak artefacts from the presence of the metal endoscope are evident, but the viewed anatomical surface can still be accurately segmented. A single threshold does not always provide accurate segmentation (b) and these regions must be segmented by hand (c).
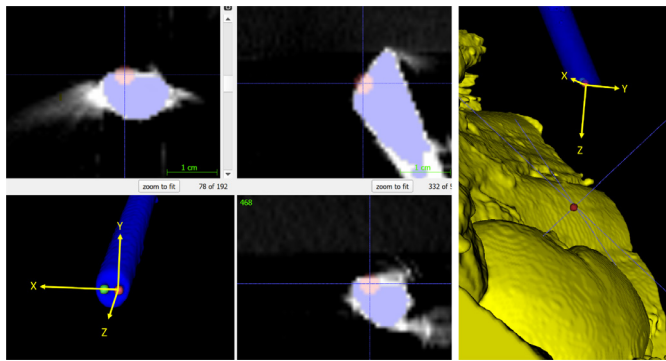
Fig. 5. Approximate left (red) and right (green) camera positions are marked on the endoscope surface in the CT scan using the segmented endoscope rendering (bottom left). The left camera position is fixed in subsequent manual alignment and a target point is identified roughly in the direction of the relevant anatomy. The resulting axes of this 30° endoscope are shown in yellow, with the Z axis pointing towards the target point. This provides an initial alignment from CT to endoscope, which can be subsequently refined manually.

surface rendered view of endoscope and anatomy, it is usually possible to identify which region of the CT is being viewed. Orienting the views accordingly, approximate positions for the left and right cameras can be marked on the end of the endoscope. It should be noted that the exact orientation is not important, as this will be adjusted manually later on. However, constraining the location of the camera to be near the end of the endoscope and largely limiting the motion to rotation about the left camera has two benefits. It reduces the complexity of the alignment task but more importantly it ensures that the distance from the camera to the anatomical surface is that described by the CT scan. Since stereoscopic disparity depends directly on this depth, a constrained motion preserving the camera position from the CT scan should lead to a more accurate reference.

In addition to the left and right cameras, a further point along the endoscope is manually identified to provide an initial approximate viewing direction. The left and right camera positions and this point define an axis system and an initial estimate of the rigid transformation from CT to stereo endoscopic camera coordinates (see Fig. 5). The purpose of this initial rough alignment is partly to ensure that in the next phase of manual alignment, the relevant anatomy can be seen in a virtual rendering from the endoscope position. By constraining the endoscope camera position, we ensure that the anatomy is viewed from the correct perspective, which is key to making an accurate reference.

2.5. Verification using a calibration object

To establish an upper limit of accuracy for our method we have collected stereo images of the calibration object with a corresponding CT scan. Six coloured spherical bearings are used as fiducial markers. These can be readily identified in both the CT scan and the endoscopic view. To establish the location of the chessboard relative to the markers, a surface was produced from multiple images from an iPhone 7 using the 3D reconstruction software, Meshroom[6]. Marking of both the fiducial markers and the chessboard points on this model provides a correctly scaled, aligned model and allows the fiducials to be expressed in chessboard coordinates (Fig. 6(a)). Registration of the fiducial markers from the Meshroom model to the CT scan gives a residual alignment error of <0.5 mm and the resulting chessboard points transformed to CT align well with the plane of the board.
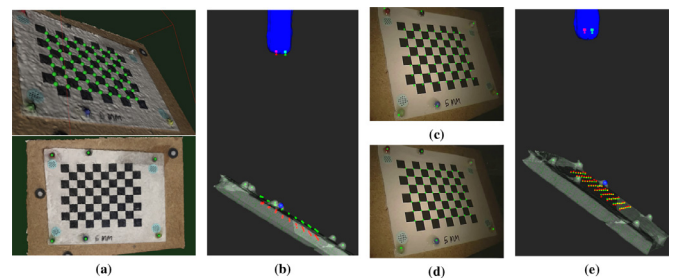
---

[6] https://github.com/alicevision/meshroom.

Fig. 6. The 6 coloured bearings and the chessboard corners are identified on a Meshroom reconstruction of the calibration object (a). This provides 3D registration between the markers and chessboard corners, and registration to the bearings locates the chessboard corners in the CT scan. Fixing the left camera position at the end of the endoscope and performing our constrained alignment to the bearings, results in the alignment shown in (b), where the green dots are the chessboard positions from CT and the red dots are the same points from stereo reconstruction. There is a small misregistration in depth resulting in a visible scaling error (c). Translation of the cameras into the endoscope shaft results in good alignment (d) and accurate reconstruction (e) with typical errors of ≈0.4 mm. This highlights the fact that the unknown position of the effective pinhole of the endoscope cameras can be estimated by allowing a small translation into the body of the endoscope. This degree of freedom is incorporated into our manual alignment process (see supplementary video).

Stereo reconstruction of the chessboard points (Fig. 6(b)) exhibits a small scaling error (Fig. 6(c)) which can be corrected for by moving the camera position back along the endoscope (Fig. 6(d)). The initial camera positions are placed on the end of the endoscope surface, but the lens arrangement and optics of the endoscope is not known, so the effective pinholes will be deeper inside the shaft. The resulting chessboard reconstruction after translation into the shaft is accurate (∼0.4 mm) and the position of the cameras within the endoscope is visually reasonable (Fig. 6(e)).

The process of manually adjusting the endoscope position and orientation to match the chessboard is the same as the alignment procedure discussed in the next section for registration of the anatomy, but the chessboard points give a measure of accuracy. Registration of the fiducial markers from the Meshroom model to the CT scan, provides the chessboard points in CT coordinates, which align well with the segmented plane of the board (Fig. 6(e)).

2.6. Manual alignment of the endoscope orientation to match the anatomical surface

In this section we describe the manual alignment process boxed in green on the flowchart (Fig 1). To achieve an accurate alignment the human eye is a very useful tool. We are able to fuse even the difficult stereo images from surgical scenes and can accurately assess depth. To make use of this human ability, we devised an interactive application that overlays the anatomical surface from the CT scan onto the stereo endoscope view. Some example renderings are shown in Fig. 7. The surface can be turned on and off, faded in and out, rendered solid textured or as lines or points. Rendering and interaction use the Visualization Toolkit (Python VTK version 8.1.2) from Schroeder et al. (2004). The VTK cameras corresponding to the left and right views are adjusted to match the OpenCV stereo rectification using elements from the SciKit-Surgery library from Thompson et al. (2020) (scikit-surgeryvtk version 0.18.1). The underlying rectified left and right images are shown as a background to the rendering. The surface can also be moved, but only the three angles of the endoscope rotation about the left camera can be adjusted. The endoscope position is constrained. This ensures that the perspective from the endoscope is maintained, ensuring that occlusions and the distance to the anatomical surface should be correct. (Fig. 8).
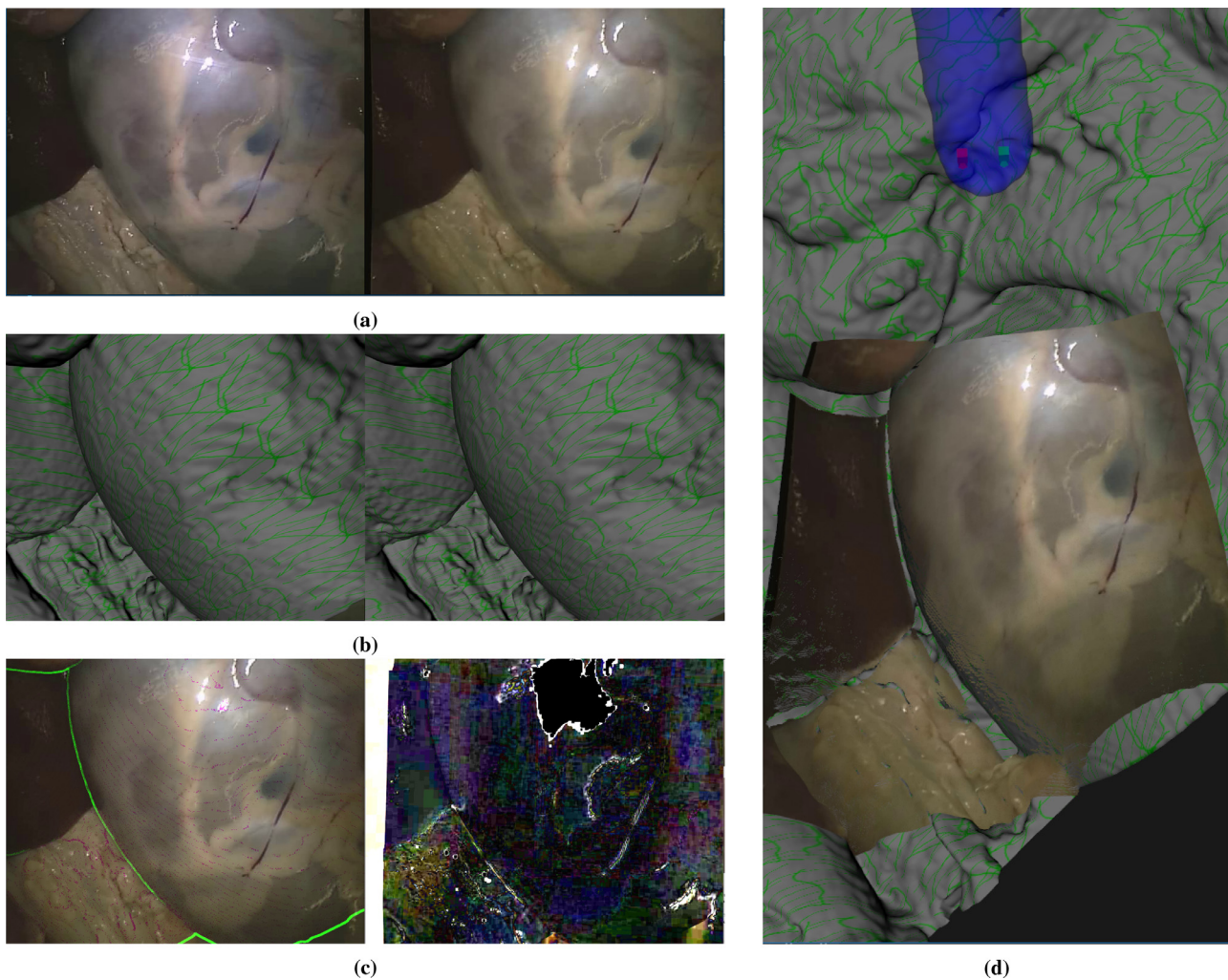
Fig. 7. The manual alignment method, showing the original image pair (a), the corresponding rendering of the CT surface (b) and processed images to help with registration (c), which include a rendering of the boundary of the CT surface in green on the left and a subtraction image showing the difference between left and right corresponding pixels according to the current registration. These hints are updated live during the manual interaction. Overlays can be displayed using different renderings (surface, wireframe or points) and manipulated to provide a more accurate alignment. Projection of the left image as a texture on the CT surface after alignment is shown in (d). This process can be seen in the supplementary videos. *In all cases the left image of a stereo pair is shown on the right to facilitate cross-eyed fusion.*

The X and Y rotations about the endoscope camera are actually very similar to translation in Y and X for small angles. The Z rotation orients the endoscope about its axis, which approximates to a 2D rotation of the image. A small translation along the endoscope axis also allowed to account for the true position of the effective pinhole within the endoscope shaft. The effect is similar to an overall scaling of the image. The left and right images are swapped for display to facilitate cross-eyed fusion of the stereo pair. Coarse and fine adjustments are made until the user is happy with the registration. This is not an easy process, requiring some concentration and skill, but usually takes only a few minutes for each image set (see Fig. 7 and the supplementary video).

2.6.1. Assessment of manual registration using repeat alignments

To assess the accuracy of manual alignment we performed repeated registrations. We use the bad3 statistic for comparison, which is the proportion of the image with greater than 3 pixels disparity error. Two people performed the manual alignment 3 times each for every image pair. Based on the evaluation study (see section 3), we chose the two best performing networks trained in generic data – HSM (Level 1) and DeepPruner – to provide independent estimates of the disparity. All error metrics were calculated and the bad3 results from each alignment can be seen in Fig. 9.

It is clear from these results that each network performs better on some images. DeepPruner is better for images 006, 007 and 008, whereas results are similar or HSM (level1) is better for all others. To reduce the effect of human error, we consider manual alignments that have a bad3 of $> 20\%$ compared to both networks as outliers. These are removed before averaging all remaining inlier alignments.

The process of averaging deserves some attention. The ith manual alignment results in a rigid transformation from CT to the left endoscope camera which can be represented by a rotation, R_i and translation T_i. We want to find the mean, $\bar{R}, \bar{T}$, of $\{R_i, T_i\}$.

Averaging a set of rotations is a well studied problem and the eigen analysis solution from Markley et al. (2007) that is freely available as a NASA report[7] is widely accepted as an optimal solution. We compute this average using the quaternion representation and provide a mean rotation matrix, $\bar{R}$.

It is sometimes suggested that averaging the translations is trivial or can be achieved by simply taking the mean of $\{T_i\}$, but this

---

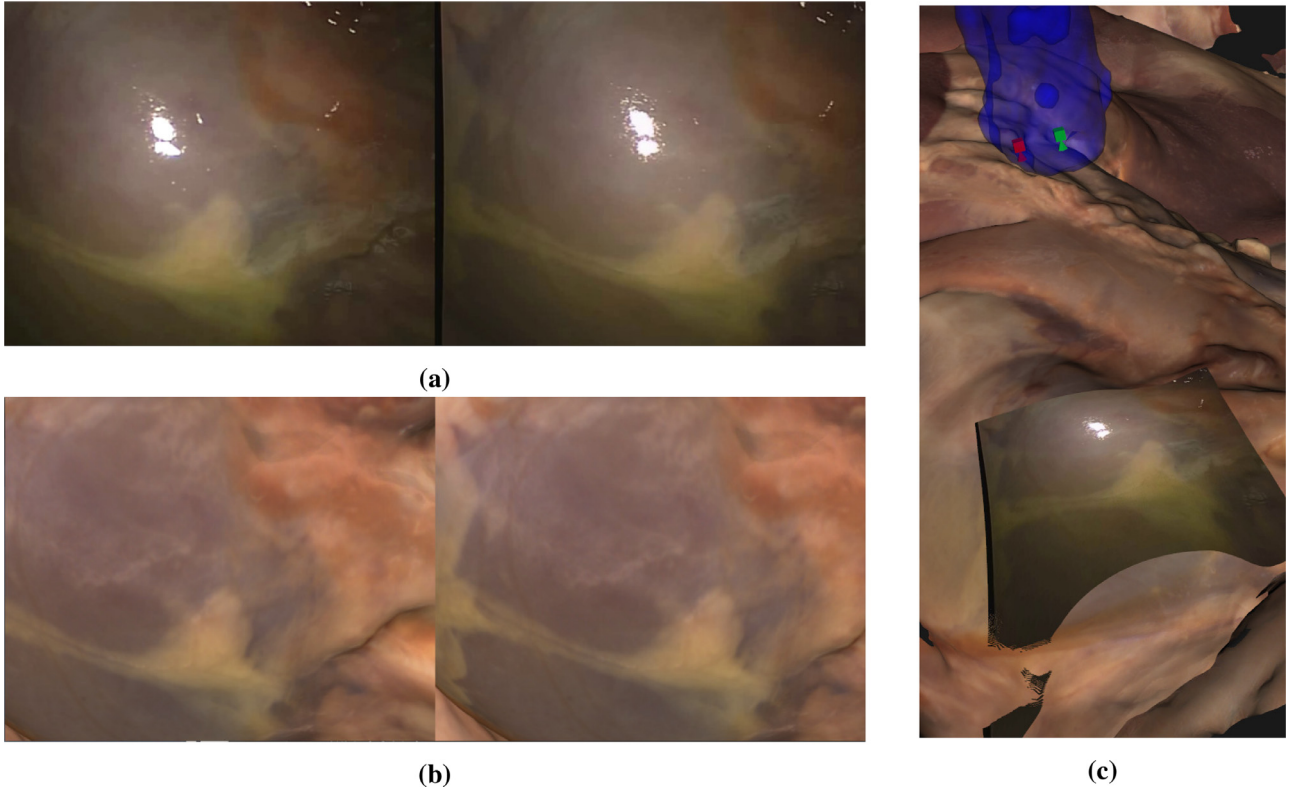[7] https://ntrs.nasa.gov/citations/20070017872.

Fig. 8. Manual alignment for a smooth surface where there is little geometrical variation to register on (a). It is possible to use the Creaform scanner RGB surface. This provides surface texture and can be readily aligned to features in the image (b). Overlay of the endoscope image projected onto the RGB surface is also shown (c). *In all cases in the paper the left image of a stereo pair is shown on the right to facilitate cross-eyed fusion.*

is not the case. This would provide the average translation of the CT origin. Our anatomical surface comes from a CT scan where both anatomy and endoscope must be visible, which means that both are pushed towards the periphery of the scan and are some distance from the origin. Using the mean of $\{T_i\}$ leads to a transformed surface that is not centrally placed relative to the manually aligned surfaces.

To obtain a more suitable average translation, we require a more relevant center of rotation in the CT space, C_{ct}. In most graphical applications this could be the centroid of an object. In our case a suitable choice would be a point on the viewed CT surface near the centre of endoscope image. For each image pair, such a point was identified on the anatomical CT surface with approximately average depth and near to the centre of the endoscope view.

Having chosen C_{ct}, calculation proceeds as follows:

$$\bar{y}_i = \frac{1}{n}\sum_i y_i \tag{4}$$

where $y_i = R_i C_{ct} + T_i$ (5)

then $\bar{T} = \bar{y}_i - \bar{R}C_{ct}$ (6)

With $\bar{R}, \bar{T}$ calculated in this way we achieve a mean rigid transformation that provides a natural average of the manual inputs. Fig. 9 shows the bad3 and depth errors for the mean (last bar in pink) compared to each of the individual manual alignments. In all cases the mean provides an error close to the best. While this does not constitute a statistical proof of correctness, the average calculated in this way is more valid than simply choosing the lowest error, which may bias towards one of the algorithms. Fig. 10 shows the results for each of the frames, with errors compared to the best

performing metric of around 2 pixels or 2 mm depth. This mean is used in all subsequent calculations and to provide the depth and disparity maps for the released version of SERV-CT.

2.6.2. Calculation of depth maps, disparity maps and regions of correspondence

Once the CT scan is aligned to the stereo endoscopic view we can produce depth maps and subsequently disparity images. The OpenGL Z-buffer provides a depth value for every pixel. For perspective projection, the normalised buffer value, Z_{gl}, has a non-linear relationship to actual depth that can be readily calculated from the near and far clipping planes, Z_{near} and Z_{far}:

$$Z_{world} = \frac{-2Z_{near}Z_{far}}{2(Z_{gl}-0.5)(Z_{far}-Z_{near})-Z_{near}-Z_{far}} \tag{7}$$

The Q matrix equations (Eqns. (2) and (3)) can also be easily inverted to provide disparity from depth.

$$\delta = \frac{T_x f}{Z_{world}} - (C_x^1 - C_x^2) \tag{8}$$

This gives a disparity for every pixel, which is the x displacement between the corresponding point in the left and right images.

A depth map is produced for both the left and right images. Occlusions can be calculated by looking at the depth for a pixel in the left image and its corresponding, based on disparity, pixel in the right image. In a rectified setup, the Z coordinate of a point should be the same in the left and right camera coordinates. Any points that do not have the same Z value within an error margin are considered occluded regions visible in the left image but not in the right image. We calculate occlusions for both the left and right images for completeness. A depth equivalent to the back clipping plane corresponds to a point not visible in the CT scan. The resulting images can be seen in Fig. 11 with left occlusions in red and
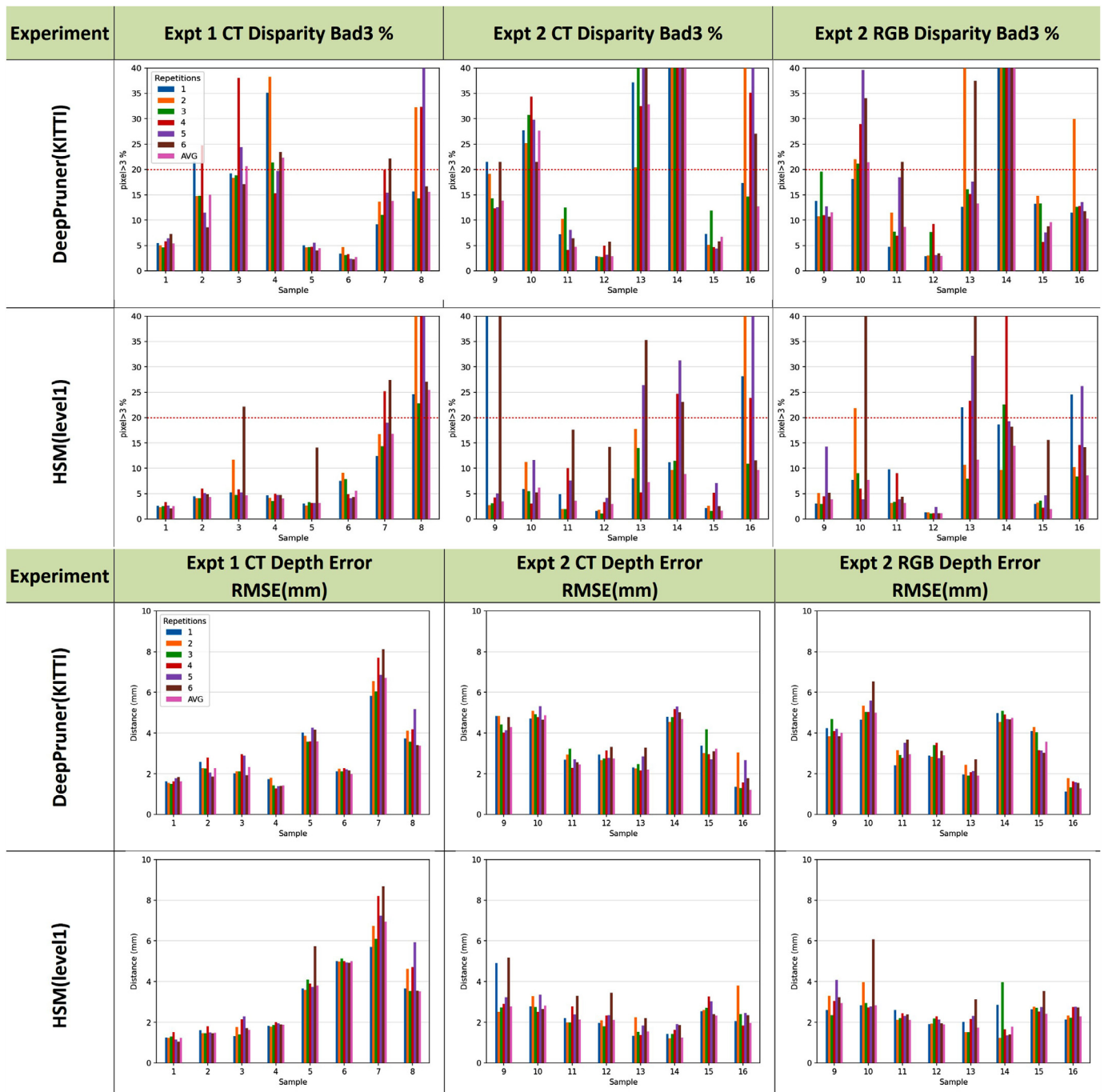
Fig. 9. Depth error and bad3 metric (percentage of disparities >3 pixels error) for multiple manual alignments. Results with a bad3 of >20% for **both** HSM (level 1) and DeepPruner are considered as outliers due to human error. Outliers were not taken into consideration when computing the average manual alignment (in pink). This average produces a consistently low bad3 error for the best performing network and the depth error is around 2 mm for all but the most challenging images.

right occlusions in green. The useful pixels for stereo can be identified and cover the majority of the image as there are only small regions of occlusion.

2.7. Distribution format

Since there are a number of different packages, formats and coordinate systems for stereo calibration, we distribute the dataset in a much simplified form. We provide only the rectified left and right images with corresponding left and right depth images and disparity from left to right. In addition, we provide a combined mask image for both cameras that shows regions of non-overlap,

areas not covered by the 3D model and also identifies occluded pixels which can be seen in one view but not the other. Stereo rectification parameters are provided in a single JSON file containing the P_1, P_2 and Q matrices for each image. The parts of the distributed form are boxed in red in the schematic flowchart (Fig. 1). This format can be directly used by both stereo correspondence algorithms and monocular, single frame depth estimation approaches. We feel that this format significantly simplifies the process and will enable rapid use of the dataset. The original images, full calibration parameters, segmented CT scans, surfaces and scripts for registration are also made available in a separate archive.
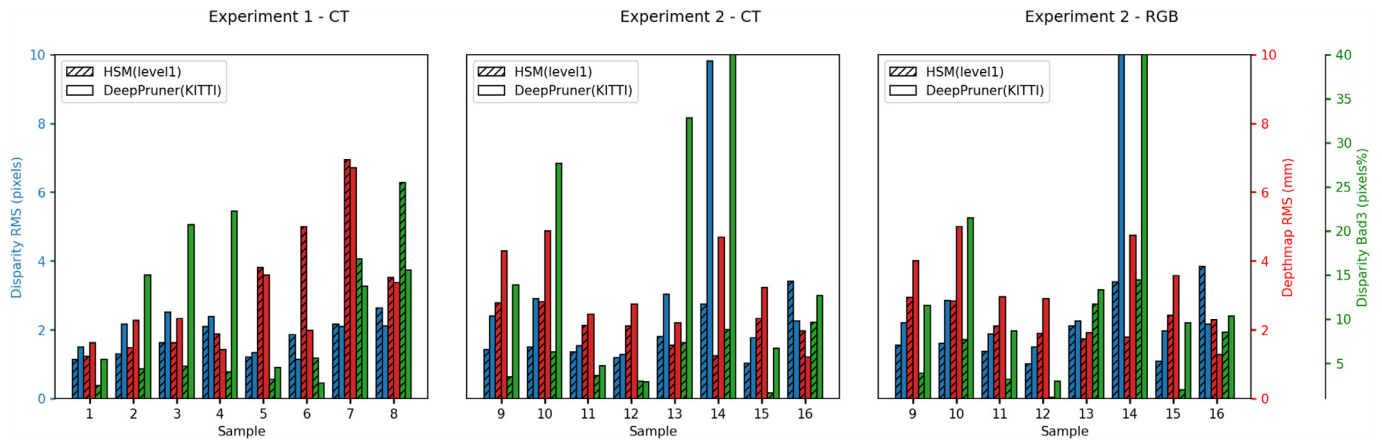
Fig. 10. Per frame error metrics of the two best performing networks, evaluated on the dataset created using the average manual registration of each image. Some very high errors can be attributed to failure of the networks on these difficult images, but at least one network succeeds for each sample. The Disparity has a consistent RMS error of around 2 pixels or less compared to the best performing network. Depth error is also around 2 mm in most cases. Samples 5 and 7 have higher depth error compared to disparity due to greater absolute depth (see Fig. 3). The difficult smooth specular images (14 and 16) have slightly higher error, which is likely to be, at least in part, due to errors from the networks rather than the reference.

3. Evaluation study

We use SERV-CT to evaluate the performance of different stereo algorithms. To evaluate disparity outputs we use the bad3 % error and root mean square disparity error (RMSE) since these are the popular metrics in stereo evaluation platforms (Scharstein et al., 2014). In addition using the Q matrix, we convert estimated disparities to depth and measure the depth RMSE error in mm.

3.1. Investigated stereo algorithms

We primarily focus on the performance of fast learning based stereo algorithms as these perform best in different challenges. The models were selected based on their accuracy and inference time, as reported on KITTI-15 leader-board. We chose methods that are publicly available and can provide inference at more than 10 frames per second. We also included PSMNET despite its slower performance since this was the winning technique of a grand challenge in endoscopic video (Allan et al., 2021). GitHub repositories containing both implemented networks and pre-trained models used in this study, are provided as footnotes. Inference runtime performance and GPU memory consumption for each network on a single NVIDIA Tesla V100 can be found in Table 1. We measure inference time excluding loading time from a storage device to CPU RAM but including prepossessing.

Experimentally we found that KITTI-15 fine-tuned methods were able to produce satisfactory results in our evaluation set. To make our result reproducible we chose to only use networks whose pretrained weights in a large stereo dataset are available online. This will also allow us to showcase how these networks adapt to a completely different domain from the their training

sample and how they compare with the best methods for surgical stereo. All the networks we are testing are regression end-to-end architectures, meaning that they allow gradients to flow freely from the output to their inputs and also produce subpixel disparities. We test some of the fastest deep stereo architectures as inference time performance is an important aspect of any surgical reconstruction system. We also include in our comparison a traditional vision method devised specifically for endoscopic surgical video, the quasi-dense stereo (Stereo-UCL) algorithm by Stoyanov et al. (2010), which was used in the the TMI study by Maier-Hein et al. (2014) and its implementation is available online and as part of OpenCV contrib since version 4.1.

3.1.1. DispNetC

Along with the introduction of large synthetic stereo dataset for training deep neural networks, Mayer et al. (2016) introduced the first end-to-end stereo network architectures, which achieved similar results the best contemporary methods while being orders of magnitude faster. We will focus on DispNetC[8] which consists of a feature extraction module, a correlation layer and a encoder decoder part with skip connections which aims to refine the cost volume and compute disparities. The feature extraction part of this network downsamples and extracts unary features for each image separately. Those features get correlated together in the horizontal dimension, building a cost volume, which in turn gets further refined. This last refinement and disparity computation sub-module, follows an 2D encoder-decoder architecture, with skip connections, which is in place to allow matching for large disparities and provides subpixel accuracy.

3.1.2. PSMNet

Influenced by work in semantic segmentation literature, Chang and Chen (2018) introduced PSMNet[9], achieving then leading performance on the KITTI leaderboard. Although it cannot be considered real time, its novelty lies on the incorporation of spatial pyramidal pooling (SPP) module. This architecture enables such networks to extract unary features that take into account global context information, something that is crucial in surgical stereo applications due to homogeneous surface texture or the presence

Table 1
Neural network average computational memory and speed when producing disparity maps on the SERV-CT frames.

Model	GPU Memory (MB)	Frames/sec.
PSMNet	4345	3
DeepPruner	1311	16
HSM(level1)	1273	20
HSM(level2)	1209	28
HSM(level3)	1187	32
HAPNet	1139	20
DispnetC	2311	20
MADNet	**1043**	**35**

---

[8] https://github.com/CVLAB-Unibo/Real-time-self-adaptive-deep-stereo (commit b20eb32).

[9] https://github.com/JiaRenChang/PSMNet (commit a5c2704).
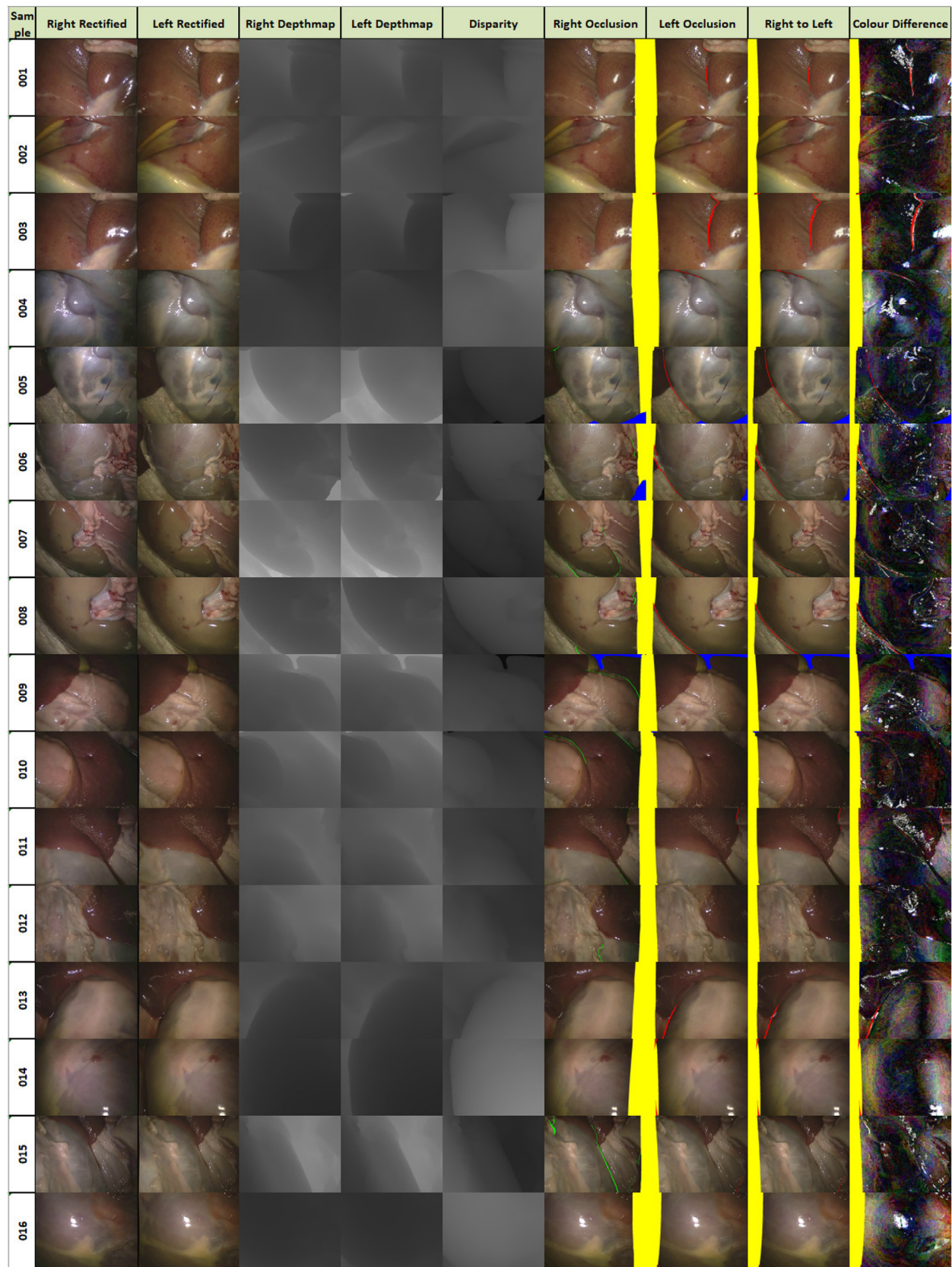
Fig. 11. The images that comprise the SERV-CT testing set. Depth maps are constructed using the OpenGL Z-buffer. Disparity comes from the depth and the stereo rectified Q matrix. Occluded regions are those that have different depths for left and right once correspondence is established using the disparity (non overlap is shown in yellow, occlusion is in red and green, for right and left images respectively, and areas outside the surface model are in blue). Resampled right-to-left and amplified colour difference images are also shown.

of specular highlights. The SPP module achieves this by extracting features at different scales and later concatenates them before forming a 4D feature volume. This 4D feature volume gets refined by a stacked hourglass architecture which further improves disparity estimation.

3.1.3. HSM

Yang et al. (2019), in an effort to develop a network that can infer depth fast in close range, to be used in autonomous vehicles platforms, developed the hierarchical deep stereo matching (HSM) network[10]. This allows fast and memory efficient disparity computation enabling it to process high resolution images. During the feature extraction process, the down-sampled feature images form feature volumes, each, corresponding to a different depth scale. Feature volumes of coarser scales get refined, upsampled and concatenated to the one of the next finer scale, hierarchically refining the disparity estimation. This enables the network to make fast queries from intermediate scales in the expense of depth accuracy.

3.1.4. DeepPruner

DeepPruner[11] is another fast and memory efficient deep stereo architecture from Duggal et al. (2019). Based on the PatchMatch algorithm from Barnes et al. (2009), they created a fully differentiable version of it making ideal to incorporate it in a end-to-end neural network. The PatchMatch module prunes the disparity search space, enabling the network to search in a smaller disparity range, which in turn reduces the memory consumption, as well as the time to build and process this feature volume. The feature volume gets processed by a refinement network to increase matching performance. The differentiable PatchMatch module, though it does not contain any learnable parameters, enables gradient flow, facilitating end-to-end training. In our test we use the fast configuration of this method as described in the original paper.

3.1.5. MADNet

To tackle the domain shift problem most deep learning architecture experience, Tonioni et al. (2019) introduced a modularly adaptive network, MADNet[12], which can be used with the modular adaptation (MAD) algorithm enabling the network to adapt in a different target domain from the one that it's trained on. The architecture is one of the fastest in the literature and the online modular adaptation scheme is efficient enough to run in real time. MADNet is based on a hierarchical pyramid and cost volume architecture, which enables it to employ the adaptation scheme at inference time, without reducing real time performance significantly. In our experiments we do not use the MAD because the number of available samples are limited.

3.1.6. HAPNet

3D convolutions and manual feature alignment are the two least efficient operations deep stereo neural networks perform. Working towards a real time stereo matching network, Brandao et al. (2020) introduced HAPNet[13], an architecture that extract features in different scales, concatenates them and find correspondences using a 2D hourglass encoder decoder block. The disparity estimation process is done in a hierarchical fashion, where low resolution feature maps are used to find low resolution disparities. Those low resolution disparities get up-convolved and concatenated with the features of the next scale to hierarchically re-

fine and regress the final disparity. The 2D hourglass encodes correspondences from the concatenated features, effectively enlarging the receptive field of the network.

3.1.7. Stereo-UCL

To estimate depth in surgical environments that specular reflections and homogeneous texture make most stereo algorithms fail, Stoyanov et al. (2010) developed an algorithms we will refer to as Stereo-UCL[14] to be compatible with the TMI evaluation study from Maier-Hein et al. (2014). The algorithms produces semi-dense disparities based on a best first region growing scheme. It is the only algorithm out of the methods of this study that can robustly estimate matches between unrectified surgical stereo pairs. In an initial step, the method finds sparse features in the left frame and then it uses optical flow to match with pixels in the right view. Those pixels are used as inputs seeds in a region growing algorithm that propagates disparity information from known disparities to adjacent pixels.

3.2. Evaluation details

For each deep learning model and method we run inference across all samples, using models trained on mainstream computer vision datasets (Menze and Geiger, 2015; Mayer et al., 2016). All deep learning based methods are configured to search for matches up to at least 192 pixels, and results are stored as 16-bit PNG images, normalized appropriately to encode subpixel information. We use the depthmap supplied as part of SERV-CT as previously described and compare this to a triangulated depthmap from the network output. For disparity evaluation, since we have disparities from both SERV-CT and from the stereo algorithms, we can directly measure the error without further processing. Additionally, using the occlusion images, we provide separate results for occluded and non-occluded pixels. We separate the results for the first *ex vivo* sample (Expt. 1 - CT) and the second sample where we have both the CT surface (Expt. 2 - CT) and the RGB surface from the Creaform scanner (Expt. 2 - RGB).

4. Results

Results are split into three groups - *ex vivo* sample 1 with CT reference (Expt. 1 – CT), and *ex vivo* sample 2 with reference from either the CT scan (Expt. 2 - CT) or Creaform RGB surface registered to CT (Expt. 2 – RGB).

Numerical results are summarised in Tables 2, 3 and 4. There is a general trend that learning based methods trained only on synthetic data produce disproportionately high error when compared to versions trained on real data. Networks fine-tuned on real data perform slightly better than the domain specific classical stereo method (Stereo-UCL). HSM and DeepPruner consistently performed the best across all dataset cases. We do not consider networks trained on synthetic data in subsequent analysis.

The greater consistency of the 3D error across our experiments is an encouraging indicator of the reliability of the reference, since 3D error metrics give an idea of the performance in real world distances. The bad3 metric is harder to optimise for as it is highly dependent on the distribution of errors, but gives an indication of the proportion of the image where there are errors. Disparity metrics enable comparison of matching performance on data from potentially very different camera setups.

Figs. 12, 13 and 14 provide error images showing the difference between the reference and the result from each algorithm for every pixel. Error metrics are lower for Expt. 1. Some more challenging images are presented in Expt. 2, particularly samples 013,

---

[10] https://github.com/gengshan-y/high-res-stereo (commit 669e5d5).

[11] https://github.com/uber-research/DeepPruner (commit 7cfd5e6).

[12] https://github.com/CVLAB-Unibo/Real-time-self-adaptive-deep-stereo (commit b20eb32).

[13] https://github.com/patrickrbrandao/HAPNet-Hierarchically-aggregated-pyramid-network-for-real-time-stereo-matching (commit cc39708).

[14] Available in OpenCV version 4.1 and above.

Fig. 12. Signed disparity error in pixels of each algorithm compared to the CT reference for Expt. 1 (hotcold colormap from endolith).
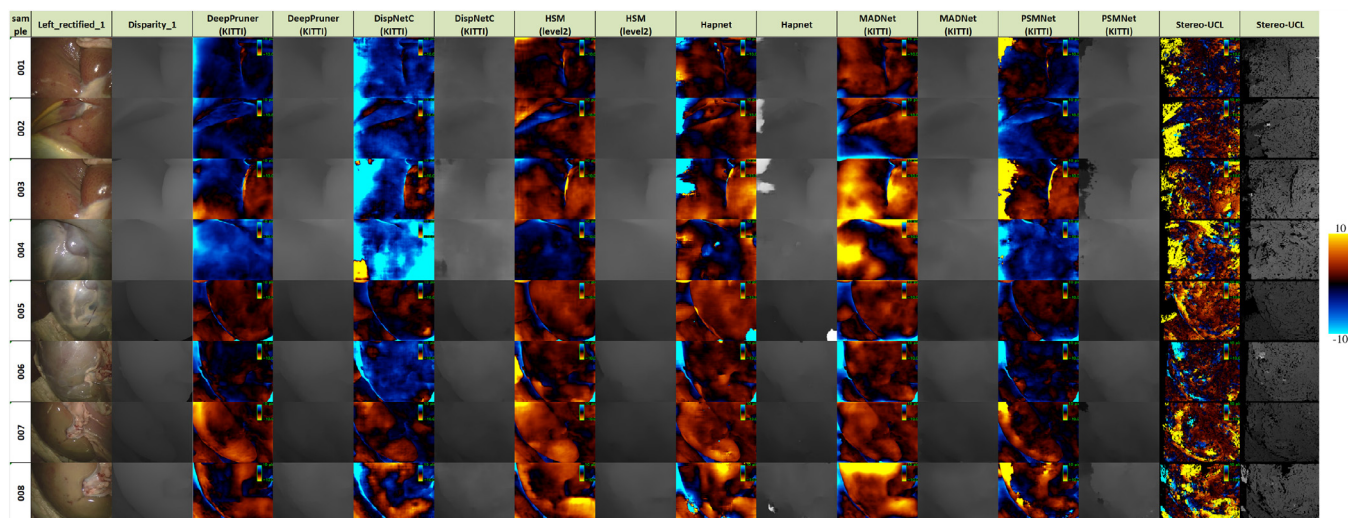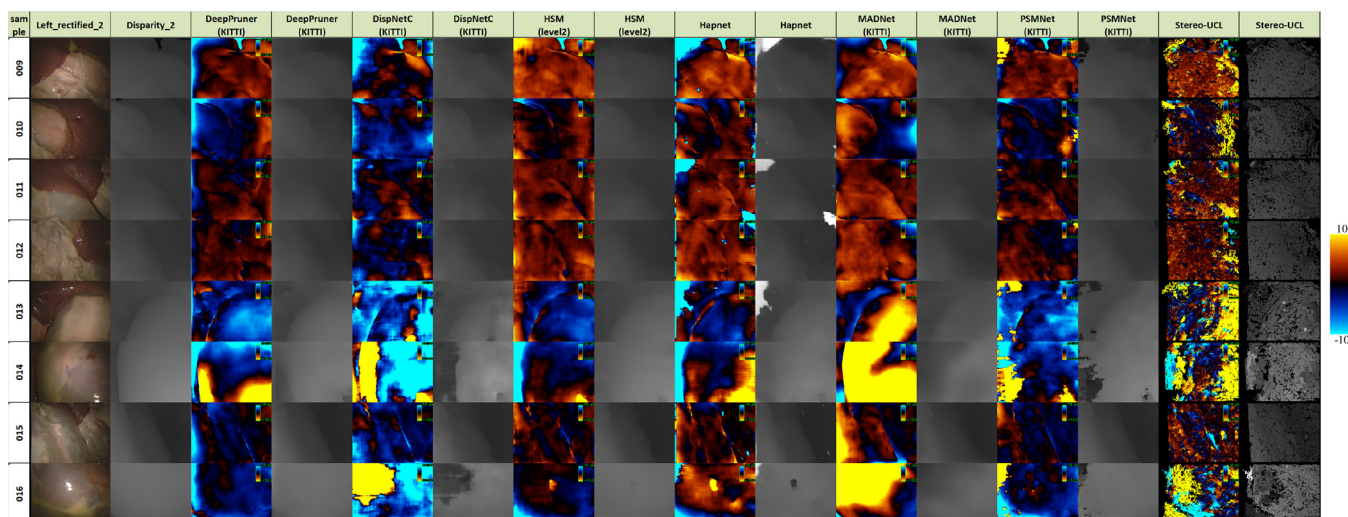

Fig. 13. Signed disparity error in pixels of each algorithm compared to the CT reference for Expt. 2 (hotcold colormap from endolith).
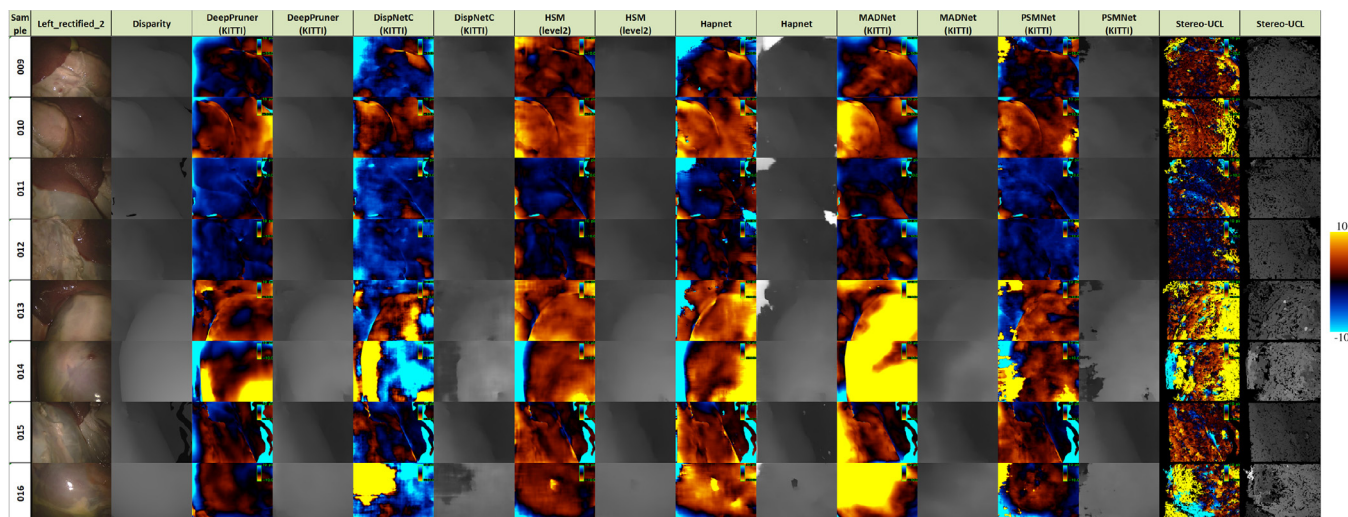

Fig. 14. Signed disparity error in pixels of each algorithm compared to the Creaform RGB reference for Expt. 2 (hotcold colormap from endolith).

Table 2
Expt. 1 - CT results.

	Mean bad3 Error %		Mean RMSE 3D Distance (mm)		Mean RMSE Disparity (pixels)	
Occlusions: Method	not included	included	not included	included	not included	included
DeepPruner(KITTI)	12.50 (±7.44)	17.18 (±8.18)	**2.91** (±1.71)	**3.77** (±1.65)	1.91 (±0.51)	2.47 (±0.60)
DeepPruner(SceneFlow)	53.63 (±20.39)	56.27 (±19.70)	13.01 (±8.60)	17.09 (±8.79)	18.29 (±14.44)	24.35 (±15.69)
DispNetC(KITTI)	40.09 (±26.41)	42.79 (±27.32)	4.58 (±0.76)	5.66 (±0.95)	4.24 (±2.70)	5.20 (±3.29)
DispNetC(SceneFlow)	26.78 (±17.77)	31.00 (±19.13)	4.66 (±1.52)	6.22 (±2.14)	4.23 (±2.59)	5.75 (±4.20)
HSM(level1)	**8.34** (±8.31)	**10.84** (±8.93)	3.18 (±2.03)	4.43 (±2.91)	**1.75** (±0.53)	**2.19** (±0.69)
HSM(level2)	12.28 (±10.20)	14.76 (±10.21)	3.53 (±2.16)	4.64 (±2.65)	2.13 (±0.69)	2.50 (±0.74)
HSM(level3)	63.90 (±12.27)	63.00 (±11.93)	10.09 (±7.07)	10.62 (±7.16)	5.42 (±1.29)	5.50 (±1.37)
Hapnet	17.85 (±13.31)	21.35 (±13.11)	6.01 (±4.07)	8.27 (±3.94)	7.41 (±7.40)	12.47 (±11.38)
MADNet(KITTI)	26.58 (±18.11)	30.09 (±18.28)	4.23 (±1.42)	5.05 (±1.50)	3.44 (±1.64)	3.83 (±1.62)
MADNet(SceneFlow)	34.01 (±16.31)	39.52 (±15.49)	13.22 (±13.31)	16.48 (±13.69)	15.75 (±14.89)	19.78 (±14.95)
PSMNet(KITTI)	12.16 (±7.12)	17.37 (±8.15)	9.53 (±9.40)	18.15 (±21.23)	3.48 (±2.21)	5.59 (±4.32)
PSMNet(SceneFlow)	98.31 (±1.23)	98.11 (±1.11)	19.63 (±6.33)	22.52 (±7.13)	24.60 (±6.22)	31.61 (±9.23)
Stereo-UCL	33.24 (±10.09)	34.43 (±10.18)	26.40 (±19.55)	36.46 (±32.03)	9.24 (±2.96)	10.89 (±4.29)

Table 3
Expt. 2 - CT results.

	Mean bad3 Error %		Mean RMSE 3D Distance (mm)		Mean RMSE Disparity (pixels)	
Occlusions: Method	not included	included	not included	included	not included	included
DeepPruner(KITTI)	19.13 (±16.95)	24.58 (±15.75)	3.21 (±1.31)	3.82 (±1.65)	3.12 (±2.77)	4.03 (±2.71)
DeepPruner(SceneFlow)	87.73 (±15.79)	88.99 (±14.13)	29.77 (±11.76)	33.14 (±12.52)	55.27 (±16.00)	68.05 (±13.15)
DispNetC(KITTI)	47.87 (±27.58)	50.72 (±25.40)	7.07 (±4.56)	7.49 (±4.70)	7.53 (±7.33)	8.12 (±7.20)
DispNetC(SceneFlow)	47.60 (±32.15)	50.25 (±31.02)	7.68 (±3.68)	8.77 (±4.67)	15.21 (±17.76)	17.09 (±20.14)
HSM(level1)	**5.46** (±2.96)	**9.22** (±5.20)	**2.12** (±0.54)	**2.98** (±1.29)	**1.81** (±0.83)	**2.73** (±2.07)
HSM(level2)	9.73 (±3.96)	13.68 (±5.72)	2.78 (±1.17)	3.74 (±1.24)	2.07 (±0.63)	3.08 (±1.92)
HSM(level3)	46.95 (±11.61)	49.77 (±9.73)	6.47 (±3.68)	7.43 (±3.09)	4.57 (±1.00)	5.68 (±2.35)
Hapnet	23.64 (±17.61)	27.71 (±16.56)	8.45 (±4.99)	10.33 (±5.24)	13.09 (±8.88)	18.82 (±12.96)
MADNet(KITTI)	38.24 (±30.26)	40.39 (±27.63)	6.31 (±3.13)	6.72 (±3.89)	7.21 (±7.02)	7.68 (±7.25)
MADNet(SceneFlow)	69.49 (±19.68)	71.61 (±18.34)	30.17 (±15.19)	31.24 (±12.45)	46.47 (±21.83)	50.59 (±18.03)
PSMNet(KITTI)	15.41 (±13.02)	19.58 (±15.52)	9.19 (±7.67)	12.94 (±12.58)	5.09 (±4.99)	6.65 (±5.85)
PSMNet(SceneFlow)	96.18 (±1.58)	96.38 (±1.30)	23.78 (±9.26)	25.17 (±8.76)	35.85 (±17.75)	40.98 (±18.94)
Stereo-UCL	42.74 (±22.25)	43.05 (±22.44)	35.19 (±37.62)	35.63 (±37.49)	12.45 (±8.01)	13.10 (±8.82)

Table 4
Expt. 2 - Creaform results.

	Mean bad3 Error %		Mean RMSE 3D Distance(mm)		Mean RMSE Disparity(pixels)	
Occlusions: Method	not included	included	not included	included	not included	included
DeepPruner(KITTI)	15.31 (±12.91)	21.62 (±12.51)	3.30 (±1.30)	3.91 (±1.59)	3.11 (±2.84)	3.99 (±2.76)
DeepPruner(SceneFlow)	87.06 (±17.17)	88.44 (±15.26)	29.36 (±11.75)	32.86 (±12.55)	54.62 (±16.35)	67.77 (±13.13)
DispNetC(KITTI)	48.24 (±24.92)	51.05 (±22.92)	7.19 (±4.43)	7.60 (±4.60)	7.46 (±7.15)	8.08 (±7.13)
DispNetC(SceneFlow)	47.99 (±30.72)	50.48 (±29.93)	7.71 (±3.46)	8.86 (±4.60)	15.01 (±17.34)	16.92 (±19.80)
HSM(level1)	**6.58** (±4.82)	**10.40** (±7.05)	**2.25** (±0.45)	**3.17** (±1.54)	**2.00** (±1.06)	**2.96** (±2.38)
HSM(level2)	10.98 (±5.99)	15.12 (±8.13)	2.96 (±1.20)	3.95 (±1.52)	2.29 (±0.91)	3.32 (±2.23)
HSM(level3)	49.78 (±13.02)	52.30 (±10.49)	6.48 (±3.53)	7.51 (±3.07)	4.84 (±1.19)	5.98 (±2.62)
Hapnet	27.07 (±20.93)	31.15 (±20.08)	8.72 (±5.27)	10.60 (±5.46)	13.46 (±9.14)	19.11 (±13.08)
MADNet(KITTI)	40.28 (±32.16)	42.27 (±29.22)	6.49 (±3.25)	6.92 (±4.00)	7.55 (±7.28)	8.06 (±7.53)
MADNet(SceneFlow)	68.25 (±19.47)	70.58 (±18.16)	30.17 (±14.94)	31.26 (±12.20)	46.38 (±22.09)	50.52 (±18.37)
PSMNet(KITTI)	14.05 (±8.03)	18.47 (±10.70)	9.34 (±7.48)	13.08 (±12.34)	5.16 (±4.85)	6.76 (±5.71)
PSMNet(SceneFlow)	96.26 (±1.22)	96.47 (±0.96)	23.85 (±9.67)	25.28 (±9.16)	36.11 (±17.95)	41.21 (±19.10)
Stereo-UCL	43.72 (±23.50)	44.02 (±23.64)	35.03 (±37.87)	35.48 (±37.72)	12.68 (±8.35)	13.34 (±9.14)

014 and 016. These depict smooth surfaces that are either feature-less or contain significant specular highlights and clearly present a problem for some of the networks. The difficulty of these images may account for much of the increase in error.

The bad3 % error results for our dataset are higher than those generally reported in the computer vision literature. This is mainly because the tested methods are not fine-tuned for this specific dataset. The limited sample size does not facilitate training on this dataset. The challenging images presented to the algorithms and also any inaccuracies in the CT or Creaform reference surfaces will contribute to this error. It is hard to separate algorithm fitting error from reference error, but the spread of performance from all methods suggests that inaccuracies from the reconstruction algorithms may dominate.

The analysis has slightly higher error for the Creaform surfaces for Expt. 2, which may result from the extra registration process

from the RGB surface to the CT scan. The only significant improvement is with frame 16, where there is less variation and a lower mean error when aligning with the RGB surface. However, the ability to register using visible surface features that have no corresponding geometric variation is potentially useful and may offer a route to automated alignment.

Overall, we can see that algorithmic performance can be compared using our dataset and that the images included, present different challenges to what is offered by existing datasets. Although the dataset has limitations, we believe the value of providing direct disparity maps for evaluation will significantly ease the process of evaluating new algorithms and support the community by establishing benchmarks.

5. Discussion

Standardised datasets have been instrumental in accelerating the development of algorithms for a wide range of problems in computer vision and medical image computing. They not only alleviate the need for data generation but also provide a means of transparently and measurably benchmarking progress. Despite recent progress and emergence of datasets in endoscopy, for example for gastroenterology (the GIANA challenge[15] and the KVASIR dataset from Pogorelov et al. (2017)), instrument and scene segmentation (Allan et al., 2019; Maier-Hein et al., 2020; Al Hajj et al., 2019), and for 3D reconstruction (Maier-Hein et al., 2014; Rau et al., 2019; Penza et al., 2018; Allan et al., 2021), there is still a need for improved datasets in surgical endoscopy, which presents specific challenges due to the lack of clearly identifiable features, highly deformable tissue and the presence specular highlights, smoke, tools and blood.

5.1. Comparison with the SCARED Challenge Dataset

The most closely related work to that described in this paper is the SCARED challenge dataset (Allan et al., 2021). This dataset is an excellent achievement that provides keyframes with for a range of anatomy that also incorporates tools in some of the frames. It is the best available endoscopic stereo reconstruction benchmark to date.

However, there are some issues arising with the data. We list some of the ways the method described here compares favourably to SCARED.

1. SCARED provides reference in form of 3D images where each pixel encodes the 3D location of the point that gets projected to it. Although creating depthmaps out of this format is trivial, some processing is needed to stereo rectify the RGB images and construct the disparity reference, making the dataset compatible with stereo matching algorithms. The output of this process heavily depends on the rectification parameters and implementation. By providing our reference in both depth and disparity domains, we ensure that reported results across different papers using SERV-CT will only reflect the performance of the algorithms and will be invariant of any data processing pipeline.
2. The quality of the endoscope's calibration in SCARED varies across the sub-datasets. We provide a consistent stereo calibration.
3. Identification and reconstruction of occluded regions. The full CT surface provides true reference depth for regions that are visible only from a single camera and the appropriate masks to identify those occluded regions. Some algorithms will provide better monocular reconstruction in these areas and SERV-CT's occlusion masks provide a way to evaluate performance
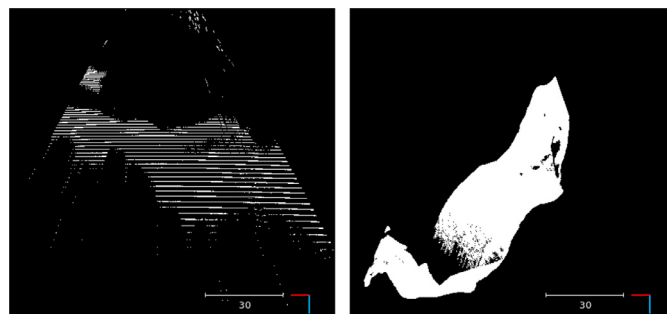
Fig. 15. X-Z view of a SCARED sample (left) and a SERV-CT sample (right). Samples provided from the SCARED challenge, exhibit significant outliers and step artefact. SERV-CT provides smooth smooth ground truth reference captured using a CT scanner.

in these regions. Although depth information for occluded pixels may be provided in the interpolated depth sequences of SCARED, appropriate occlusion masks are absent making it hard to evaluate the performance stereo matching algorithms in the absence of pixel correspondences.

4. The SCARED dataset includes outliers and exhibits a significant step artefact, especially in the periphery. SERV-CT provides a smooth surface without significant holes or artefacts (see Fig. 15).
5. Acquisition time The time taken to produce a scan with the O-arm[TM] is 30s. Though the scan time is not described in the SCARED paper, there are 40 images that need to be projected to grab a frame with the projection system: (10bits x 2(orientation) x2 (negative/positive)). To cover the whole scene or initially out of focus regions requires repositioning and refocussing of the projector. We estimate that the collection time will be several minutes, during which the tissue may move or deform under gravity and introduce ambiguities in the ground truth. Any such motion will be reduced using the faster O-arm[TM].

It is unclear how the staircase artefact in particular could influence the performance of learning based algorithms trained on the SCARED dataset. Due to the same artifact we cannot get a clear picture of marginal performance differences between two algorithms.

During training, SERV-CT can be used in combination with SCARED to create models robust to data shift and potentially one dataset can help mitigate issues introduces by the other. Additionally, rather than using SERV-CT for training, we can leverage the dense and smooth annotation of its samples to measure marginal performance differences between two methods and their robustness to data shift.

5.2. Alternative protocols for dataset reconstruction

The protocol followed in this work involved moving the endoscope under robotic motion to provide multiple views in a manner similar to surgery. An alternative would be to clamp the endoscope and move the phantom. If the endoscope can be held sufficiently rigid with respect to the scanner, then registration from CT to endoscope should be maintained between scans. Multiple scans can be taken and the phantom moved, deformed or cut between scans. For our bulky torso phantom and with the constraints of the O-arm[TM] volume this significant movement was not considered feasible. Small motions and deformations of the scene would be possible, but these will need to be stable over the period of the scan. Any cutting, deformation or resection under this protocol would only be viewed from one perspective. This was not considered practical in our setup, but for other arrangements such a protocol might provide a means of increasing the sample size.

Another option would be to take a scan, then place a calibration object or a set of markers in the scan volume and rescan. Again, the endoscope would have to be fixed with respect to the O-arm™ for the duration of the scans. With a limited number of markers in close proximity for our calibration object this was not sufficiently accurate. But again, with alternative setups such a method could be used. The point registration to the calibration frame and any instability of the endoscope clamping will introduce their own errors, but such a method could be used in principle.

As a further alternative, we initially attempted to use marker bearings attached to the endoscope to provide tracking frame of reference. Due to artifact from the endoscope itself, these could not be located with sufficient accuracy for precise hand-eye calibration.

There are structured light scanning methods that provide sub-pixel reconstruction, avoiding step artefact. These methods may increase the data collection time further, however. One method that may offer promise is to perform a combined approach, with an accurate external surface from one device being aligned with with a similar reconstruction through the endoscope. There are many potential avenues for future reference dataset construction and we hope that this discussion and the account of SERV-CT offer useful guidance to the research community.

5.3. Note on human visual adjustment and verification of stereo reconstruction

Any reference dataset will involve a reconstruction process and perhaps registration too. All such methods will have associated errors. In addition to analysis of error metrics and loss functions, stereo human visual inspection of reference surfaces should be a standard part of the verification process for future datasets. Where there are errors, human visual correction and adjustment can be facilitated under constraints that ensure that aspects well measured by the process are maintained. This was our strategy with SERV-CT. As methods become more accurate, human manual adjustment may no longer be needed, but visual inspection is still advised to provide insight into the errors that remain.

6. Conclusions

With this paper, we have developed and reported a validation dataset based on CT images of the endoscope and the viewed anatomical surface. The location of the endoscope constrains the perspective from which the anatomical surface is viewed. Subsequent rotation of the view is established by manual alignment, followed by a small Z translation to account for the true lens position within the endoscope (see section 2.6). Constraining the endoscope location in this way, ensures that the distance from camera to anatomy comes from the CT dataset. This method of constrained alignment could be used for any 3D modelling system that covers both the endoscope and the viewed anatomy. The dataset and algorithms for processing the raw data and the generated disparity maps are openly available[16]. In addition, we report the performance of various fast computational stereo algorithms that are open source and provide the full parameter settings alongside the trained model weights to allow experiments to be reproduced easily.

SERV-CT adds to existing evaluation sets for stereoscopic surgical endoscopic reconstruction. The validation covers the majority of the image and there is considerable variation of depth in the viewed scenes (see Fig. 3). The analysis of several stereo reconstruction algorithms has been performed and demonstrates the feasibility of SERV-CT as a validation set, but also highlights challenges. The results suggest that learning based stereo methods,

trained on real world scenes, are promising candidates for surgical stereo-endoscopic reconstruction.

There are limitations to the work. A relatively small number of 16 frames with corresponding reference are provided. The manual alignment relies heavily on operator skill and is not a trivial process. Automating this part of the procedure would be desirable but presents an open registration problem in its own right. There is variation of anatomy and considerable variation of depth in the images presented, but further realism could be provided by the introduction of tools, smoke, dissection or resection and blood. The endoscope images from the original da Vinci are comparatively low contrast and resolution compared to newer endoscopes and there is noticeable colour difference between the eyes.

Datasets are not only important for validation but also for training deep learning models. The sample size of our dataset, alone, is not sufficient to train accurate models. A possible way of addressing this is through simulation (Rau et al., 2019; Pfeiffer et al., 2019), but fine tuning may still be required. In our future work plans, we intend to extend this dataset significantly in a variety of ways. Kinematic or other tracking may extend the CT alignment to multiple frames, significantly increasing the number of stereo pairs available and also providing a reference for video-based reconstruction and localisation methods such as SLAM. We intend to gather more such datasets under different conditions and will also investigate the use of other devices for measurement of the tissue surface, such as laser range finders and further structured light techniques.

Despite the recognised limitations of the SERV-CT dataset, we have established the feasibility of this methodology of reference generation. The method could be applied to any measurement system that can provide the location of both the endoscope and the viewed anatomy in the same coordinate system. It may potentially be a way of approaching the bottlenecks in image-guided surgery through preoperative and intraoperative surface registration. We also hope that this work encourages further development of such reference surgical endoscopic datasets to facilitate research in this important area which may help provide surgical guidance and is likely to be of significant benefit in the development of future robotic surgery.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Odin Vision Ltd.

Digital Surgery.

CRediT authorship contribution statement

P.J. Eddie Edwards: Conceptualization, Methodology, Formal analysis, Data curation, Software, Writing – original draft. **Dimitris Psychogyios:** Validation, Methodology, Formal analysis, Data curation, Writing – original draft. **Stefanie Speidel:** Writing – review & editing. **Lena Maier-Hein:** Writing – review & editing. **Danail Stoyanov:** Writing – review & editing, Supervision, Funding acquisition.

Acknowledgments

---

[16] https://www.ucl.ac.uk/interventional-surgical-sciences/serv-ct.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2021.102302.

References

Al Hajj, H., Lamard, M., Conze, P.H., Roychowdhury, S., Hu, X., Maršalkaitė, G., Zisimopoulos, O., Dedmari, M.A., Zhao, F., Prellberg, J., Sahu, M., Galdran, A., Araújo, T., Vo, D.M., Panda, C., Dahiya, N., Kondo, S., Bian, Z., Vahdat, A., Bialopetravičius, J., Flouty, E., Qiu, C., Dill, S., Mukhopadhyay, A., Costa, P., Aresta, G., Ramamurthy, S., Lee, S.W., Campilho, A., Zachow, S., Xia, S., Conjeti, S., Stoyanov, D., Armaitis, J., Heng, P.A., Macready, W.G., Cochener, B., Quellec, G., 2019. Cataracts: Challenge on automatic tool annotation for cataract surgery. Medical Image Analysis 52, 24–41.

Allan, M., Mcleod, J., Wang, C., Rosenthal, J. C., Hu, Z., Gard, N., Eisert, P., Fu, K. X., Zeffiro, T., Xia, W., Zhu, Z., Luo, H., Jia, F., Zhang, X., Li, X., Sharan, L., Kurmann, T., Schmid, S., Sznitman, R., Psychogyios, D., Azizian, M., Stoyanov, D., Maier-Hein, L., Speidel, S., 2021. Stereo correspondence and reconstruction of endoscopic data challenge. arXiv:2101.01133.

Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y. H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., Herrera, L., Li, W., Iglovikov, V., Luo, H., Yang, J., Stoyanov, D., Maier-Hein, L., Speidel, S., Azizian, M., 2019. 2017 robotic instrument segmentation challenge. arXiv:1902.06426.

Bae, G., Budvytis, I., Yeung, C.K., Cipolla, R., 2020. Deep multi-view stereo for dense 3d reconstruction from monocular endoscopic video. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention–MICCAI 2020. Springer International Publishing, Cham, pp. 774–783.

Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B., 2009. Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. 28, 24.

Bernhardt, S., Nicolau, S.A., Soler, L., Doignon, C., 2017. The status of augmented reality in laparoscopic surgery as of 2016. Medical image analysis 37, 66–90.

Bertolo, R., Hung, A., Porpiglia, F., Bove, P., Schleicher, M., Dasgupta, P., 2020. Systematic review of augmented reality in urological interventions: the evidences of an impact on surgical outcomes are yet to come. World journal of urology 38, 2167–2176.

Bourdel, N., Collins, T., Pizarro, D., Bartoli, A., Da Ines, D., Perreira, B., Canis, M., 2017. Augmented reality in gynecologic surgery: evaluation of potential benefits for myomectomy in an experimental uterine model. Surgical endoscopy 31, 456–461.

Brandao, P., Psychogyios, D., Mazomenos, E., Stoyanov, D., Janatka, M., 2020. Hapnet: hierarchically aggregated pyramid network for real-time stereo matching. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 1–6. doi:10.1080/21681163.2020.1835561.

Chang, J.R., Chen, Y.S., 2018. Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5410–5418.

Chang, P.L., Stoyanov, D., Davison, A.J., Edwards, P.E., 2013. Real-time dense stereo reconstruction using convex optimisation with a cost-volume for image-guided robotic surgery. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (Eds.), Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, Berlin, Heidelberg, pp. 42–49.

Duggal, S., Wang, S., Ma, W.C., Hu, R., Urtasun, R., 2019. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4384–4393.

Edwards, P., Chand, M., Birlo, M., Stoyanov, D., 2021. The challenge of augmented reality in surgery. In: Atallah, S. (Ed.), Digital Surgery. Springer, Cham, pp. 121–135. doi:10.1007/978-3-030-49100-0_10.

Ghazi, A., Campbell, T., Melnyk, R., Feng, C., Andrusco, A., Stone, J., Erturk, E., 2017. Validation of a full-immersion simulation platform for percutaneous nephrolithotomy using three-dimensional printing technology. Journal of Endourology 31, 1314–1320.

Grasa, O.G., Bernal, E., Casado, S., Gil, I., Montiel, J., 2013. Visual slam for handheld monocular endoscope. IEEE transactions on medical imaging 33, 135–146.

Hansen, C., Wieferich, J., Ritter, F., Rieder, C., Peitgen, H.O., 2010. Illustrative visualization of 3d planning models for augmented reality in liver surgery. International journal of computer assisted radiology and surgery 5, 133–141.

Hughes-Hallett, A., Mayer, E.K., Marcus, H.J., Cundy, T.P., Pratt, P.J., Darzi, A.W., Vale, J.A., 2014. Augmented reality partial nephrectomy: examining the current status and future perspectives. Urology 83, 266–273.

Lin, B., Sun, Y., Qian, X., Goldgof, D., Gitlin, R., You, Y., 2016. Video-based 3d reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey. The International Journal of Medical Robotics and Computer Assisted Surgery 12, 158–178.

Liu, X., Sinha, A., Unberath, M., Ishii, M., Hager, G.D., Taylor, R.H., Reiter, A., 2018. Self-supervised learning for dense depth estimation in monocular endoscopy. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. Springer, pp. 128–138.

Mahmood, F., Durr, N.J., 2018. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. Medical image analysis 48, 230–243.

Mahmood, N., Collins, T., Hostettler, A., Soler, L., Doignon, C., Montiel, J.M.M., 2018. Live tracking and dense reconstruction for handheld monocular endoscopy. IEEE transactions on medical imaging 38, 79–89.

Maier-Hein, L., Groch, A., Bartoli, A., Bodenstedt, S., Boissonnat, G., Chang, P.L., Clancy, N.T., Elson, D.S., Haase, S., Heim, E., Hornegger, J., Jannin, P., Kenngott, H., Kilgus, T., Müller-Stich, B., Oladokun, D., Röhl, S., dos Santos, T.R., Schlemmer, H., 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, 1913–1930.

Maier-Hein, L., P, M., A, B., El-Hawary, 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, 974–996.

Maier-Hein, L., Wagner, M., Ross, T., Reinke, A., Bodenstedt, S., Full, P. M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T. N., Bruno, P., Kisilenko, A., Müller, B., Davitashvili, T., Capek, M., Tizabi, M., Eisenmann, M., Adler, T. J., Gröhl, J., Schellenberg, M., Seidlitz, S., Lai, T. Y. E., Pekdemir, B., Roethlingshoefer, V., Both, F., Bittel, S., Mengler, M., Mündermann, L., Apitz, M., Speidel, S., Kenngott, H. G., Müller-Stich, B. P., 2020. Heidelberg colorectal data set for surgical data science in the sensor operating room. arXiv:2005.03501.

Markley, F.L., Cheng, Y., Crassidis, J.L., Oshman, Y., 2007. Averaging quaternions. Journal of Guidance, Control, and Dynamics 30, 1193–1197.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T., 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4040–4048.

Menze, M., Geiger, A., 2015. Object scene flow for autonomous vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3061–3070.

Mountney, P., Stoyanov, D., Yang, G.Z., 2010. Three-dimensional tissue deformation recovery and tracking. IEEE Signal Processing Magazine 27, 14–24.

van Oosterom, M.N., van der Poel, H.G., Navab, N., van de Velde, C.J., van Leeuwen, F.W., 2018. Computer-assisted surgery: virtual-and augmented-reality displays for navigation during urological interventions. Current opinion in urology 28, 205–213.

Penza, V., Ciullo, A.S., Moccia, S., Mattos, L.S., De Momi, E., 2018. Endoabs dataset: Endoscopic abdominal stereo image dataset for benchmarking 3d stereo reconstruction algorithms. The International Journal of Medical Robotics and Computer Assisted Surgery 14, e1926.

Pfeiffer, M., Funke, I., Robu, M. R., Bodenstedt, S., Strenger, L., Engelhardt, S., Roß, T., Clarkson, M. J., Gurusamy, K., Davidson, B. R., et al., 2019. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. Springer. International Conference on Medical Image Computing and Computer-Assisted Intervention, 119–127.

Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.T., Lux, M., Schmidt, P.T., et al., 2017. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on Multimedia Systems Conference, pp. 164–169.

Quero, G., Lapergola, A., Soler, L., Shahbaz, M., Hostettler, A., Collins, T., Marescaux, J., Mutter, D., Diana, M., Pessaux, P., 2019. Virtual and augmented reality in oncologic liver surgery. Surgical Oncology Clinics 28, 31–44.

Rau, A., Edwards, P.E., Ahmad, O.F., Riordan, P., Janatka, M., Lovat, L.B., Stoyanov, D., 2019. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. International journal of computer assisted radiology and surgery 14, 1167–1176.

Röhl, S., Bodenstedt, S., Suwelack, S., Kenngott, H., Müller-Stich, B.P., Dillmann, R., Speidel, S., 2012. Dense gpu-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration. Medical Physics 39, 1632–1645.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P., 2014. High-resolution stereo datasets with subpixel-accurate ground truth. Springer. German conference on pattern recognition, 31–42.

Scharstein, D., Szeliski, R., 2003. High-accuracy stereo depth maps using structured light. IEEE. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., I–I,

Schroeder, W.J., Lorensen, B., Martin, K., 2004. The visualization toolkit: an object-oriented approach to 3D graphics. Kitware.

Stoyanov, D., 2012. Surgical vision. Annals of biomedical engineering 40, 332–345.

Stoyanov, D., Scarzanella, M. V., Pratt, P., Yang, G. Z., 2010. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. Springer. International Conference on Medical Image Computing and Computer-Assisted Intervention, 275–282.

Thompson, S., Dowrick, T., Ahmad, M., Xiao, G., Koo, B., Bonmati, E., Kahl, K., Clarkson, M.J., 2020. Scikit-surgery: compact libraries for surgical navigation. International Journal of Computer Assisted Radiology and Surgery 15, 1075–1084.

Thompson, S., Schneider, C., Bosi, M., Gurusamy, K., Ourselin, S., Davidson, B., Hawkes, D., Clarkson, M.J., 2018. In vivo estimation of target registration errors during augmented reality laparoscopic surgery. International journal of computer assisted radiology and surgery 13, 865–874.

Tonioni, A., Tosi, F., Poggi, M., Mattoccia, S., Stefano, L.D., 2019. Real-time self-adaptive deep stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 195–204.

Visentini-Scarzanella, M., Sugiura, T., Kaneko, T., Koto, S., 2017. Deep monocular 3d reconstruction for assisted navigation in bronchoscopy. International journal of computer assisted radiology and surgery 12, 1089–1099.

Yang, G., Manela, J., Happold, M., Ramanan, D., 2019. Hierarchical deep stereo matching on high-resolution images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5515–5524.

Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31, 1116–1128.

Zhou, H., Jagadeesan, J., 2019. Real-time surface deformation recovery from stereo videos. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention–MICCAI 2019. Springer, Cham, pp. 339–347.