Simultaneous Multi-Structure Segmentation and 3D Non-Rigid Pose Estimation in Image-Guided Robotic Surgery

Masoud S. Nosrati, *Student member, IEEE,* Rafeef Abugharbieh, *Senior member, IEEE,* Jean-Marc Peyrat, Julien Abinahed, Osama Al-Alao, Abdulla Al-Ansari, Ghassan Hamarneh, *Senior member, IEEE*

***Abstract*—In image-guided robotic surgery, labeling and segmenting the endoscopic video stream into meaningful parts provides important contextual information that surgeons can exploit to enhance their perception of the surgical scene. This information can provide surgeons with real-time decision-making guidance before initiating critical tasks such as tissue cutting. Segmenting endoscopic video is a very challenging problem due to a variety of complications including significant noise and clutter attributed to bleeding and smoke from cutting, poor color and texture contrast between different tissue types, occluding surgical tools, and limited (surface) visibility of the objects' geometries on the projected camera views. In this paper, we propose a multi-modal approach to segmentation where preoperative 3D computed tomography scans and intraoperative stereo-endoscopic video data are jointly analyzed. The idea is to segment *multiple* poorly visible structures in the *stereo/multi-channel* endoscopic videos by fusing reliable prior knowledge captured from the preoperative 3D scans. More specifically, we estimate and track the pose of the preoperative models in 3D and consider the models' *non-rigid* deformations to match with corresponding visual cues in multi-channel endoscopic video and segment the objects of interest. Further, contrary to most augmented reality frameworks in endoscopic surgery that assume known camera parameters, an assumption that is often violated during surgery due to non-optimal camera calibration and changes in camera focus/zoom, our method embeds these parameters into the optimization hence correcting the calibration parameters within the segmentation process. We evaluate our technique in several scenarios: synthetic data, *ex vivo* lamb kidney datasets, and *in vivo* clinical partial nephrectomy surgery with results demonstrating high accuracy and robustness.**

I. Introduction

Minimally invasive surgery (MIS) is prized for its many advantages over traditional open operations including decreased risk of infection due to minimal incisions and faster recovery times resulting in quicker patient return to normal life [11]. The disadvantages of MIS are mostly associated with the loss of direct 3D view of the surgical scene as well as cumbersome and non-intuitive tool manipulation. Minimally invasive robotic surgery (MIRS), which deploys a sophisticated robotic arm remotely controlled from a surgeons console, has undergone rapid development in recent years. In 2012 about 200,000 operations were conducted worldwide using the da Vinci robots (Intuitive Surgical Inc., Sunnyvale, CA) [30]. By July 2013, more than 2,500 da Vinci systems have been installed worldwide [32], which signals the fast growth in demand for such systems. MIRS systems offer several advantages over traditional laparoscopic surgery including greater surgical precision due to finer scaling of movement, improved dexterity with more degrees of freedom (DoF), stereo vision, and enhanced comfort for the surgeon who works in a sitting position [33]. One of the crucial steps during a typical surgery (including open surgeries, MIS and image guided MIRS), is for the surgeons to distinguish between different tissues in the intraopertative view and carefully localize pathology, e.g. delineate the tumour and identify the resection margins. For example during a partial nephrectomy, tumour delineation is particularly critical to ensure the entire cancer has been removed while sparing as much healthy tissue as possible is highly desired to preserve kidney function. However, identifying tissues in intraoperative images is difficult due to many complications including noise associated with clutter such as bleeding or smoke from cutting, poor endoscopic image color/texture contrast between different structures, occluding surgical tools, and limited 3D visibility of the structures of interest where only surfaces are observable from the camera feed.

Typically, surgeons rely on previously viewed preoperative 3D scans, e.g. computed tomography (CT) or magnetic resonance imaging (MRI), to try to mentally reconstruct locations of various structures during surgery. Advances in intraoperative imaging have introduced some other modalities into the operating room, e.g. ultrasound (US) and X-ray fluoroscopy. However, the pre- and intraoperative modalities remain in separate spatial frames of reference. To fuse both modalities and assist the surgeon, numerous approaches have been proposed for augmenting the intraoperative view with 3D data obtained from preoperative scans, e.g. fusing 3D preoperative volume with 2D intraoperative US [3], [6], [7], [16], intraoperative MR [12], and 2D X-ray [24], [40] (Table I). Pose estimation (position, orientation and scale) is one of the main challenges in 2D slice to 3D volume registration. One way to overcome this challenge is to use external intraoperative markers and track them [15], [16]. However, the feasibility,

Masoud S. Nosrati and Ghassan Hamarneh are with the School of Computing Science at Simon Fraser University, BC, Canada (e-mail: {smn6, hamarneh}@sfu.ca)

Rafeef Abugharbieh is with the Department of Electrical and Computer Engineering, University of British Columbia, BC, Canada (e-mail: rafeef@ece.ubc.ca)

Jean-Marc Peyrat and Julien Abinahed are with Qatar Robotic Surgery Centre, Doha, Qatar (e-mail: {jmpeyrat, jabinahed}@qstp.org.qa)

Osama Al-Alao and Abdulla Al-Ansari are with Urology Department, Hamad General Hospital, Doha, Qatar (e-mail: {oalalao, aalansari1}@hmc.org.qa)

TABLE I: Categorization and comparison of certain features (handling multiple 2D views, multiple objects segmentation, tissues occlusion, non-rigid tissues deformation, manual vs. automatic pose estimation, and camera parameter correction) between state-of-the-art methods and our proposed method. The symbol "-" indicates that the corresponding methods align the whole 2D image to a 3D volume and do not deal with object segmentation and/or camera information. MR: magnetic resonance; US: ultrasound; FL: fluoroscopy; NMP: Non-medical photos, EN: endoscopy.

Methods	Type of 2D	multi. 2D views	Multi. objects seg.	Occlusion	Non-rigid deformation	Auto. pose est.	Cam. correction
Osechinskiy et al. [23]	MR	✗	-	✗	✓	✓	-
Gill et al. [12]	MR	✗	-	✗	✗	✓	-
Dalvi et al. [3]	US	✗	-	✗	✗	✗	-
Zikic et al. [40]	FL	✓	-	✗	✗	✓	-
Pickering et al. [24]	FL	✗	-	✓	✗	✓	-
Estepar et al. [6]	US	✗	✓	✗	✗	✗	-
Hernes et al. [15]	US	✗	✓	✗	✓	✓	-
Yim et al. [39]	EN	✗	✓	✗	✗	✓	✗
Su et al. [31]	EN	✗	✓	✓	✗	✗	✗
Merritt et al. [18]	EN	✗	✗	✓	✗	✓	✗
Prisacariu et al. [27]	NMP	✗	✗	✓	✓	✓	✗
Sandhu et al. [29]	NMP	✗	✗	✓	✓	✓	✗
Dambreville et al. [4]	NMP	✗	✗	✓	✓	✓	✗
Prisacariu et al. [26]	NMP	✓	✓	✓	✗	✓	✗
Nosrati et al. [22]	EN	✓	✓	✓	✓	✗	✗
Proposed method	EN	✓	✓	✓	✓	✓	✓

quality, or information content of intraoperative X-ray and US still markedly lags behind the typically high resolution 3D preoperative data, and endoscopic imaging remains the staple modality in MIS. The current approach used in the operating room of mentally reconstructing locations of various structures during surgery by transferring the mental abstraction from 3D to 2D data, is an error-prone procedure especially if the surgeon's level of experience is limited. Many efforts have been made towards addressing this issue and augmenting the endoscopic views. These method vary from directly segmenting the endoscopic scene (e.g. using level sets) to registering the preoperative data onto the intraoperative endoscopic scene as discussed in the following section.

A. Related Works

Many minimally invasive operations benefit from endoscopic cameras for visual examination of the interior of a body cavity or hollow organs. Endoscopic cameras are equipped by a visible light source that enables surgeons to see the internal organs via light reflection. As such, many efforts have been made towards augmenting endoscopic intraoperative scenes by segmenting them into meaningful parts. Some recent works proposed active contour-based methods for endoscopic video segmentation [8], [9] while other methods focused on parameter-sensitive morphological operations and thresholding techniques [5], [19]. However, such approaches rely on color/intensity image information and thus often fail due to noise and clutter from bleeding and smoke. In addition, these methods focus on segmenting one object only in an endoscopic scene.

Other works have incorporated preoperative data such as CT. Some excellent reviews on augmented reality in laparoscopic surgery and image-guided interventions have been provided by [2], [20], [21]. Among methods that utilized preoperative information, some works focused on manually registering 3D preoperative data on 3D surfaces reconstructed from stereo endoscopic video [25], [31] or on reconstructed transparent 3D intraoperative cone-beam image [34]. Other works focused on feature tracking in which corresponding points on endoscopic video and preoperative data are assumed to be known [28]. While the registration in [25], [28], [34] is performed manually, the methods proposed in Yim et al. [39] and Merritt et al. [18] are able to automatically find the 3D pose of the objects and rigidly register 3D CT data to a 2D endoscopy image. None of the aforementioned methods can handle free-form deformation of tissues that usually happens due to respiratory motion and/or surgical intervention.

Recently, we proposed an efficient segmentation and 3D pose tracking technique and presented a closed-form solution for our formulation [22]. We also demonstrated how our framework allows for the inclusion of laparoscopic camera motion model to stabilize the segmentation/pose tracking in the presence of a large occlusion. However, our method in [22] requires manual pose estimation and 3D to 2D alignment for the first frame of the video.

In the (non-medical) computer vision area, Prisacariu et al. [26] proposed a variational method for object segmentation by optimizing a Chan-Vese energy functional with respect to six pose parameters of the object model in 3D. However, in MIS six degrees of freedom are not enough as tissues typically deform non-rigidly. Unlike [26], Sandhu et al. [29] derived a gradient flow for the task of non-rigid pose estimation for a *single* object and used kernel PCA to capture the variance in the shape space. Similar to [29], the method proposed later by Prisacariu et al. [27] allowed for non-rigid pose estimation in a *single* 2D view. In [27], the authors captured 3D shape variance by learning non-linear probabilistic low dimensional latent spaces using the Gaussian process latent variable dimensionality reduction technique. All three aforementioned works ([26], [27], [29]) assumed that the camera parameters are known, which is not always the case in robotic surgery as surgeons often change the zoom and focus. In addition, applying these methods to robotic surgery problems is not straightforward as images in endoscopic videos are highly cluttered and noisy.

B. Contributions

Inspired by [26], we propose a unified level set-based framework that allows for segmenting *multiple structures* in *multi-view* endoscopic video that integrates prior knowledge from preoperative data. Our framework estimates the pose of the preoperative models in 3D as well as considers their *non-rigid* deformation to match them with their corresponding visual cues in multi-channel endoscopic video and segment the objects of interest. Furthermore, our method *corrects the camera calibration parameters* (when needed during the surgery) to compensate for the error in calibration parameters caused by changes in zoom and/or focus. Deploying a region-based formulation (in contrast to local feature-based methods), prior information from the preoperative data, and a random decision forest used to estimate the probability of different organs in each frame of the endoscopic video, enables our framework to handle occlusions (e.g. by surgical tools) as well as noise and clutter in the endoscopic environment. In contrast to our previous work [22] where manual 3D to 2D aligning was necessary for the first frame of the video, in this work we estimate the 3D preoperative pose *automatically* through an optimization framework. In fact, this work can be seen as the (complementary) first step for [22]. In addition, [22] is unable to handle changes in focus/zoom, while in this work, we incorporate the camera parameters into our optimization framework to handle the focus/zoom changes during the surgery.

Our surgical application of interest is robotic partial nephrectomy where the goal is to remove a cancerous tumour while preserving as much healthy kidney tissue as possible. Our proposed fusion of preoperative data with the intraoperative stereo endoscopic view can thus guide the surgeon in localizing the tumour and identifying the resection margins. Table I compares certain features of our work with state-of-the-art methods.

The remainder of this paper is organized as follows: in Section II, we present the details of our optimization-based algorithm for multi-structure segmentation in multiple endoscopic views and introduce our objective function and optimization strategy. We provide the implementation details in Section III followed by several experiments on synthetic, *ex vivo* and *in vivo* data in Section IV.

II. Methods

A. Problem setup and notation

Let $\mathcal{S}_{pre} = \{S_1, \cdots, S_N\}$ be the set of N segmented structures of interest in the preoperative spatial domain, $\Omega_{pre} \subset \mathbb{R}^3$, where S_i represents the surface of the i^{th} structure. Also, let $\boldsymbol{P}_\ell^{pre,i} = (X_\ell^{pre,i}, Y_\ell^{pre,i}, Z_\ell^{pre,i}) \in S_i$ be the coordinates of the ℓ^{th} point on S_i. During the surgery, due to tools' pressure, abdominal insufflation and organs' physiological motions, tissues deform non-rigidly and the segmented structures do not appear in the intraoperative scene as they do in the preoperative data. To handle this non-rigid deformation, we first generate a catalog of possible 3D deformed shapes for each structure of interest, using the DeformIt software [13]. Having the catalog of 3D shapes for each structure, we model the variability in the i^{th} structure's shape through principal component analysis (PCA). We then estimate a novel 3D shape of i^{th} structure by

$$\boldsymbol{P}^{pre,i} = \overline{\boldsymbol{P}^{pre,i}} + \boldsymbol{U}^i \boldsymbol{w}^i \ , \tag{1}$$

where $\boldsymbol{U}^i = \{u_1^i, \cdots, u_{K_i}^i\}$ are the modes of variation, $\boldsymbol{w}^i = \{w_1^i, \cdots, w_{K_i}^i\}$ are the shape weights and K_i is the number of i^{th} structure's principal modes of variation.

In our clinical application, the number of camera views is limited to the two channels of the stereo endoscopic camera. Nevertheless, here we present a more general formulation for an arbitrary number of camera views. Assuming we have M camera views, we represent the 3-channel RGB image of the m^{th} camera by $I_m : \Omega_{2D}^m \subset \mathbb{R}^2 \to \mathbb{R}^3$ where Ω_{2D}^m is the domain of m^{th} camera view.

Our objective is to find an optimal global transformation, $\boldsymbol{T}$, an affine mapping transform described in the next section in (2), that brings the deformed preoperative models $\boldsymbol{P}^{pre}$ into the cameras' domain. In addition, we find the optimal shape parameters $\boldsymbol{w}^i$ for each structure of interest in the 3D preoperative domain such that the projection of each transformed and deformed 3D model, using the endoscopic camera parameters, align with the corresponding structures in the 2D images $I_1, \cdots, I_M$ and label each structure.

We set the first camera ($m = 1$) domain to be the reference surgical spatial domain ($\Omega_{srg} \subset \mathbb{R}^3$). Let $\boldsymbol{P}_\ell^{srg} = (X_\ell^{srg}, Y_\ell^{srg}, Z_\ell^{srg})$ be a point in the 3D surgical scene and $\boldsymbol{p}_\ell^m = (x_\ell^m, y_\ell^m)$ be its corresponding projected point on I_m.

Projecting a point from the camera domain to Ω_{2D}^m requires the camera parameters. The m^{th} camera parameters are the camera's focal point (f_m^x, f_m^y), camera's principal point (c_m^x, c_m^y), radial distortion parameters (k_m, k_m') and decentering distortion coefficients (p_m, p_m'). In addition, in our notation, the rotation matrix $\boldsymbol{R}_m$ and the translation vector $\boldsymbol{t}_m$ denote the extrinsic parameters between the m^{th} and the 1^{st} cameras.

Accurate projection of the preoperative models onto the intraoperative images requires calibration of the cameras prior to the operation. The calibration remains valid until the focus and/or zoom of the cameras are altered. Since some of the camera parameters change during the operation, we treat the camera parameters as unknowns to be estimated, along with $\boldsymbol{T}$ and $\boldsymbol{w}$. By finding the correct $\boldsymbol{T}$, $\boldsymbol{w}$, and camera parameters (for 3D to 2D projection), we are able to delineate the structures of interest in M intraoperative images. We find these parameters by optimizing an energy functional that is defined in the next section. The surgical setup along with our notations are summarized in Figure 1.

B. Energy minimization formulation

1) ***Unknowns:*** In this section we introduce the unknowns of our problem: global transformation, shape and camera parameters, and the segmentation boundaries.

a) Segmentation boundaries: We define $\boldsymbol{c}_m^i$ to be the boundary of the projection of the i^{th} segmented structure in 3D onto I_m (Figure 1). Each boundary $\boldsymbol{c}_m^i$ partitions the

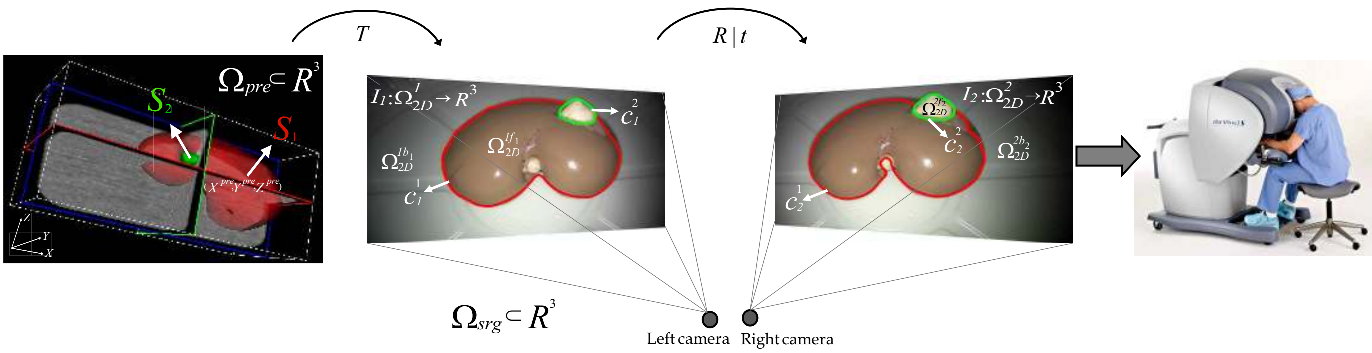


Fig. 1: Surgical setup and notation. From left to right: Segmented 3D preoperative CT, left endoscopic channel, right endoscopic channel, da Vinci surgical robot console.

domain of I_m into foreground, Ω_{2D}^{m,f_i}, and background, Ω_{2D}^{m,b_i}. Assuming N structure types may appear in $I_1, \cdots, I_M$, we aim to find the contour $\boldsymbol{c}_m^i$ for all N structures and in all M 2D images such that it delineates the i^{th} object of interest in I_m.

b) Shape parameters and transformation: The transformation $\boldsymbol{T}$ maps the points from Ω_{pre} to Ω_{srg}. Neglecting for the moment the non-rigid deformation of organs, having the correct transformation $\boldsymbol{T}$, the projection of the transformed segmented preoperative model onto $I_1, \cdots, I_M$, would segment these images properly. However, due to difference in patient position during the operation[1], abdominal insufflation and pressure from other organs and/or laparoscopic surgical tools, different structures will be deformed non-rigidly. Hence, a rigid transformation alone is not sufficient to estimate the pose and to correctly segment the structures in the endoscopic view. Therefore, we include shape parameters $\boldsymbol{w}$ in our optimization problem using (1). To keep our framework general, we choose $\boldsymbol{T}$ to be an affine mapping that transforms $\boldsymbol{P}^{pre}$ from Ω_{pre} to the surgical coordinate frame Ω_{srg} while accounting for linear deformations (e.g. stretching or shearing):

$$\begin{bmatrix} \boldsymbol{P}^{srg} \\ 1 \end{bmatrix} = \boldsymbol{T}(\boldsymbol{P}^{pre}) = \boldsymbol{R} \begin{bmatrix} \boldsymbol{P}^{pre} \\ 1 \end{bmatrix} + \boldsymbol{t}$$
$$\begin{bmatrix} X^{srg} \\ Y^{srg} \\ Z^{srg} \\ 1 \end{bmatrix} = \begin{bmatrix} q_1 & q_2 & q_3 & 0 \\ q_4 & q_5 & q_6 & 0 \\ q_7 & q_8 & q_9 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X^{pre} \\ Y^{pre} \\ Z^{pre} \\ 1 \end{bmatrix} + \begin{bmatrix} q_{10} \\ q_{11} \\ q_{12} \\ 0 \end{bmatrix}, \qquad (2)$$

where $q_1 \cdots q_{12}$ are the 12 degrees of freedom of an affine transformation.

c) Camera parameters: In Section II-A we mentioned the importance of correcting the camera parameters during the surgery. It has been observed that the effect of zoom is negligible on three out of eight camera parameters: k', p, and p' [10], [38]. Therefore, we aim to correct the remaining five camera parameters by finding $\pi_m = (f_m^x, f_m^y, c_m^x, c_m^y, k_m)$ for all M cameras.

[1]During the preoperative scan, patients are in the supine position while during surgery they are lying sideways (lateral recumbent position).

*2) **Objective function**:* We define the energy functional E which is used to segment $I_1, \cdots, I_M$ according to the preoperative shape models obtained from (1). In our formulation, we represent the contour $\boldsymbol{c}_m^i$ implicitly by its corresponding level set function, ϕ_m^i, where

$$\begin{cases} \phi_m^i(\boldsymbol{x}) > 0, & \forall \boldsymbol{x} \in \Omega_{2D}^{m,f_i} \\ \phi_m^i(\boldsymbol{x}) < 0, & \forall \boldsymbol{x} \in \Omega_{2D}^{m,b_i} \\ \phi_m^i(\boldsymbol{x}) = 0, & \forall \boldsymbol{x} \in \partial\Omega_{2D}^{m,f_i} \end{cases} . \qquad (3)$$

The contour $\boldsymbol{c}_m^i$ is represented by the zero level sets of ϕ_m^i, i.e. $\boldsymbol{c}_m^i = \{\boldsymbol{x} \in \Omega_{2D}^m | \phi_m^i(\boldsymbol{x}) = 0\}$.

Having N structure types appear in M camera images, we define the following energy functional E to segment the images:

$$\begin{aligned} &E(\boldsymbol{\Phi}, \boldsymbol{T}, \boldsymbol{W}, \boldsymbol{\pi}; \boldsymbol{P}^{pre}, I_1, \cdots I_M) = \\ &\sum_{n=1}^{N} \sum_{m=1}^{M} \int_{\Omega_{2D}^m} \Big(g_f(n,m,\boldsymbol{x})\, H(\phi_m^n(\boldsymbol{x})) \\ &\qquad + g_b(n,m,\boldsymbol{x})\, (1 - H(\phi_m^n(\boldsymbol{x}))) \Big) d\boldsymbol{x} , \qquad (4) \end{aligned}$$

where $H(.)$ is the heaviside function and

$$\boldsymbol{\Phi} = \{\phi_1^1, \cdots, \phi_1^N, \cdots, \phi_M^1, \cdots, \phi_M^N\} , \qquad (5)$$
$$\phi_m^n(\boldsymbol{x}) = SDM\Big(\partial\Big(\mathcal{P}_m\big(\boldsymbol{T}(\boldsymbol{P}^{pre,n})\big)\Big)\Big) , \qquad (6)$$
$$\boldsymbol{W} = \{\boldsymbol{w}^1, \cdots, \boldsymbol{w}^N\} , \qquad (7)$$
$$\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_M\} . \qquad (8)$$

$SDM(.)$ in (6) is the signed distance map and $\mathcal{P}_m : \Omega_{srg} \to \Omega_{2D}^m$ is the projection from the surgical-scene frame of reference to Ω_{2D}^m. $\boldsymbol{\pi}$ are the cameras parameters and $g_f(n,m,\boldsymbol{x})$ and $g_b(n,m,\boldsymbol{x})$ are the regional terms that measure the agreement of the image pixels $\boldsymbol{x}$ with the inside and outside statistical models of the n^{th} structure in the m^{th} image (I_m). $g_f(n,m,\boldsymbol{x})$ and $g_b(n,m,\boldsymbol{x})$ are calculated as follows

$$\begin{aligned} g_f(n,m,\boldsymbol{x}) &= -\log\ p_f^n(\boldsymbol{x}|I_m(\boldsymbol{x})) \\ g_b(n,m,\boldsymbol{x}) &= -\log\ p_b^n(\boldsymbol{x}|I_m(\boldsymbol{x})) \\ n &= \{1, \cdots, N\} \qquad (9) \\ m &= \{1, \cdots, M\} , \end{aligned}$$

where $p_f^n(\boldsymbol{x}|I_m(\boldsymbol{x}))$ and $p_b^n(\boldsymbol{x}|I_m(\boldsymbol{x})) = 1 - p_f^n(\boldsymbol{x}|I_m(\boldsymbol{x}))$ are the probabilities of a given pixel $\boldsymbol{x}$ belonging respectively to the inside and to the outside the n^{th} object in I_m (for more details refer to Section III).

Given the first camera (reference camera) parameters, the projection from 3D space Ω_{srg} to the first 2D image I_1 (i.e. $\mathcal{P}_1$) is calculated as follows:

$$
\begin{aligned}
(x_1, y_1) &= \mathcal{P}_1(X^{srg}, Y^{srg}, Z^{srg}) \\
x'_1 = X^{srg}/Z^{srg} \ , &\quad y'_1 = Y^{srg}/Z^{srg} \ , \\
r_1^2 &= {x'}_1^2 + {y'}_1^2 \ , \\
x''_1 = x'_1(1 + k_1 r_1^2 + k'_l r_1^4) &+ 2p_1 x'_1 y'_1 + p'_1(r_1^2 + 2{x'}_1^2) \ , \\
y''_1 = y'_1(1 + k_1 r_1^2 + k'_1 r_1^4) &+ p_1(r_1^2 + 2{y'}_1^2) + 2p'_1 x'_1 y'_1 \ , \\
x_1 = f_1^x \ x''_1 + c_1^x \ , &\quad y_1 = f_1^y \ y''_1 + c_1^y \ ,
\end{aligned} \tag{10}
$$

where k_1 and k'_1 are the radial distortion and p_1 and p'_1 are the decentering distortion parameters of the first camera. Also (f_1^x, f_1^y) and (c_1^x, c_1^y) are first camera's focal and principal points, respectively. To calculate $\mathcal{P}_m$ for $m \neq 1$, we first need to bring the points in Ω_{srg} to the m^{th} camera's frame of reference using the camera's extrinsic parameters, $\boldsymbol{R}_m$ and $\boldsymbol{t}_m$,

$$
\begin{bmatrix} X^m \\ Y^m \\ Z^m \end{bmatrix} = \boldsymbol{R}_m \begin{bmatrix} X^{srg} \\ Y^{srg} \\ Z^{srg} \end{bmatrix} + \boldsymbol{t}_m \ . \tag{11}
$$

Then, $\mathcal{P}_m$ is calculated similar to (10) by using (X^m, Y^m, Z^m) instead of $(X^{srg}, Y^{srg}, Z^{srg})$.

C. Optimization

Given the objective function E, our optimization task is:

$$
\boldsymbol{T}^*, \boldsymbol{W}^*, \boldsymbol{\pi}^* = \underset{\boldsymbol{T}, \boldsymbol{W}, \boldsymbol{\pi}}{\text{argmin}} \ E \ . \tag{12}
$$

Note that the boundary of structures in each image ($\boldsymbol{c}_m^i$) are obtained by calculating the zero level of their corresponding level set functions ϕ_m^i. The level set functions are calculated from (6) after finding the optimal $\boldsymbol{T}$, $\boldsymbol{W}$ and $\boldsymbol{\pi}$.

For computational efficiency, for each frame, we optimize (12) with respect to each set of variables $\boldsymbol{T}$, $\boldsymbol{W}$ and $\boldsymbol{\pi}$ successively and repeat this process until convergence, i.e the change in the unknowns of all three smaller optimizations is small enough.

In the first stage, we optimize (12) for $\boldsymbol{T}$ and find the transformation parameters $q_1, \cdots, q_{12}$. To find the proper pose of the preoperative model with respect to the first frame of the video efficiently, we use a hierarchical multi-resolution approach in which the resolution of the preoperative data and 2D images are increased in a coarse to fine fashion. In the second stage, we optimize (12) for the shape parameters, $\boldsymbol{W}$, followed by optimizing (12) for the cameras' parameters, $\boldsymbol{\pi}$, in the third stage. Note that it is not necessary to update the camera parameters continuously and this third stage of our optimization framework is turned on as soon as surgeons change the focus/zoom of the cameras. These three optimization steps are repeated until convergence.

To find the pose and shape parameters, we compute the derivative of E with respect to the finite set of optimization variables $\xi = \{\xi_1, \cdots, \xi_\ell\}$, where ξ_ℓ can be the transformation parameters q_ℓ or shape parameters w_ℓ:

$$
\frac{\partial E}{\partial \xi_\ell} = \sum_{n=1}^{N} \sum_{m=1}^{M} \Bigg(\int_{\Omega_{2D}^m} \big(g_f(n,m,\boldsymbol{x}) - g_b(n,m,\boldsymbol{x})\big) \frac{\partial H(\phi_m^n(\boldsymbol{x}))}{\partial \xi_\ell} \Bigg) \ , \tag{13}
$$

where

$$
\begin{aligned}
\frac{\partial H(\phi_m^n(\boldsymbol{x}))}{\partial \xi_\ell} &= \frac{\partial H(\phi_m^n(\boldsymbol{x}))}{\partial \phi_m^n} \left(\frac{\partial \phi_m^n}{\partial x} \frac{\partial x}{\partial \xi_\ell} + \frac{\partial \phi_m^n}{\partial y} \frac{\partial y}{\partial \xi_\ell} \right) \\
&= \delta(\phi_m^n) \begin{bmatrix} \frac{\partial \phi_m^n}{\partial x} & \frac{\partial \phi_m^n}{\partial y} \end{bmatrix} \begin{bmatrix} \frac{\partial x}{\partial \xi_\ell} \\ \frac{\partial y}{\partial \xi_\ell} \end{bmatrix} \ .
\end{aligned} \tag{14}
$$

In (14), $\delta(.)$ is the Dirac delta function. We use the centered finite difference to calculate $\frac{\partial \phi}{\partial x}$ and $\frac{\partial \phi}{\partial y}$. Every 2D point $\boldsymbol{p}^m = (x^m, y^m)$ in I_m has at least one corresponding 3D point $\boldsymbol{P}^{srg} = (X^{srg}, Y^{srg}, Z^{srg})$ in the surgical domain Ω_{srg}. Each 2D point $\boldsymbol{p}^m = (x^m, y^m)$, is related to $\boldsymbol{P}^{srg}$ by the m^{th} camera parameters (cf. (10) and (11)). From (10) and for $m = 1$ we have

$$
\frac{\partial x_1}{\partial \xi_\ell} = f_1^x \frac{\partial x''_1}{\partial \xi_\ell} \ , \tag{15}
$$

$$
\frac{\partial y_1}{\partial \xi_\ell} = f_1^y \frac{\partial y''_1}{\partial \xi_\ell} \ . \tag{16}
$$

$\frac{\partial x_m}{\partial \xi_\ell}$ and $\frac{\partial y_m}{\partial \xi_\ell}$ are calculated in a similar way.

Considering (10), to calculate (15) and (16), we need to calculate $\frac{\partial X^{srg}}{\partial \xi_\ell}$, $\frac{\partial Y^{srg}}{\partial \xi_\ell}$ and $\frac{\partial Z^{srg}}{\partial \xi_\ell}$. Every point in the surgical frame of reference $\boldsymbol{P}^{srg}$ is related to a point in the preoperative domain, $\boldsymbol{P}^{pre} \in \Omega_{pre}$ by (2). For the transformation parameters, $\xi = \{q_1, \cdots, q_{12}\}$, the derivative of a 3D point $\boldsymbol{P}^{srg}$ in the surgical domain with respect to q_i is summarized in Table II.

TABLE II: Partial derivatives of 3D points in the surgical frame $\boldsymbol{P}^{srg}$ with respect to the transformation parameters, q_i in (2).

q_i	$\frac{\partial X^{srg}}{\partial q_i}$	$\frac{\partial Y^{srg}}{\partial q_i}$	$\frac{\partial Z^{srg}}{\partial q_i}$
q_1	X^{pre}	0	0
q_2	Y^{pre}	0	0
q_3	Z^{pre}	0	0
q_4	0	X^{pre}	0
q_5	0	Y^{pre}	0
q_6	0	Z^{pre}	0
q_7	0	0	X^{pre}
q_8	0	0	Y^{pre}
q_9	0	0	Z^{pre}
q_{10}	1	0	0
q_{11}	0	1	0
q_{12}	0	0	1

For the shape parameters $\xi = \{w_1, \cdots, w_\ell\}$, recalling (1), the derivative of a 3D point $\boldsymbol{P}^{srg}$ in the surgical domain (Ω_{srg}) with respect to w_ℓ is

$$
\begin{aligned}
\frac{\partial \boldsymbol{P}^{srg}}{\partial w_\ell} &= \frac{\partial \boldsymbol{R}}{\partial w_\ell} \boldsymbol{P}^{pre} + \boldsymbol{R} \frac{\partial \boldsymbol{P}^{pre}}{\partial w_\ell} \\
&= \boldsymbol{R} \frac{\partial \boldsymbol{P}^{pre}}{\partial w_\ell} \\
&= \boldsymbol{R} \cdot u_\ell ,
\end{aligned} \tag{17}
$$

where u_ℓ is the ℓ^{th} mode of variation in $\boldsymbol{U}$. Considering the extrinsic parameters between cameras, $\frac{\boldsymbol{P}^m}{\partial w_\ell}$ is calculated in a similar way, i.e. $\frac{\partial \boldsymbol{P}^m}{\partial w_\ell} = \boldsymbol{R}_m \boldsymbol{R} \cdot u_\ell$.

In the last stage of our optimization, we optimize (12) with respect to the camera parameters, π_m, mentioned in Section II-B1c. The derivative of E with respect to π_m is computed similar to (13) and (14). Table III summarizes the derivatives of 2D points $\boldsymbol{p}^m = (x^m, y^m)$ in I_m with respect to the camera parameters.

TABLE III: Partial derivatives of 2D points $\boldsymbol{p}^m$ in I_m with respect to camera parameters π_m (cf. (10)).

π^m	f_m^x	f_m^y	c_m^x	c_m^y	k_m
$\frac{\partial \boldsymbol{x}^m}{\partial \pi_m}$	x''_m	0	1	0	$f_m^x x'_m r_m^2$
$\frac{\partial \boldsymbol{y}^m}{\partial \pi_m}$	0	y''_m	0	1	$f_m^y y'_m r_m^2$

The boundary of the segmented structures are the zero level set of their corresponding level sets function ϕ that is obtained by (6) after finding the optimal $\boldsymbol{T}$, $\boldsymbol{W}$ and $\boldsymbol{\pi}$.

III. Implementation details

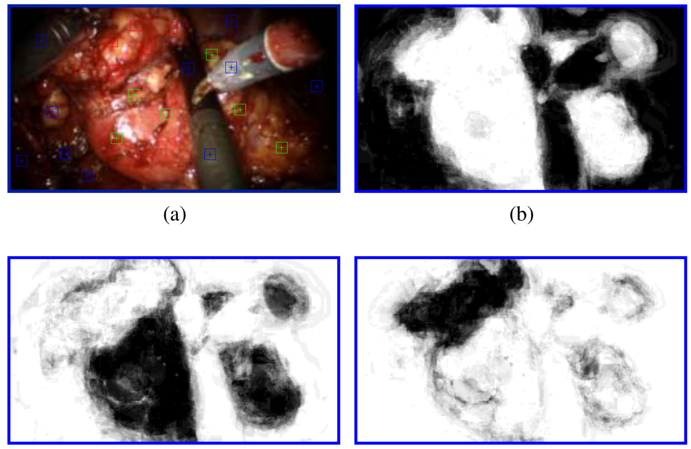
To make our method as robust as possible, we leverage a variety of image features to calculate the regional terms, $g_f(n, m, \boldsymbol{x})$ and $g_b(n, m, \boldsymbol{x})$ in (4), for different structures which may have different discriminative color and texture features.

We concatenate the three normalized RGB channels, the three YCbCr channels and their local color histogram features as regional cues into the regional appearance vector $\mathcal{A}_m$ for 2D images (i.e. left and right channels of stereoscopic video, $M = 2$). $p_f^n(\boldsymbol{x}|I_m(\boldsymbol{x}))$ and $p_b^n(\boldsymbol{x}|I_m(\boldsymbol{x})) = 1 - p_f^n(\boldsymbol{x}|I_m(\boldsymbol{x}))$ in (9) are estimated by training a random forest (RF) consisting of N_b binary decision trees (here $N_b = 20$). To train the RF, we select several patches in I_1 and I_2 from different structures (i.e. supervised training) in the first 2% of all frames, i.e. the first ∼10 frames out of ∼600 frames. Figure 2(a) shows a sample seeding on a sample video frame of real clinical data. In practice, surgeons may select these patches with the help of surgical tools. After training, for each pixel $\boldsymbol{x}$, the feature channels, $\mathcal{A}_m(\boldsymbol{x})$, are propagated through each RF tree resulting in the probability $p_j(\boldsymbol{x} \in Structure_i | \mathcal{A}_m(\boldsymbol{x}))$, for the j^{th} tree. These probabilities are combined into a forest's joint probability $p_f^i(\boldsymbol{x}|I_m(\boldsymbol{x})) = \frac{1}{N_b}\sum_{j=1}^{N_b} p_j(\boldsymbol{x} \in Structure_i | \mathcal{A}_m(\boldsymbol{x}))$ to determine the probability of $\boldsymbol{x}$ belonging to i^{th} structure. Figure 2(b-d) illustrates examples of regions probability for the frame shown in Figure 2(a). A lower intensity in Figures 2(b-d) corresponds to higher probability.

We emphasize that unlike feature-based methods (e.g. [28], our method does not require any correspondence between 3D CT and the 2D intraoperative data. The surgeon only needs to provide a few clicks on the object of interest (e.g. kidney/tumour) and background without knowing their corresponding points in the preoperative CT.

IV. Experiments

In this section, we provide several experimental results over synthetic, *ex vivo* and *in vivo* datasets to demonstrate the performance of our framework applied to partial nephrectomy. We also evaluate the robustness of our framework to initialization, noise, non-rigid deformation and occlusion.



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

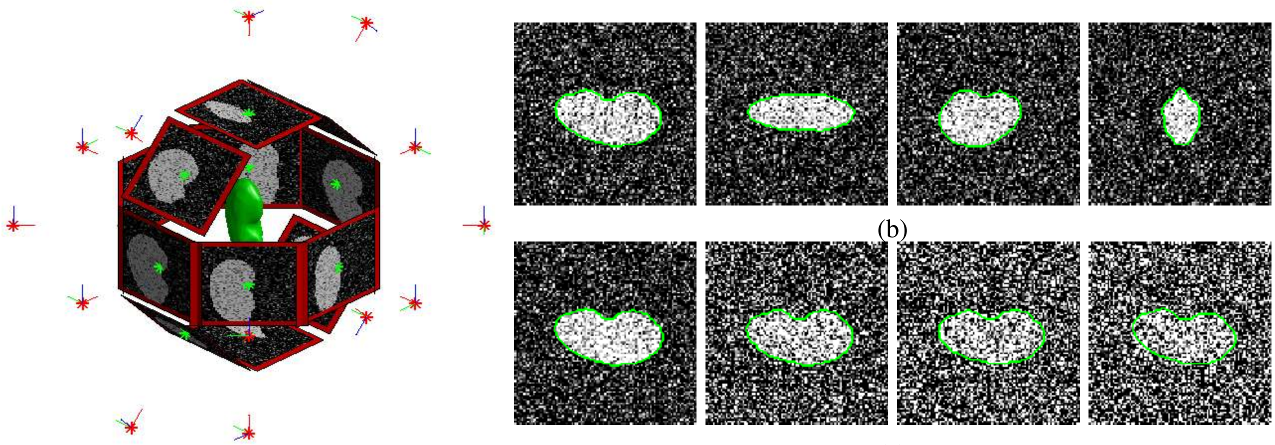
Fig. 2: Examples of regions probability. (a) Seed selection from a random *in vivo* frame (blue: background/tools, green: kidney, red: tumour). (b-d) Probability of background, kidney and tumour for the frame shown in (a). A lower intensity in (b-d) corresponds to higher probability.

A. Synthetic data

For our first set of synthetic experiments, we created a virtual kidney with 14 2D images by rotating a virtual camera around the synthetic kidney and calculating the projection of the kidney on the 2D planes (Figure 3(a)). Each 2D plane was polluted with additive white Gaussian noise with a standard deviation of $\sigma = 0.75$. These 2D images have been created by projecting the 3D object onto each 2D plane using eq. (10).

Our first synthetic test evaluates the pose recovery capabilities (robustness to initial pose) of our method. We perturbed the correct pose of our model in θ_x, θ_y, θ_z, t_x, t_y, t_z, s_x and s_y, which are the rotations around x, y and z axes, translations in x, y and z directions and scaling in x and y directions, respectively. With regard to translation, our framework is able to recover t_x, t_y and t_z as long as the projection of the 3D model intersects the object in the 2D images. Table IV shows the range of perturbation for θ_x, θ_y, θ_z, s_x and s_y over which our method is able to recover the correct pose. To show the benefit of multi-view feature of our method over other methods with a single view (e.g. monocular camera), we performed the pose recovery test using 1 up to 14 camera views. As evident from Table IV, as the number of camera views increases, our method is able to handle a wider range of disturbance. Figure 3(b) illustrates the simultaneous segmentation of multiple views.

Our second synthetic experiment demonstrates the robustness of our framework to noise. We polluted the 2D images with additive white Gaussian noise with different standard deviations, σ. Figure 3(c) shows qualitative results of successful segmentation of 2D images with different noise levels. We compared our framework with active contour without



(a) (b) (c)

Fig. 3: Pose recovery and robustness to noise in the synthetic experiment set-up. (a) Multiple views of a synthetic kidney. (b) Simultaneous segmentations of multiple views polluted with additive white noise with $\sigma = 0.25$. (c) Robustness to noise. Images from left to right are polluted by Gaussian noise with $\sigma = \{0.25, 0.75, 1.25, 1.75\}$. Note that all 14 views were polluted by Gaussian noise and in (c) we only show a single view as an example. The advantage of multiple view compared to single view in pose estimation is shown in Table IV.

TABLE IV: Pose recovery range (the amount of pose deviation in which our method can still recover the correct pose) w.r.t. number of cameras/views. θ_x, θ_y, and θ_z are rotation parameters (in degree) around x, y and z axes. s_x and s_y are anisotropic scaling parameters along x and y axes, respectively.

No. of cams.	1	4	7	14
θ_x	-40 - $+40$	-40 - $+40$	-50 - $+50$	-90 - $+90$
θ_y	-80 - $+60$	-80 - $+80$	-90 - $+90$	-90 - $+90$
θ_z	-60 - $+20$	-90 - $+90$	-90 - $+90$	-90 - $+90$
s_x	0.05 - 4	0.05 - 6	0.01 - 7	0.01 - 8
s_y	0.01 - 9	0.01 - 10	0.01 - 12	0.01 - 12

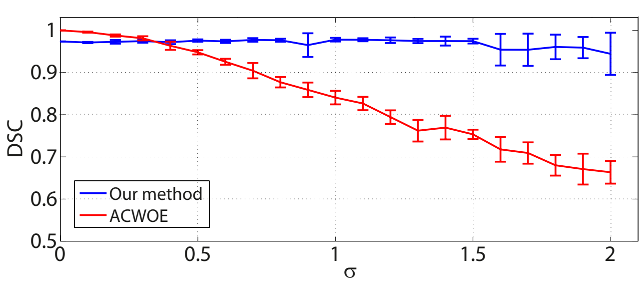
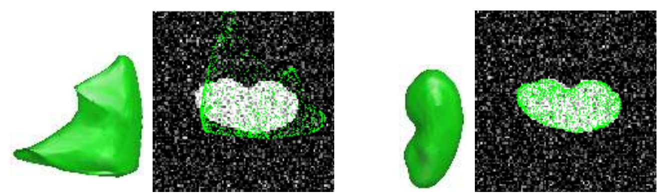


Fig. 4: Robustness to noise for our method and ACWOE. DSC w.r.t. different standard deviation (σ) of the Gaussian noise.

edges (ACWOE) [37] in terms of robustness to noise. We chose ACWOE since it uses the same energy functional we use (cf. (4)), however, ACWOE optimizes (4) with respect to ϕ instead of the pose and shape parameters. Figure 4 compares the robustness of our method and ACWOE with respect to different noise levels. As seen in Figure 4, since ACWOE only uses the intensity information to separate the foreground from the background and does not use any prior information, ACWOE fails in presence of high noise. We note that with low levels of noise, ACWOE performed slightly better than our method as ACWOE only has a single parameter to optimize for. Since our method has multiple parameters to optimize, there is a higher chance of parameters getting stuck in local minima. In fact, with low noise levels, visual cues alone are enough to segment objects of interest. However, in the presence of large noise, prior information (e.g. shape) is necessary to obtain a feasible result.



(a) (b)

Fig. 5: Shape recovery after randomly perturbing the principal modes. (a) Perturbed and (b) recovered 3D shape and its overlay on the corresponding 2D image.

In our third synthetic experiment, we randomly perturbed the principal shape parameters of our model ($\boldsymbol{w}$) and used our optimization framework to recover the correct shape. During this experiment, the pose parameters were fixed. Although the perturbed shape was not in the training set (used in PCA), our method was able to recover the shape even when the shape is largely perturbed (Figure 5). Such an exaggerated deformation is rare, nonetheless, our method was able to recover the proper shape parameters. However, due to our local optimization framework, we do not claim that our method is able to estimate the correct shape parameters for *all* severely perturbed shapes.

Figure 6 illustrates further analysis of our shape recovery technique in which the tolerable range of shape perturbation (standard deviation from the mean shape) is computed as a function of both noise level and number of views. The benefit

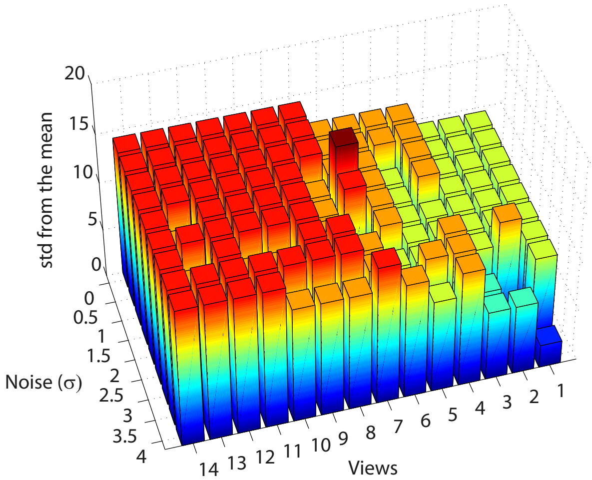


Fig. 6: Tolerable range of shape perturbation (standard deviation from the mean shape) in terms of noise level and number of views.

of handling more camera views is obvious from Figure 6. More views enable our method to tolerate more shape perturbation. Our method is fairly robust to noise. However, in the presence of severe noise and only one or two cameras, our method cannot recover large shape perturbation.

To evaluate our camera parameters estimation module, we simulated the change of zoom of the synthetic cameras by changing the focal and principal points over time and with additive noise. According to [1] and [17], the distortion parameter can be approximated by a 2nd-order polynomial function of $1/f$ where f is the camera's focal point. To have a ground truth for k, we followed [1] and approximated k using the following polynomial equation: $k = 0.226 - 0.0719(1/f) - 0.214(1/f)^2$.
Figure 7 illustrates the estimated focal and principal points over time. According to Figure 7, our method is able to estimate the focal and principal points fairly accurately and does a reasonable job in estimating the distortion parameter. We emphasize that the influence of the distortion parameter is negligible compared to the camera's focal and principal points.

B. Ex vivo lamb kidney data

In the second part of our experiments, we prepared a set of 10 *ex vivo* phantoms using lamb kidneys and implanted artificial tumours outside and inside each kidney to emulate a partially exophytic and completely endophytic tumour, respectively. To more closely simulate a real surgical environment, we print a snapshot from a real robot-assisted partial nephrectomy and used it as the background for our *ex vivo* lamb kidneys (Figures 8 and 10). A Siemens Somatom CT scanner was used to acquire a high resolution CT volume of our phantoms. We also recorded a stereo endoscopy video of each phantom using da Vinci S system at full HD 1080i resolution with 30 FPS. While capturing the stereo endoscopic video, we changed the camera's zoom repeatedly.

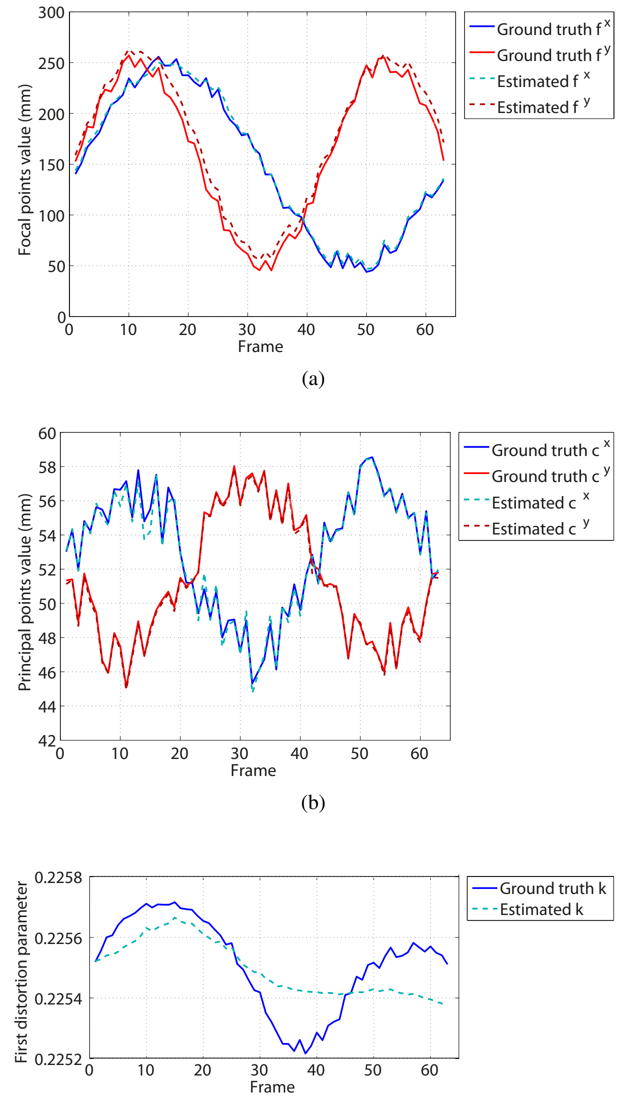


Fig. 7: Estimation of cameras' focal and principal points and the first radial distortion parameter (k).

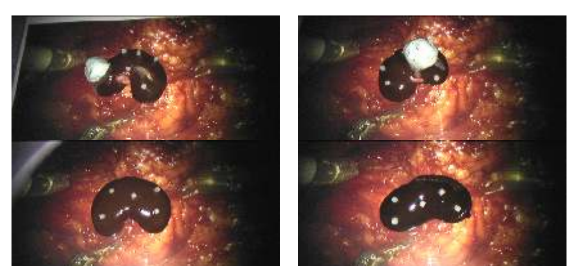


Fig. 8: Single endoscopic frame of four (out of 10) lamb kidney phantoms with exophytic (top row) and endophytic (bottom row) tumours.

We segmented the kidney and tumour in each CT using TurtleSeg [35]. To create the 3D shape catalog, we deformed the segmented kidneys and tumours using DeformIt [13]. We created pseudo-realistic deformations by carefully simulating external forces (pulling and pushing) on the surface of the

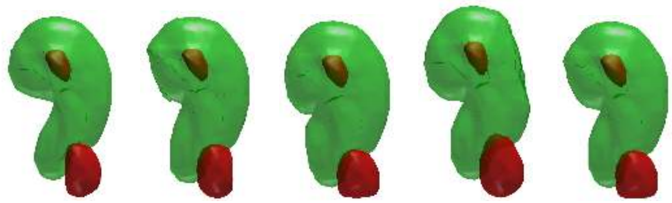
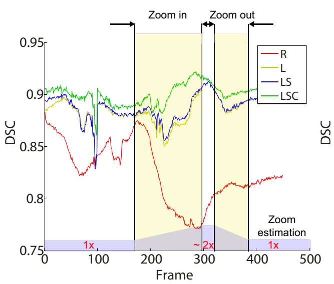
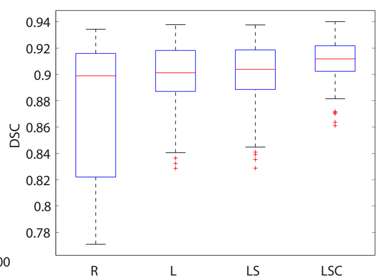


Fig. 9: Examples of shape variations of a kidney and its two tumours after deformations using DeformIt software [13]. Both endophytic and exophytic tumours are in red and kidneys are in green.

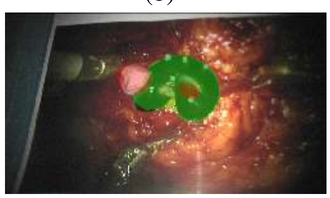


(a)

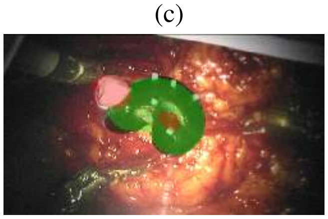


(b)

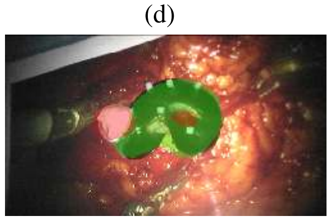
(c)



(d)



(e)



(f)

Fig. 10: Quantitative and qualitative result for an *ex vivo* phantom. (a) DSC vs. frame (time) for different settings of our framework: R: rigid; L: linear; LS: linear with shape optimization; LSC: linear with shape optimization and camera parameters correction. (b) Box plot representation of the Dice score over time (frame) of our prior-based segmentation for different settings. (c) Initial pose (before convergence). (d) Recovered pose and shape after convergence. (d-f) Corresponding qualitative results on three (out of 480) frames (Green: kidney; Red: tumour). Large and small red labels represent exophytic and endophytic tumours, respectively.

segmented objects and then calculated the corresponding deformations [13]. Each kidney and tumour were deformed in $\sim$40 and $\sim$15 different ways, respectively. Figure 9 shows a few samples of deformed objects in the catalog of a phantom example. To obtain the segmentation ground truth for our stereo video data, we used the "Rotobrush" tool of Adobe After Effect CS6 (Adobe Systems Inc.) as a semi-automatic video segmentation tool allowing for visual inspection and correction. On average, the Rotobrush segmentation of each stereo video took about 15 minutes.

We automatically segmented 10 phantom stereo videos using the proposed framework with four different settings:

1) $\boldsymbol{T}$ is rigid, no shape optimization, no camera parameters correction (R),
2) $\boldsymbol{T}$ is linear, no shape optimization, no camera parameters correction (L),
3) $\boldsymbol{T}$ is linear, with shape optimization, no camera parameters correction (LS),
4) $\boldsymbol{T}$ is linear, with shape optimization, with camera parameters correction (LSC).

We then reported the Dice similarity coefficient (DSC) versus time for all visible tissues (kidney and/or partially visible tumours). Figure 10(a) shows a sample DSC vs. time for one of our phantom cases with sample qualitative results on three (out of 480) frames. The frames with zoom transition as well as the zoom estimation for each frame have been highlighted. As expected, as we increase the number of degrees of freedom (rigid vs. linear vs. non-rigid) the results become more accurate. Also, including the camera parameters in the optimization procedure improves the accuracy, particularly where the zoom starts to change. Figure 10(b) illustrates the box plot results for each of the above mentioned settings for the entire *ex vivo* phantom data. Since we do not have the ground truth segmentation (and it is nontrivial how one may obtain it) for endophytic tumours, the DSC includes kidney and exophytic tumors only. Figure 10(c) shows the initial pose, which despite being not well placed, results in a reasonable pose as shown in Figure 10(d). We emphasize that this phantom has both exophytic and endophytic tumours. Since there is no visual cues for the endophytic tumour, we did not consider this tumour in our optimization procedure. However, we showed the endophytic tumour along with kidney and exophytic tumour for visualization purposes. As it is seen in Figure 10(d-f), preoperative based endoscopic segmentation, enables surgeons to roughly locate the underlying tissues (endophytic tumour in this example) even though they are not visible with the naked eye, which can facilitate their decision making.

C. *In vivo clinical study*

We also applied our framework on five different clinical cases of robot assisted partial nephrectomy. For each patient, we tested our method on 20 seconds of stereo endoscopic videos from the tumour demarcation stage captured by a da Vinci surgical system with a frame rate of 30 FPS where the tumours and kidneys were segmented in the corresponding CT data using the TurtleSeg software [35]. To obtain the ground truth segmentation for each patient, we observed the tumour demarcation and resection stages of each partial nephrectomy video and localized the tumour and kidney boundaries with the help of a urologist. After confirming the tumour and kidney boundaries for several frames with the urologist, we segmented the remaining frames semi-automatically, using Adobe After Effects CS6 software. Each stereo video took $\sim$3.5 hours to segment semi-automatically.

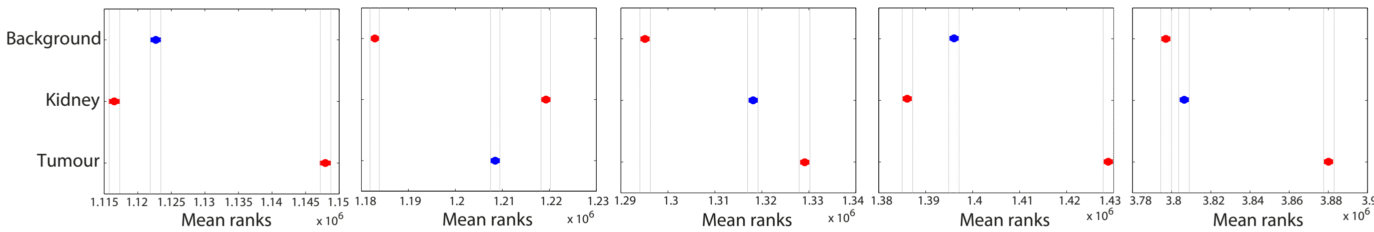
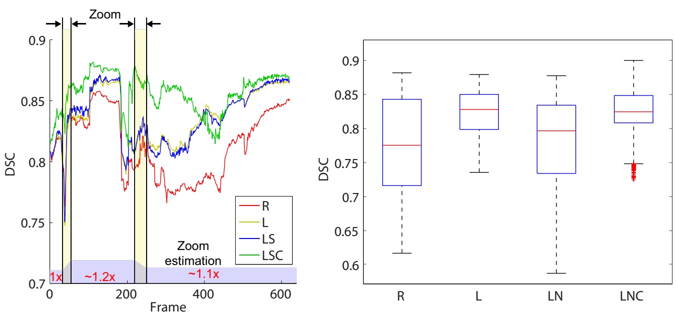


Fig. 11: The Kruskal-Wallis test for all five clinical cases rejects the null hypothesis that the data in each category (background, kidney, and tumour) comes from the same distribution. Every pairs of groups have mean ranks significantly different from the third group.



(a) (b)

Fig. 12: Our framework evaluation on *in vivo* partial nephrectomy surgery. (a) DSC over time (frame) of an *in vivo* partial nephrectomy for different settings of our framework. (b) Box plot representation of the DSC over time (frame) on three *in vivo* cases for different setting. R: rigid; L: linear; LS: linear with shape optimization; LSC: linear with shape optimization and camera parameters correction.

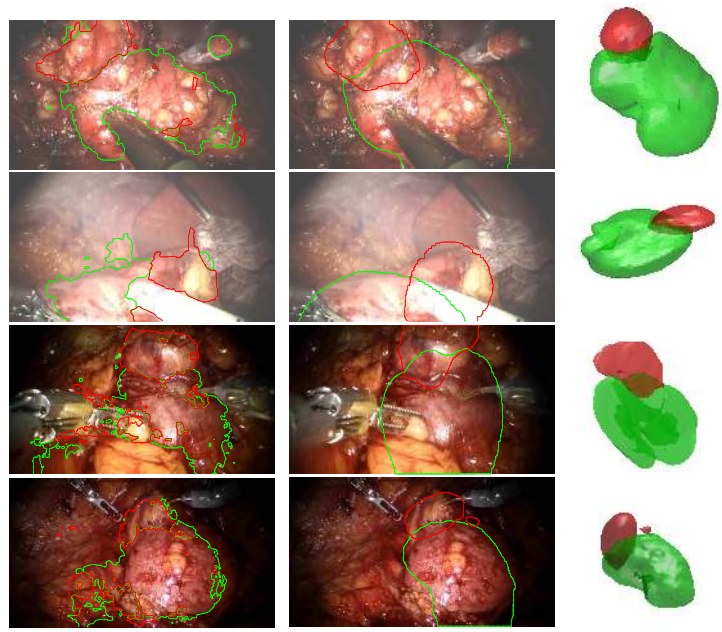


Fig. 13: Robustness to occlusion in *in vivo* real surgery cases. 1^{st} column: segmentation results without prior information from the preoperative data using multiphase active contour without edges [37]. 2^{nd} column: Our results with corresponding 3D rendering in the 3^{rd} column.

We note that the motivation of our proposed technique was to address the tumour demarcation stage of a partial nephrectomy. This is a critical stage where surgeons mark the boundaries of tumours before initiating any cutting. The appearance of kidney/tumour remains the same during the tumour demarcation stage. Therefore, training the structures' appearance only once would be enough for our system.

To show that the kidney, tumour and the background have different distribution (using our features explained in Section III), we performed the Kruskal-Wallis test and reported the p-value for the null hypothesis that the data in each category comes from the same distribution. The p-value we obtained for each of our five clinical *in vivo* cases was close to zero. Our additional follow-up test to confirm that data samples of different groups (background, kidney and tumour) come from different distributions are shown in Figure 11).

The Dice similarity coefficient versus time (for one of the patients) along with the box plot results (for all three patients) are shown in Figure 12 for different experimental settings. The frames with zoom transition as well as the zoom estimation for each frame have been highlighted in Figure 12(a). The overall DSC is lower than our *ex vivo* lamb kidney experiments which is mainly due to the noise, lighting effects, tools crossing and appearance similarity between different tissues. Nevertheless, our method was able to achieve DSC close to 0.85 for real *in vivo* clinical cases.

In our last experiment, we qualitatively compared our method with the popular level sets method [37] using the same energy functional as (4). Figure 13 shows how our prior-based method is able to properly segment the endoscopic image even with tissues occlusions (due to tools crossing), unlike the level sets-based method in the first row.

During our experiments on *in vivo* cases, we chose to optimize the camera parameters during the entire duration of the videos (LSC setting). We observed that in the cases where tools occlude most of the tissues, the weak data term (due to the occlusion) mislead the camera optimization procedure and results got worse in the consecutive frames due to the wrong camera parameters. Therefore, as we suggested before, the best time to turn on the camera correction module is when surgeons change the focus/zoom. This information can easily be obtained from the da Vinci surgical system's API which was not available in our experiments.

V. Discussion and conclusions

We proposed a novel technique to segment multiple structures in multiple views of endoscopic videos that leverages both preoperative data, as a source of prior knowledge, as well as endoscopic visual cues (by training a random decision forest). Our method enables accurate segmentation in the highly noisy and cluttered environment of an endoscopic video including tissues occlusion. In our framework, surgeons will be able to identify and locate deep/hidden structures (e.g. endophytic tumours) which will guide in their decision making. In addition, we incorporated a camera parameter correction module into our unified optimization-based framework that can be used when the focus/zoom changes. Our results on synthetic, *ex vivo* and *in vivo* clinical cases of partial nephrectomy illustrate the great potential of the proposed framework for augmented reality applications in MIS.

The strengths of our method can be summarized by that it: (i) segments multiple structures; (ii) handles multi-view endoscopic videos; (iii) estimates the 3D pose of the structures in the 3D preoperative space; (iv) accounts for non-rigid deformation of structures during the operation; (v) corrects camera parameters (as needed) when zoom and/or focus changes; (vi) does not require any point correspondence between preoperative and intraoperative data for pose estimation (a few clicks on the objects of interest, e.g. kidney/tumour, and background is sufficient to guide the overlaying process); and (vii) is highly parallelizable.

In addition, using a region-based formulation helps our method handle occlusions better compared to methods that are based on detecting and tracking few number of features/salient points.

Note that although surgeons only see the visible part of the kidney and tumour, overlaying the 3D preoperative data onto the 2D scene has the potential to increase surgeons' confidence and reduce operation time during the time-sensitive portions of the surgery. More importantly, this 3D to 2D alignment helps surgeons to appreciate where the invisible structures hidden underneath of tissues lie. The small red artificial tumour in Figure 10 and part of the boundaries of the exophytic tumours in Figure 13 are in fact hidden structures that cannot be seen with the naked eye in the endoscopic video. We emphasize that, in this work we assumed that correct segmentation and overlay of kidney and exophytic tumours results in a reasonable alignment of internal structures (i.e. structures inside the kidney). However, as we did not have the ground truth for internal structures we could not evaluate them quantitatively.

Our method has some notable limitations. Our method relies on a local optimization framework, therefore the final segmentation and pose recovery results will depend on the initial pose. Obviously a poor pose initialization can result in an incorrect solution. However, our experiments suggest that even with rough initialization that is distant from the desired solution we can obtain reasonable results (Figure 10(c)).

Another challenge was how to calculating the visual cues. Due to the large variability among patients, training the RF on other patients and testing on the current patient did not give reasonable results. Therefore, we used the current patient data and selected image patches from the first few frames of the video sequence to train patient specific random forest models. This is an important and perhaps the most challenging problem in image-guided intervention problems that is worthy of future exploration. One possible direction for future research to address this issue is to leverage recent research in transfer learning and domain adaptation [36], [14].

Although we limited the shape parameters to vary not more than three times their standard deviation, the global transformation and camera parameters are still unconstrained. Constraining these variables in the optimization not only can increase the convergence speed but also can improve the final results. In future work, such constraints can be learned via examining many number of surgical cases.

In this work, we tried to create realistic deformation of structures; however, we believe that faithfully capturing such complex shape spaces is a highly challenging area of research, especially when there is limited number of sample structures available. We foresee that a more elaborate approach (e.g. based on elasticity properties or more relevant training samples) can lead to improved results.

Using non-optimized MATLAB code on a standard single core 3.40 GHz CPU, training and testing the random forest took about 52 and 2 seconds, respectively. Nonetheless, the training task can be parallelized to speed up the procedure. Also, it took less than 6 seconds for our algorithm to segment each stereo frame of an endoscopic video. We emphasize that the proposed method is highly parallelizable and GPU implementation for real-time video segmentation is possible in future work. In addition to GPU implementation, encoding depth information into the energy functional as well as using more advanced shape learning techniques (e.g. kernel PCA) are two important directions for future work. Experimenting on more clinical cases and in a more challenging situation (e.g. during kidney bleeding) as well as setting up an experiment to evaluate the correct alignment of internal structures quantitatively are other two important direction for expanding this work in future.

Acknowledgment

This work was partly supported by NPRP Grant #4-161-2-056 from the Qatar National Research Fund (a member of the Qatar Foundation) and partly by The Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] Luis Alvarez, Luis Gómez, and Pedro Henríquez. Zoom dependent lens distortion mathematical models. *Journal of Mathematical Imaging and Vision*, 44(3):480–490, 2012.

[2] Matthias Baumhauer, Marco Feuerstein, Hans-Peter Meinzer, and J Rassweiler. Navigation in endoscopic soft tissue surgery: perspectives and limitations. *Journal of Endourology*, 22(4):751–766, 2008.

[3] Rupin Dalvi and Rafeef Abugharbieh. Fast feature based multi slice to volume registration using phase congruency. In *IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 5390–5393. 2008.

[4] Samuel Dambreville, Romeil Sandhu, Anthony Yezzi, and Allen Tannenbaum. Robust 3D pose estimation and efficient 2D region-based segmentation from a 3D shape prior. In *European Conference on Computer Vision (ECCV)*, pages 169–182. 2008.

[5] Basanna V Dhandra, Ravindra Hegadi, Mallikarjun Hangarge, and Virendra S Malemath. Analysis of abnormality in endoscopic images using combined HSI color space and watershed segmentation. In *International Conference on Pattern Recognition (ICPR)*, volume 4, pages 695–698, 2006.

[6] Raúl San José Estépar, Carl-Fredrik Westin, and Kirby G Vosburgh. Towards real time 2D to 3D registration for ultrasound-guided endoscopic and laparoscopic procedures. *International Journal of Computer Assisted Radiology and Surgery*, 4(6):549–560, 2009.

[7] Enzo Ferrante and Nikos Paragios. Non-rigid 2D-3D medical image registration using Markov random fields. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 163–170. 2013.

[8] Isabel N Figueiredo, Pedro N Figueiredo, Georg Stadler, Omar Ghattas, and Aderito Araujo. Variational image segmentation for endoscopic human colonic aberrant crypt foci. *IEEE Transactions on Medical Imaging*, 29(4):998–1011, 2010.

[9] Isabel N Figueiredo, Juan Carlos Moreno, VB Surya Prasath, and Pedro N Figueiredo. A segmentation model and application to endoscopic images. In *Image Analysis and Recognition*, pages 164–171. 2012.

[10] Clive S Fraser and Salaheddin Al-Ajlouni. Zoom-dependent camera calibration in digital close-range photogrammetry. *Photogrammetric Engineering and Remote Sensing*, 72(9):1017, 2006.

[11] Inderbir S Gill, Louis R Kavoussi, Brian R Lane, Michael L Blute, Denise Babineau, J Roberto Colombo Jr, Igor Frank, Sompol Permpongkosol, Christopher J Weight, Jihad H Kaouk, et al. Comparison of 1,800 laparoscopic and open partial nephrectomies for single renal tumors. *The Journal of Urology*, 178(1):41–46, 2007.

[12] Sean Gill, Purang Abolmaesumi, Siddharth Vikal, Parvin Mousavi, and Gabor Fichtinger. Intraoperative prostate tracking with slice-to-volume registration in MR. In *International Conference of the Society for Medical Innovation and Technology*, pages 154–158, 2008.

[13] Ghassan Hamarneh, Preet Jassi, and Lisa Tang. Simulation of ground-truth validation data via physically-and statistically-based warps. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 459–467. 2008.

[14] Tobias Heimann, Peter Mountney, Matthias John, and Razvan Ionasec. Learning without labeling: Domain adaptation for ultrasound transducer localization. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 49–56. Springer, 2013.

[15] Nagelhus Hernes, A Toril, Frank Lindseth, Tormod Selbekk, Arild Wollf, Ole Vegard Solberg, Erik Harg, Ola M Rygh, Geir Arne Tangen, Inge Rasmussen, et al. Computer-assisted 3D ultrasound-guided neurosurgery: technological contributions, including multimodal registration and advanced display, demonstrating future perspectives. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 2(1):45–59, 2006.

[16] Johann Hummel, Michael Figl, Michael Bax, Helmar Bergmann, and Wolfgang Birkfellner. 2D/3D registration of endoscopic ultrasound to CT volume data. *Physics in Medicine and Biology*, 53(16):4303, 2008.

[17] Tung-Ying Lee, Tzu-Shan Chang, Chen-Hao Wei, Shang-Hong Lai, Kai-Che Liu, and Hurng-Sheng Wu. Automatic distortion correction of endoscopic images captured with wide-angle zoom lens. *IEEE Transactions on Biomedical Engineering*, 60(9):2603–2613, 2013.

[18] Scott A Merritt, Lav Rai, and William E Higgins. Real-time CT-video registration for continuous endoscopic guidance. In *Medical Imaging*, pages 614313–614313, 2006.

[19] Philip W Mewes, Dominik Neumann, Oleg Licegevic, Johannes Simon, Aleksandar Lj Juloski, and Elli Angelopoulou. Automatic region-of-interest segmentation and pathology detection in magnetically guided capsule endoscopy. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 141–148. 2011.

[20] Daniel J Mirota, Masaru Ishii, and Gregory D Hager. Vision-based navigation in image-guided interventions. *Annual review of biomedical engineering*, 13:297–319, 2011.

[21] Stéphane Nicolau, Luc Soler, Didier Mutter, and Jacques Marescaux. Augmented reality in laparoscopic surgical oncology. *Surgical oncology*, 20(3):189–201, 2011.

[22] Masoud S Nosrati, Jean-Marc Peyrat, Julien Abinahed, Osama Al-Alao, Abdulla Al-Ansari, Rafeef Abugharbieh, and Ghassan Hamarneh. Efficient multi-organ segmentation in multi-view endoscopic videos using pre-operative priors. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 324–331. 2014.

[23] Sergey Osechinskiy and Frithjof Kruggel. Slice-to-volume nonrigid registration of histological sections to MR images of the human brain. *Anatomy Research International*, 2011, 2010.

[24] Mark R Pickering, Abdullah A Muhit, Jennie M Scarvell, and Paul N Smith. A new multi-modal similarity measure for fast gradient-based 2D-3D image registration. In *IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5821–5824, 2009.

[25] Philip Pratt, Erik Mayer, Justin Vale, Daniel Cohen, Eddie Edwards, Ara Darzi, and Guang-Zhong Yang. An effective visualisation and registration system for image-guided robotic partial nephrectomy. *Journal of Robotic Surgery*, 6(1):23–31, 2012.

[26] Victor A Prisacariu and Ian D Reid. PWP3D: Real-time segmentation and tracking of 3D objects. *International Journal of Computer Vision*, 98(3):335–354, 2012.

[27] Victor Adrian Prisacariu, Aleksandr V Segal, and Ian Reid. Simultaneous monocular 2D segmentation, 3D pose recovery and 3D reconstruction. In *Asian Conference on Computer Vision (ACCV)*, pages 593–606. 2013.

[28] Gustavo A Puerto-Souza and Gian Luca Mariottini. Toward long-term and accurate augmented-reality display for minimally-invasive surgery. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5384–5389, 2013.

[29] Romeil Sandhu, Samuel Dambreville, Anthony Yezzi, and Allen Tannenbaum. A nonrigid kernel-based framework for 2D-3D pose estimation and 2D image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1098–1115, 2011.

[30] Babbage Science and technology. Surgical robots: The kindness of strangers. *The Economist Online*, 2012-01-18.

[31] Li-Ming Su, Balazs P Vagvolgyi, Rahul Agarwal, Carol E Reiley, Russell H Taylor, and Gregory D Hager. Augmented reality during robot-assisted laparoscopic partial nephrectomy: toward real-time 3D-CT to stereoscopic video registration. *Urology*, 73(4):896–900, 2009.

[32] Intuitive Surgical. da vinci products FAQ. 2013-09-7.

[33] Gerald Y Tan, Raj K Goel, Jihad H Kaouk, and Ashutosh K Tewari. Technological advances in robotic-assisted laparoscopic surgery. *Urologic Clinics of North America*, 36(2):237–249, 2009.

[34] Dogu Teber, Selcuk Guven, Tobias Simpfendörfer, Mathias Baumhauer, Esref Oguz Güven, Faruk Yencilek, Ali Serdar Gözen, and Jens Rassweiler. Augmented reality: a new tool to improve surgical accuracy during laparoscopic partial nephrectomy? preliminary in vitro and in vivo results. *European urology*, 56(2):332–338, 2009.

[35] Andrew Top, Ghassan Hamarneh, and Rafeef Abugharbieh. Active learning for interactive 3D image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 603–610. 2011.

[36] Annegreet van Opbroek, M Arfan Ikram, Meike W Vernooij, and Marleen de Bruijne. A transfer-learning approach to image segmentation across scanners by maximizing distribution similarity. In *Machine Learning in Medical Imaging*, pages 49–56. 2013.

[37] Luminita A Vese and Tony F Chan. A multiphase level set framework for image segmentation using the Mumford and Shah model. *International Journal of Computer Vision*, 50(3):271–293, 2002.

[38] Anthony G Wiley and Kam W Wong. Geometric calibration of zoom lenses for computer vision metrology. *Photogrammetric Engineering and Remote Sensing*, 61(1):69–74, 1995.

[39] Yeny Yim, Mike Wakid, Can Kirmizibayrak, Steven Bielamowicz, and James K Hahn. Registration of 3D CT data to 2D endoscopic image using a gradient mutual information based viewpoint matching for image-guided medialization laryngoplasty. *Journal of Computing Science and Engineering*, 4(4):368–387, 2010.

[40] Darko Zikic, Ben Glocker, Oliver Kutter, Martin Groher, Nikos Komodakis, Ali Khamene, Nikos Paragios, and Nassir Navab. Markov random field optimization for intensity-based 2D-3D registration. In *SPIE Medical Imaging*, pages 762334–762334, 2010.