IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 13, 2020, accepted March 5, 2020, date of publication March 11, 2020, date of current version March 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980025

Automatic Segmentation of Multiple Structures in Knee Arthroscopy Using Deep Learning

YAQUB JONMOHAMADI[1], YU TAKEDA[2], FENGBEI LIU[3], FUMIO SASAZAWA[4], GABRIEL MAICAS[3], ROSS CRAWFORD[5], JONATHAN ROBERTS[1], (Senior Member, IEEE), AJAY K. PANDEY[1], AND GUSTAVO CARNEIRO[3]

[1]School of Electrical Engineering and Robotics, Science and Engineering Faculty, Queensland University of Technology, Brisbane, QLD 4000, Australia
[2]Department of Orthopaedic Surgery, Hyogo College of Medicine, Nishinomiya 663-8501, Japan
[3]School of Computer Science, Australian Institute for Machine Learning, The University of Adelaide, Adelaide, SA 5005, Australia
[4]Faculty of Medicine and Graduate School of Medicine, Hokkaido University, Sapporo 060-0808, Japan
[5]Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, QLD 4000, Australia

Corresponding authors: Yaqub Jonmohamadi (y.jonmo@qut.edu.au) and Ajay K. Pandey (a2.pandey@qut.edu.au)

This work was supported by the Australia India Strategic Research Fund (Project AISRF53820) and in part by the Australian Research Council through under Grant DP180103232. The cadaver studies is covered by the Queensland University of Technology Ethics Approval under project1400000856.

ABSTRACT Minimally invasive surgery (MIS) is among the preferred procedures for treating a number of ailments as patients benefit from fast recovery and reduced blood loss. The trade-off is that surgeons lose direct visual contact with the surgical site and have limited intra-operative imaging techniques for real-time feedback. Computer vision methods as well as segmentation and tracking of the tissues and tools in the video frames, are increasingly being adopted to MIS to alleviate such limitations. So far, most of the advances in MIS have been focused on laparoscopic applications, with scarce literature on knee arthroscopy. Here for the first time, we propose a new method for the automatic segmentation of multiple tissue structures for knee arthroscopy. The training data of 3868 images were collected from 4 cadaver experiments, 5 knees, and manually contoured by two clinicians into four classes: Femur, Anterior Cruciate Ligament (ACL), Tibia, and Meniscus. Our approach adapts the U-net and the U-net++ architectures for this segmentation task. Using the cross-validation experiment, the mean Dice similarity coefficients for Femur, Tibia, ACL, and Meniscus are 0.78, 0.50, 0.41, 0.43 using the U-net and 0.79, 0.50, 0.51, 0.48 using the U-net++. While the reported segmentation method is of great applicability in terms of contextual awareness for the surgical team, it can also be used for medical robotic applications such as SLAM and depth mapping.

INDEX TERMS Arthroscopy, artificial intelligence, auto segmentation, deep learning, endoscopy, surgery.

I. INTRODUCTION

Unlike open surgery, which involves cutting multiple tissue layers to access the surgical area of interest inside the human body, Minimally invasive surgery (MIS) is conducted via small incisions to reduce surgical trauma and post-operation recovery time. Significant progress has been achieved to make MIS safer and more accurate for better patient outcomes. Despite increasing demand for MIS, there are some common drawbacks, namely: limited access to the operating space, reduced field of view (FoV), the lack of haptic feedback, diminished hand-eye coordination, and prolonged learning curves and training periods. This leads to extended operation times and increased cost to patients [1].

It is expected that the ability to automatically segment and label tissues present in the camera view, similar to what happens in preoperative CT or MRI images [2], can simplify the long learning curve associated with MIS [3]. In arthroscopy, unlike laparoscopy, the tissues are located typically very close to the camera (e.g., 10 mm away), resulting in only a fraction of joint structures appearing in the camera FoV. Hence, quite often surgeons fail to identify tissue structures and recourse to visual surveying: moving the camera around to identify and gain tissue awareness. In fact, for knee arthroscopy, given a video frame, the clinician can only identify Femur with confidence, while other structures, such as Meniscus, Tibia, ACL, and nonstructural tissues (such as fat) remain

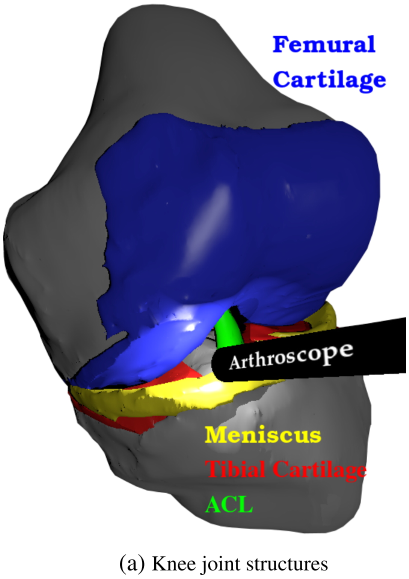
The associate editor coordinating the review of this manuscript and approving it for publication was Hazrat Ali.

a challenge for them. This phenomenon happens repeatedly during surgery, which could prolong the operation time and lead to unintentional damage to critical tissues due to the excessive and untracked camera movements. This justifies the need for an automatic segmentation and tissue labeling approach, which could aide surgeons by providing contextual awareness of the surgical scene, reduce the operation time, and reduce the long learning curve usually associated with training to become a surgeon.

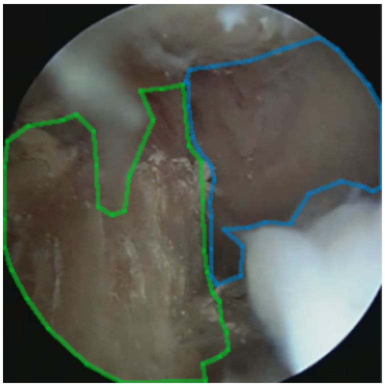
Research on computer vision and machine learning techniques for improving MIS are on the rise and real-time segmentation and localization of tissue and surgical tools in 3D is the main focus area. There have been three main approaches to this: 1) segmentation of the video frames into different sections using active-contour based methods [4] or parameter sensitive morphological operations and thresholding approaches [5]; 2) 3D segmentation by incorporating the preoperative images such as computational tomography (CT) onto the intra-operative data of endoscope [6]–[8] or stereo endoscope [9]–[14] – the registration of the CT images into the endoscopic frames is achieved either manually using the known landmarks in both modalities or automatically [13]–[15]; 3) estimation of the 3D surface of the surgical scene using structure from point clouds and segmentation of the 3D surface into subsections using the geometrical discontinuities and structural cues [16].

Methods relying on color intensity and texture features are challenging due to the level of noise present in image-guided robotic surgery [12], whereas 3D segmentation using CT or stereo imaging have been remarkably successful in laparoscopy. In the case of arthroscopy, to the best of our knowledge, none of the approaches mentioned above have been explored. Key landmarks present inside the knee structure including Femur, ACL, Meniscus, and Tibia are shown in Fig. 1-a. Fig. 1-b shows a sample endoscopic frame, where the small FoV displays a small portion of the ACL and the Femur and at the same time the view is partially obscured by floating nonstructural tissues, such as the fat shown in the bottom right corner of the frame.

Compared to laparoscopy, there are three main limitations in arthroscopy, which makes registration between pre and intra-operative images challenging: A) the anatomical construct of the knee is as such that the gap between bone joint under flexon is usually less than 10 mm, this makes the available FoV much smaller than laparoscopy. Therefore, arthroscopic frames capture only a small portion of the total joint structure and do not provide large enough surfaces, as desired for the registration process; B) the reduction of the field of view caused from nonstructural tissues, such as fat – these tissues are created either during the incision process or are a result of the damaged/degraded joint structures; and C) during knee arthroscopy, surgeons require to visualise the knee joint at different flexion, making the surgical area a non-rigid space – key structures present in the knee cavity change their position with respect to each other and appear differently under flexion. As a result of these limitations, the application of computer vision methods in arthroscopy is challenging. Therefore it is not surprising that medical robotics and computer vision literature on MIS for knee arthroscopy is quite limited.



(a) Knee joint structures



(b) Sample arthroscopic frame

FIGURE 1. Figure (a) shows left knee joint structures including ACL in green, Meniscus in yellow, the Tibial cartilage in red, and the Femoral cartilage in blue. Our paper aims to automatically segment Femoral cartilage, ACL, Meniscus, and Tibia. A typical knee arthroscopic frame is shown in (b), where a small part of the ACL (contoured in green) and the Femur (contoured in blue) are visible. The floating white tissues are blocking the view on the bottom right and top left corners of the frame.

In recent years, our research group at the Queensland University of Technology (QUT) has made significant progress to circumvent these limitations [17]–[19]. Efforts to use Simultaneous Localization and Mapping (SLAM) have recently been applied to arthroscopy, but the extraction of key landmark features from images still remains an open challenge. Other approaches include applying Augmented Reality (AR) to arthroscopy, as reported in [20]. However, this was limited to the projection of an expert surgeon hand onto the screen of the other surgeons. Similarly, the AR for wrist arthroscopy described in [3] was to depict surgical tools in the 3D space by tracking them using an electromagnetic localization system.

It is important to highlight that literature precedence on the segmentation of knee joint structures for contextual awareness or to be able to isolate features of interest is largely unavailable.

During the past 5 years, the use of deep neural networks has revolutionized computer vision [21]. Deep learning refers to a composition of many simple functions parameterized by variables [22], [23], trained by stochastic gradient descent which can be computed using the back propagation procedure [24], [25].

The well-known convolutional neural networks (CNN) represent an effective type of neural networks due to learning the hierarchical feature representations of the image in a purely data-driven manner. This means that the features that are good for classification are learned from the images [26], [27]. This approach of learning has been extensively applied to biomedical imaging applications [28]–[30], as well as classification purposes in endoscopy [31] and colonoscopy [32]. The fully CNN (FCNN) introduced by Long *et al.* [33] provided the opportunity of pixel-wise classification and has been applied for the semantic segmentation of medical images [21], [30], [34]–[36]. In laparoscopy, the FCNN has been successfully applied for the segmentation of liver [36]–[39] and surgical tools [40]–[43]. Aside from the methodological breakthroughs in biomedical image analysis, advances in safe/secure data acquisition have also been reported in the literature [44].

In this work, we report a fully automatic approach for tissue segmentation from knee arthroscopy video. More specifically, we propose an instance-based segmentation method that can automatically segment Femur, Anterior Cruciate Ligament (ACL), Tibia, and Meniscus. We relied on a training data comprising 3868 images collected from 4 cadaver experiments, five knees, and manually annotated by two clinicians into the four classes mentioned above. Our approach consists of an adaptation of the U-net [45] and the U-net++ [46]. Using a cross validation experiment, the mean Dice coefficients for Femur, Tibia, ACL, and Meniscus are 0.78, 0.50, 0.41, 0.43 using the U-net and 0.79, 0.50, 0.51, 0.48 using the U-net++. This method represents the first step to improve our previously proposed medical robotic SLAM and depth mapping methods [17], [18].

II. METHODS

Throughout this manuscript, the italic lower or upper cases refers to scalars (z *and* Z), bold italic lower case refers to vectors ($\boldsymbol{z}$), and bold italic upper case refers to matrices ($\boldsymbol{Z}$).

A. U-NET AND U-NET++

Among the FCNN models, the U-net [45] is a well-known model which was proposed for segmentation of biomedical images and have shown to work well with small data sets. Its distinct architecture includes skip connections from the encoders (which extract information from the input images) to decoders (which project the information into the image space), and concatenation and deconvolution of these features to obtain up-sampled features map. A more recent variant of U-net, called U-net++ was proposed in [46], where the skip connections are replaced by nested and dense skip connections to create a more powerful architecture. The motivation for this modification was to decrease the semantic gap of the feature maps between the encoder and decoder. In our work, we assess the performance of both approaches for the fully automatic tissue segmentation from knee arthroscopy video. The simplified diagrams of the two architectures are shown in Fig. 2. In the encoder part of the two networks, padded 3 × 3 convolution and 2 × 2 max pooling (stride 2) with 64 initial features were used. For the decoding path, 2 × 2 upsampling the features and 3 × 3 convolution were used. The data activation within layers was done using rectified linear unit method (ReLU) [47], defined as

$$R(z) = max(0, z). \tag{1}$$

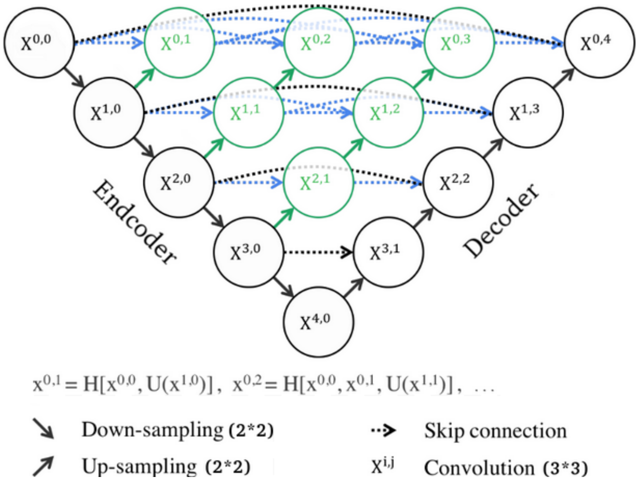


FIGURE 2. The simplified diagram of U-net and U-net++. The U-net++ includes all the colors whereas the U-net is represented by the blocks in black. The blue and green colors refer to the skip connections. The image is modified and from [46].

The sigmoid activation for the final segmentation layer was used for both networks, and is defined as

$$\sigma(z) = 1/(1 + e^{z}). \tag{2}$$

It is known that the semantic segmentation is not accurate enough to separate multiple classes of objects when the boundaries are fine and closely packed [43] or when multiple objects of the same class are close to each other [48]. Both scenarios are typical surgical scenes, and we aim to perform multi-class instance segmentation, similarly to [48].

B. DATA ACQUISITION

The arthroscopy data set used in this study was obtained from four cadaveric experiments performed at our Medical and Engineering Research Facility and it consists of three females and one male. The data set comprises the left knee from the female cadavers, and both left and right knees from the male cadaver. In each case, two incision points were made at the

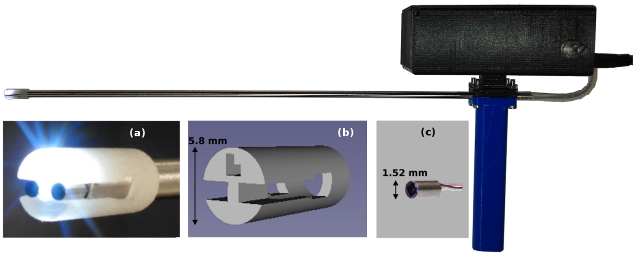


FIGURE 3. The custom built camera prototype. Figure (a) shows the closeup view at the tip (b) and the 3D design (c) and the muC103A camera. The endoscope circuits and extra wiring are contained inside the black box and connected to the computer using two USB cables.

bottom left and bottom right. Video sequences were recorded using two types of arthroscopes: Stryker endoscope (4.0 mm diameter) and a custom built stereo arthroscope based on muc103 camera module (6 mm diameter). The resolution of the Stryker camera was 1280 × 720 and field of view (FoV) 30 degrees, whereas the custom camera had a 384 × 384 resolution and FoV of 87.5 degrees. The Stryker arthroscope had a circular field of view with a diameter of approximately 800 pixels. Hence, the video frames from this camera were cropped into 720 × 720 frames and down-sampled to 384 × 384 to have the same dimension as the custom built camera. The training images were obtained by frame grabbing of the video sequences recorded by the two cameras every two seconds. The custom built camera prototype is shown in Fig. 3, which is comprised of two muC103A cameras together with their C8262 UVC interface modules, a white LED (T0402W) for illumination all housed in a custom built 3D printed camera head for mounting cameras and the LED. All electronics were placed inside a 3D printed box at the far end of the insertion tube. The advantage of the custom built camera over the commercial endoscopes is the broader field of view as well as the stereo vision (not discussed here). The diameter of the camera tip is 5.8 mm, which is 2.52 mm smaller than the smallest commercially available stereo endoscope by da Vinci.

The contrast-limited adaptive histogram equalization was then applied on the RGB channel individually to improve the contrast of the images. Ground truth was obtained using manual contouring of the training images by two clinicians using the MevisLab toolbox [49]. Both clinicians had access to the original video sequences from which the frames were extracted. This was required because it is challenging for the clinicians to identify key structures captured in a single frame, as most of the time only a small part of the knee structure is visible. Overall, 3868 images were manually contoured.

C. LIMITATIONS

TABLE 1. The table provides statistical information on the number of the images from each cadaveric.

Structure	Femur	ACL	Tibia	Meniscus	Number of
Cadaver knee					images
1	40%	0%	7%	0%	99
2	32%	20%	5%	9%	1043
3	30%	14%	8%	10%	1768
4-left	47%	3%	4%	6%	459
4-right	33%	8%	9%	12%	489
total	33%	13%	7%	9%	3868

Not all the cadaver experiments were conducted in a single session, so discrepancies associated with access to key landmarks and changes to lighting conditions had some effect on the quality of images used in this study. For instance, according to the needs of surgical flow, the illumination during different experiments was provided from different incision points, and later experiments relied on in-built LED as the lighting source (LED T0402W). This means that in the endoscopic videos recorded from different cadaveric experiments, structures had a slightly different color temperature due to the type of illumination used. The angle of the lighting with respect to the camera was not controlled. We first focused on getting as many representative images as possible for the key landmark feature of the knee cavity, i.e the Femur. Therefore the Femur is over represented in our image data set. The data set also captured age and gender related discrepancies associated with the knee anatomy. Both male and female cadavers were included in this study with age ranging from 56-93 years. Some of the videos captured had a severely degenerated form of Femur. The statistical specification of the data in terms of the proportion of the structures observed in each of the cadavers is shown in Table 1.

III. EXPERIMENTAL SETUP

In this section, we describe the hyper parameters and cost functions applied for the U-net and U-net++, the training, and the evaluation approach.

A. HYPER PARAMETERS

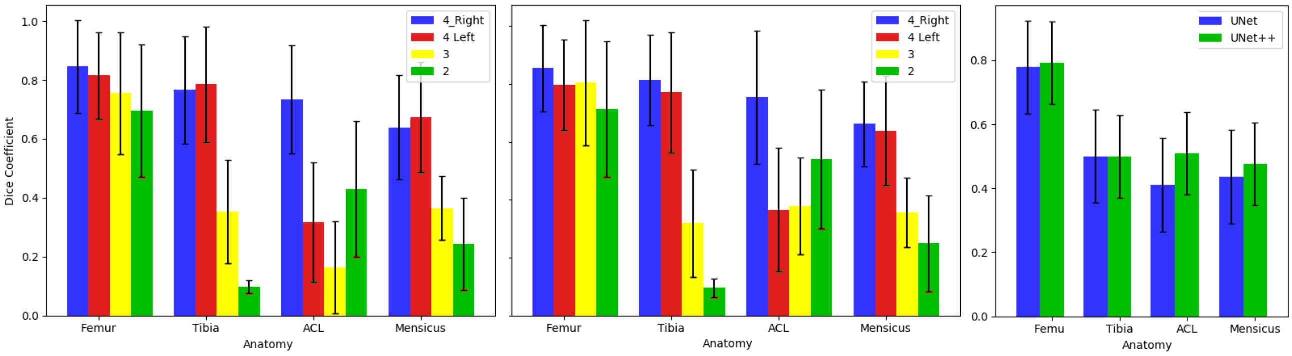
The main model used by the encoder is the Imagenet [50] pre-trained RestNet-34 [51]. The 5 layers of the U-net and U-net++ had 64, 64, 128, 256, and 512 filters. For the training of the networks, the optimization function combined the Dice coefficient loss [52] with the cross entropy loss, as suggested in [46]:

$$\mathrm{DiceCCE}(T,P) = 0.5 \times \mathrm{CCE}(T,P) + (1 - \mathrm{Dice}(T,P)), \tag{3}$$

where the Dice coefficient Dice(T,P) is defined as:

$$\mathrm{Dice}(T,P) = (2 \times \sum_{i}^{N} T_i P_i + \mathit{Smooth})/(\sum_{i}^{N} T_i + \sum_{i}^{N} P_i + \mathit{Smooth}), \tag{4}$$

were the constant *Smooth* is set to 1, matrices T and P are the ground truth and the model prediction, respectively, N is the number of the pixels. The categorical cross CCE(T,P)



(a) U-net

(b) U-net++

(c) Average

FIGURE 4. Figure (a) and (b) represent the U-net and U-net++ mean and standard deviation (STD) of segmentation accuracy measured by Dice similarity coefficient on four knee structures of Femur, Tibia, ACL, and Meniscus on four validation data sets of cadaver 2 (green bar), 3 (yellow bar), 4-left knee (red bar), and 4-right knee (blue bar). Figure (c) shows the average of the Dice similarity coefficients on four-fold cross validation data sets, where the blue bar corresponds to Figure (a) (U-net) and green bar corresponds to Figure (b) (U-net++).

entropy in equation 3 is defined as:

$$\mathrm{CCE}(T, P) = (1/N)\sum_{i}^{N}\sum_{j}^{C} T_{ij} log(P_{ij}), \tag{5}$$

where C is the number of categories, which is 4 here. For the validation of the training and the test results, the Dice similarity coefficients are reported. The networks were implemented in Pytorch [53]. The data discrepancy mentioned in the Section II-C resulted in an unbalanced data set where certain segmentation classes dominated the training label distribution. For instance, the occurrence of the label Femur is substantially larger than the other structures because it is a dominant structure in the images. To address this problem, several previous studies relied on 'class weighting', in which the loss function of a particular sample is weighted by the inverse of the proportion of the label of that sample (for example [54]). We tried several variations of such class weighting, but results were not improved, and hence the class weighting was not used.

B. TRAINING AND EVALUATION

Our results are computed based on a four-fold cross-validation experiment, where the validation images were selected from one cadaver, while images from other cadavers were used for training. The images from cadaver 1 were not used for the validation but for training only due to the small number of images available from this cadaver and the fact the vast majority of images contained mostly the Femur. In this way, validation data sets were formed from cadavers 2, 3, 4 left knee, and 4 right knee. The training images were shuffled and then randomly selected using batches of 16 images for stochastic gradient descent. The augmentation pipeline provided by [38] involves randomly flip in horizontal and vertical directions, randomly produce brightness contrast changes and non-rigid transformation including elastic transformation and optical distortion. The training images were also randomly cropped to 256 × 256 from the original images of 384 × 384. For the optimization, we used Adam optimizer [55] with starting learning rate of 1e-4 and weight decay 1e-5. The number of training epochs was 70 and we used a polynomial learning rate with factor of 0.9. In the case of U-net, every epoch took 4 minutes and 10s to process, whereas U-net++ took 9 minutes and 30s. The model was trained on an NVIDIA Tesla M40.

IV. RESULTS AND DISCUSSION

Fig. 4 shows the quantitative results of the four fold cross validation. While Fig. 4-a and -b show the mean and standard deviation (STD) of Dice similarity coefficients for U-net and U-net++ on each data set separately, Fig. 4-c is the average of the Dice coefficients results of the four data sets, i.e., blue bars in Fig. 4-c are the mean value of the Fig. 4-a and green bars are the mean value of Fig. 4-b. Analysing the results, Femur produces the highest Dice similarity in all test sets, while other structures have different results in each test set. By comparing the test results, U-net++ shows slightly higher accuracy in most test sets. According to Fig. 4-a and -b, the highest accuracy was achieved on data from cadaver 4-right (blue bars) followed by 4-left (red bars), whereas the worst accuracy was achieved on data from cadaver 2 (green bars). The segmentation of Femur was consistently achieved with high accuracy with the lowest dice coefficient being 0.64 on cadaver 2.

The qualitative segmentation results are shown in Fig. 5. Fig. 5-a shows a common scene where three structures of the Femur, Tibia, and Meniscus are visible in one frame. Since the illumination for cadaver 2 was provided from another incision and was different from the LED used for later experiments, the difference in the coloring of the scene is visible in the first row of Fig. 5-a and -b compared with the three images of the other cadavers in the next rows. Fig. 5-b shows the images, where ACL is the main structure visible in the frame. As it is clear, the shape of the ACL could change substantially due to the angle of the camera and the angle of the knee joint. While Fig. 5-a and -b are depicting images obtained from the

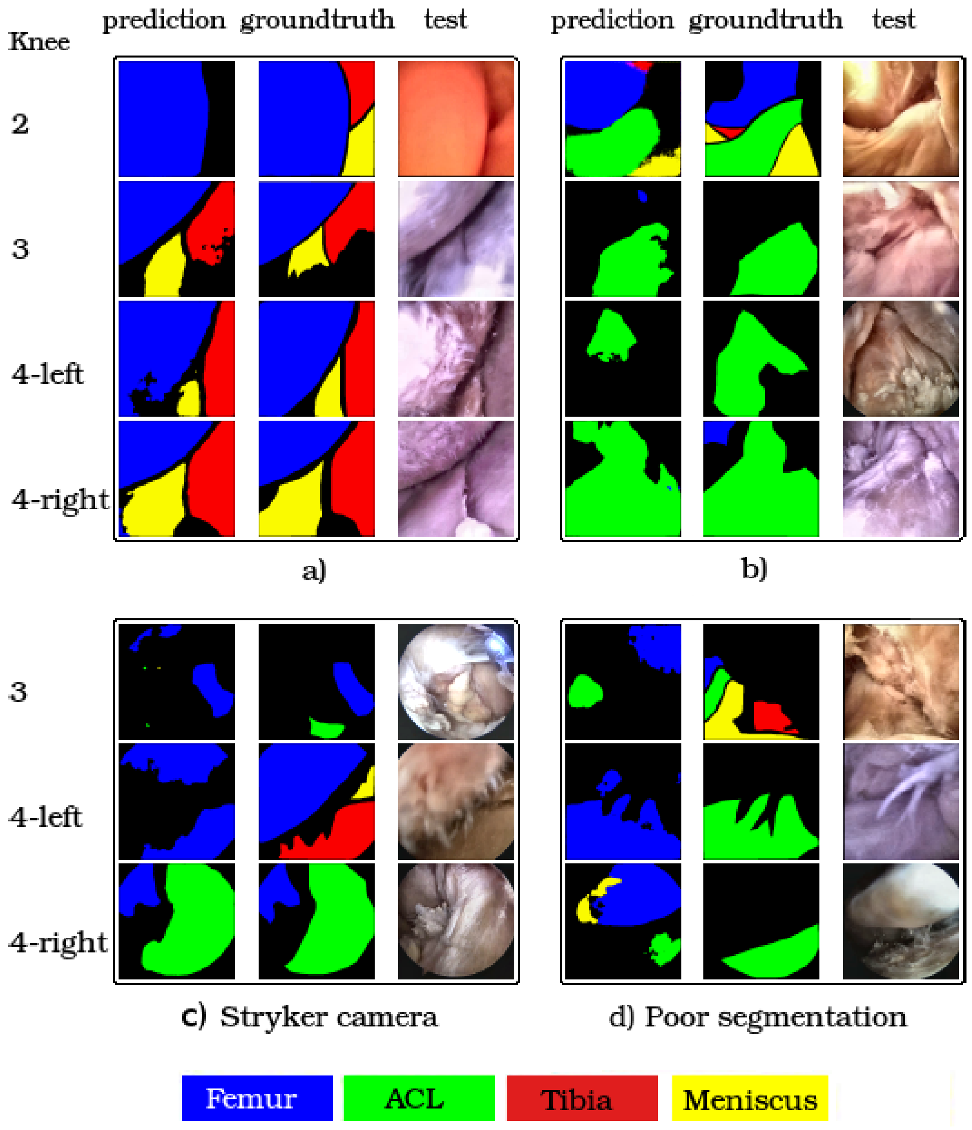


FIGURE 5. Qualitative results for the multi-structure segmentation of the arthroscopic video frames. The numbers on the left-hand side refer to the cadaver knee index. Figure a) shows a typical Femur, Tibia, and Meniscus combination during the knee arthroscopy. The ACL is shown in Figure b) as the main structure. Figures a) and b) are from the custom build camera, whereas Figure c) is obtained from the Stryker arthroscope. In the first row of Figure c) the tip of the custom build camera is visible. Figures a to c show good to moderate segmentation results, whereas Figure d) shows samples of poor segmentation.

custom-built camera, Fig. 5-c shows samples of the images using the Stryker arthroscope. In the first row of the Fig. 5-c, the tip of the custom build camera is present in the frame. In the second row of the Fig. 5-c, the highly degenerated Femur of the 92-year old cadaver is clearly visible. The same Femur is also visible in the 3rd row of Fig. 5-a. This degeneration changes the appearance of the left Femur of this cadaver lin comparison with a normal Femur. The right knee Femur is less degenerated (Fig. 5-a, 4th row). Examples of poor segmentation performance are illustrated in Fig. 5-d. Two examples of floating tissue are visible in the second row of Fig. 5-a and the second row of Fig. 5-d. In the latter case, although the floating tissue is successfully removed from the image, the network has failed to distinguish ACL from Femur.

V. CONCLUSION

In this paper, we propose the first automatic segmentation method of key structures present in the knee cavity. Using trained fully convolutional neural networks, we successfully segmented knee arthroscopic video frames into four structures: Femur, ACL, Tibia, and Meniscus. There are two possible uses for this type of segmentation. Firstly, the automatic segmentation of the arthroscopic frames provides contextual awareness for the surgeons and could be used for clinical training intra-operatively. Secondly, it can be used for medical robotics for tissue and tool tracking in full 3D. The U-net and U-net++ architectures were used as baseline models, and results indicate that the U-net++ had marginally higher accuracy for the segmentation results. This, however, comes with the cost of the training time of U-net++ being twice that of the U-net.

Dice similarity results on the four-fold cross validation experiment indicate that Femur was consistently segmented with high accuracy compared with the other three structures. Part of the reason for this is that the data was imbalanced and the Femur comprised about a third of the total pixels in the training data. Moreover, the Femur has a rather distinct spherical shape which is easily distinguishable from other structures and it has been shown that the dice scores vary because of tissue geometry alone [56]. Recognition of the other landmark tissue would become better with a larger representation and this forms part of our future research.

AUTHOR CONTRIBUTIONS STATEMENT

Y.J. contributed to data collection, data preprocessing, initial training of deep learner, and manuscript draft preparation. Y.T. and FS contributed by performing cadaveric experiments and manual contouring of training images. F.L. and G.M. contributed in further optimization of the deep learning approach developed by Y.J.. R.C. and J.R. were involved in cadaveric experiment planning and ethic approval. A.P. and G.C. conceptualized the idea and supervised the project. All authors have read and commented on the manuscript.

REFERENCES

[1] R. Smith, A. Day, T. Rockall, K. Ballard, M. Bailey, and I. Jourdan, "Advanced stereoscopic projection technology significantly improves novice performance of minimally invasive surgical skills," *Surgical Endoscopy*, vol. 26, no. 6, pp. 1522–1527, Jun. 2012.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[3] B. Münzer, K. Schoeffmann, and L. Böszörmenyi, "Content-based processing and analysis of endoscopic images and videos: A survey," *Multimedia Tools Appl.*, vol. 77, no. 1, pp. 1323–1362, Jan. 2018.

[4] I. N. Figueiredo, P. N. Figueiredo, G. Stadler, O. Ghattas, and A. Araujo, "Variational image segmentation for endoscopic human colonic aberrant crypt foci," *IEEE Trans. Med. Imag.*, vol. 29, no. 4, pp. 998–1011, Apr. 2010.

[5] B. V. Dhandra, R. Hegadi, M. Hangarge, and V. S. Malemath, "Analysis of abnormality in endoscopic images using combined HSI color space and watershed segmentation," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2006, pp. 695–698.

[6] D. J. Mirota, M. Ishii, and G. D. Hager, "Vision-based navigation in image-guided interventions," *Annu. Rev. Biomed. Eng.*, vol. 13, no. 1, pp. 297–319, Aug. 2011.

[7] S. Nicolau, L. Soler, D. Mutter, and J. Marescaux, "Augmented reality in laparoscopic surgical oncology," *Surgical Oncol.*, vol. 20, no. 3, pp. 189–201, Sep. 2011.

[8] M. Baumhauer, M. Feuerstein, H.-P. Meinzer, and J. Rassweiler, "Navigation in endoscopic soft tissue surgery: Perspectives and limitations," *J. Endourol.*, vol. 22, no. 4, pp. 751–766, Apr. 2008.

[9] G. Y. Tan, R. K. Goel, J. H. Kaouk, and A. K. Tewari, "Technological advances in robotic-assisted laparoscopic surgery," *Urologic Clinics North Amer.*, vol. 36, no. 2, pp. 237–249, May 2009.

[10] P. Pratt, E. Mayer, J. Vale, D. Cohen, E. Edwards, A. Darzi, and G.-Z. Yang, "An effective visualisation and registration system for image-guided robotic partial nephrectomy," *J. Robot. Surg.*, vol. 6, no. 1, pp. 23–31, Mar. 2012.

[11] M. S. Nosrati, J.-M. Peyrat, J. Abinahed, O. Al-Alao, A. Al-Ansari, R. Abugharbieh, and G. Hamarneh, "Efficient multi-organ segmentation in multi-view endoscopic videos using pre-operative priors," *Med. Image Comput. Comput. Assist. Interv.*, vol. 17, no. 2, pp. 324–331, 2014.

[12] M. S. Nosrati, R. Abugharbieh, J.-M. Peyrat, J. Abinahed, O. Al-Alao, A. Al-Ansari, and G. Hamarneh, "Simultaneous multi-structure segmentation and 3D nonrigid pose estimation in image-guided robotic surgery," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 1–12, Jan. 2016.

[13] D. Stoyanov, M. V. Scarzanella, P. Pratt, and G.-Z. Yang, "Real-time stereo reconstruction in robotically assisted minimally invasive surgery," *Med. Image Comput. Comput. Assist. Interv.*, vol. 13, no. 1, pp. 275–282, 2010.

[14] S. Röhl, S. Bodenstedt, S. Suwelack, H. Kenngott, B. P. Müller-Stich, R. Dillmann, and S. Speidel, "Dense GPU-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration," *Med. Phys.*, vol. 39, no. 3, pp. 1632–1645, Mar. 2012.

[15] S. A. Merritt, L. Rai, and W. E. Higgins, "Real-time CT-video registration for continuous endoscopic guidance," *Proc. SPIE*, vol. 6143, Mar. 2006, Art. no. 614313.

[16] N. Haouchine and S. Cotin, "Segmentation and labelling of intra-operative laparoscopic images using structure from point cloud," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 115–118.

[17] A. Marmol, P. Corke, and T. Peynot, "ArthroSLAM: Multi-sensor robust visual localization for minimally invasive orthopedic surgery," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 3882–3889.

[18] A. Marmol, A. Banach, and T. Peynot, "Dense-ArthroSLAM: Dense intra-articular 3-D reconstruction with robust localization prior for arthroscopy," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 918–925, Apr. 2019.

[19] L. Wu, A. Jaiprakash, A. Pandey, D. Fontanarosa, Y. Jonmohamadi, M. Antico, M. Strydom, A. Razjigaev, F. Sasazawa, J. Roberts, and R. Crawford, "Robotic and image-guided knee arthroscopy," in *Handbook of Robotic and Image-Guided Surgery*. Amsterdam, The Netherlands: Elsevier, 2020. [Online]. Available: https://scholar.google.com/scholar?um=1&ie=UTF-8&lr&cites=3926880190490242806

[20] S. K. Baker, C. T. Fryberger, and B. A. Ponce, "The emergence of augmented reality in orthopaedic surgery and education," *Orthopaedic J.*, vol. 16, pp. 8–16, Jun. 2015.

[21] C. M. Deniz, S. Xiang, R. S. Hallyburton, A. Welbeck, J. S. Babb, S. Honig, K. Cho, and G. Chang, "Segmentation of the proximal femur from MR images using deep convolutional neural networks," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 16485.

[22] F. Unglaub and C. Spies, "Augmented reality-based navigation system for wrist arthroscopy: Feasibility," *J. Wrist Surg.*, vol. 3, no. 1, p. 66, Feb. 2014.

[23] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat, D. C. Barratt, S. Ourselin, M. J. Cardoso, and T. Vercauteren, "NiftyNet: A deep-learning platform for medical imaging," *Comput. Methods Programs Biomed.*, vol. 158, pp. 113–122, May 2018.

[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[25] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[26] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.

[27] H. R. Roth, L. Lu, A. Farag, A. Sohn, and R. M. Summers, "Spatial aggregation of Holistically-Nested networks for automated pancreas segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2016, pp. 451–459.

[28] A. A. Cruz-Roa, J. E. A. Ovalle, A. Madabhushi, and F. A. G. Osorio, "A deep learning architecture for image representation, visual interpretability and automated Basal-Cell carcinoma cancer detection," in *Proc. Adv. Inf. Syst. Eng.*, 2013, pp. 403–410.

[29] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2014, pp. 520–527.

[30] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[31] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "EndoNet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 86–97, Jan. 2017.

[32] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015.

[34] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, "A fixed-point model for pancreas segmentation in abdominal CT scans," in *Proc. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, 2017, pp. 693–701.

[35] F. Liu, Z. Zhou, H. Jang, A. Samsonov, G. Zhao, and R. Kijowski, "Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging," *Magn. Reson. Med.*, vol. 79, no. 4, pp. 2379–2391, Apr. 2018.

[36] F. Ambellan, A. Tack, M. Ehlke, and S. Zachow, "Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the osteoarthritis initiative," *Med. Image Anal.*, vol. 52, pp. 109–118, Feb. 2019.

[37] M. R. Robu, P. Edwards, J. Ramalhinho, S. Thompson, B. Davidson, D. Hawkes, D. Stoyanov, and M. J. Clarkson, "Intelligent viewpoint selection for efficient CT to video registration in laparoscopic liver surgery," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 7, pp. 1079–1088, Jul. 2017.

[38] M. R. Robu, J. Ramalhinho, S. Thompson, K. Gurusamy, B. Davidson, D. Hawkes, D. Stoyanov, and M. J. Clarkson, "Global rigid registration of CT to video in laparoscopic liver surgery," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 6, pp. 947–956, Jun. 2018.

[39] D. Lee, J. Yoo, and J. C. Ye, "Deep residual learning for compressed sensing MRI," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 15–18.

[40] L. C. García-Peraza-Herrera, W. Li, C. Gruijthuijsen, A. Devreker, G. Attilakos, J. Deprest, E. V. Poorten, D. Stoyanov, T. Vercauteren, and S. Ourselin, "Real-time segmentation of non-rigid surgical tools based on deep learning and tracking," in *Proc. Int. Workshop Comput.-Assist. Robot. Endoscopy*, 2017, pp. 84–95.

[41] A. Jin, S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein, and L. Fei-Fei, "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 691–699.

[42] A. Vardazaryan, D. Mutter, J. Marescaux, and N. Padoy, "Weakly-supervised learning for tool localization in laparoscopic videos," in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis* (LNIP), vol. 11043. Basel, Switzerland: Springer, 2018, pp. 169–179.

[43] A. Shvets, A. Rakhlin, A. A. Kalinin, and V. Iglovikov, "Automatic instrument segmentation in Robot-Assisted surgery using deep learning," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl.*, Dec. 2018, pp. 624–628.

[44] H. Zhao, P. Bai, Y. Peng, and R. Xu, "Efficient key management scheme for health blockchain," *CAAI Trans. Intell. Technol.*, vol. 3, no. 2, pp. 114–118, Jun. 2018.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science), vol. 9351. Basel, Switzerland: Springer, 2015, pp. 234–241.

[46] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Lecture Notes in Computer Science), vol. 11045. Berlin, Germany: Springer-Verlag, 2018, pp. 3–11.

[47] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[48] S. M. Kamrul Hasan and C. A. Linte, "U-NetPlus: A modified encoder-decoder U-net architecture for semantic and instance segmentation of surgical instrument," 2019, *arXiv:1902.08994*. [Online]. Available: http://arxiv.org/abs/1902.08994

[49] *MeVisLab: MeVisLab*. [Online]. Available: https://www.mevislab.de/

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[52] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[53] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pyTorch," in *Proc. NIPS*, 2017, pp. 1–4.

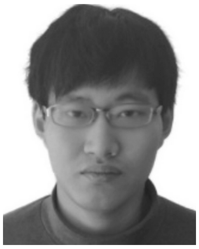
[54] M. Anthimopoulos, S. Christodoulidis, L. Ebner, T. Geiser, A. Christe, and S. Mougiakakou, "Semantic segmentation of pathological lung tissue with dilated fully convolutional networks," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 714–722, Mar. 2019.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[56] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer, "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains," *NeuroImage*, vol. 21, no. 4, pp. 1428–1442, Apr. 2004.

YAQUB JONMOHAMADI received the Ph.D. degree in neuroimaging, with focus on EEG analysis, from the University of Otago, New Zealand, in 2014. He started his first postdoctoral fellowship in multimodal neuroimaging at the University of Auckland, New Zealand, in 2015. He is currently undertaking his second postdoctoral fellowship in medical robotics and medical imaging at the Queensland University of Technology, Australia. His primary research interests are neuroimaging, biomedical imaging, computer vision, and artificial intelligence.

YU TAKEDA received the degree in medicine from the Hyogo College of Medicine, Japan, in 2009, and the Ph.D. degree from the Hyogo College of Medicine, in 2018. He is currently an Orthopaedic Surgeon. He is also working as a Researcher in autonomous knee surgery robotic applications with the Queensland University of Technology, Australia.



FENGBEI LIU received the B.S. degree (Hons.) in computer science from The University of Adelaide, where he is currently pursuing the Ph.D. degree. His research interest includes medical image analysis, computer vision, and deep learning.

FUMIO SASAZAWA received the degree from the Faculty of Engineering, The University of Tokyo, Tokyo, Japan, in 1997, and the degree from the School of Medicine, Shinshu University, Matsumoto, Japan, to obtain medical license, in 2004, and the Ph.D. degree in cellular and molecular biology from the Hokkaido University Graduate School of Medicine, in 2014. He worked as a Visiting Researcher with the Medical Robotics Team, Queensland University of Technology, Brisbane, Australia, from 2017 to 2018. He is currently an Orthopaedic Surgeon specializing in lower extremities including hip and knee joint.

GABRIEL MAICAS received the Ph.D. degree in medical image analysis from The University of Adelaide, in 2018. He is currently a Research Fellow with The Australian Institute for Machine Learning, The University of Adelaide. His main research interests are in the field of medical image analysis, computer vision, machine learning, and artificial intelligence.

ROSS CRAWFORD is currently a Professor of orthopaedic research with QUT and undertake private clinical practice at the Prince Charles and Holy Spirit Hospital. He is also a member of numerous medical committees. He has published more than 200 articles. As an Expert Surgeon, he assists with cadaver surgery experiments at the QUT Medical and Engineering Research Facility, Prince Charles Campus, and brings significant knowledge of knee arthroscopy and the use of medical robotics to this research.

JONATHAN ROBERTS (Senior Member, IEEE) received the Ph.D. degree from the University of Southampton, U.K., in 1994. He was the President of the Australian Robotics and Automation Association, from 2007 to 2008. He has been a Research Director of the Autonomous Systems Lab, CSIRO, since 2009. He is currently a Professor in robotics and autonomous systems with the Queensland University of Technology. He was a member of the IEEE Robotics and Automation Society.

AJAY K. PANDEY received the Ph.D. degree in experimental physics with mention tres honorable from the University of Angers, France, in 2007. He has worked at University of St Andrews and the University of Queensland. He leads an interdisciplinary Research Group that specializes in implementation of advanced materials and computer vision approaches to machine vision medical and neuro-imaging, intelligent bionics and soft robotics. He is currently a Senior Lecturer in robotics and autonomous systems with the School of Electrical Engineering and Robotics, QUT. He is an Editorial Board Member of *Scientific Reports* a Springer Nature Group journal.

GUSTAVO CARNEIRO received the Ph.D. degree in computer science from the University of Toronto, in 2004. He has worked at Siemens Corporate Research, University of British Columbia, and the University of California San Diego. He is currently a Professor with the School of Computer Science, The University of Adelaide, and also the Director of medical machine learning with the Australian Institute of Machine Learning. His primary research interests are in the fields of computer vision, medical image analysis, and machine learning.