Title: SurgAI: Deep Learning for Computerized Laparoscopic Image Understanding in Gynaecology

Running title: Artificial Intelligence in Gynaecology

Authors:

Sabrina MADAD ZADEH[1,2]; Tom FRANCOIS[2]; Lilian CALVET[2] Ph.D; Pauline CHAUVET[1,2] M.D; Michel CANIS[1,2] M.D, Ph.D; Adrien BARTOLI[2] Ph.D; Nicolas BOURDEL[1,2,*] M.D, Ph.D

*Corresponding author:

CHU Clermont-Ferrand, Department of Gynaecological surgery, 1 place Lucie et Raymond Aubrac, 63000 Clermont-Ferrand, France; nicolas.bourdel@gmail.com; +33676713113

[1]Department of Gynaecological surgery. CHU Clermont-Ferrand, France.

[2]EnCoV, Institut Pascal, CNRS, Université Clermont Auvergne, Clermont-Ferrand, France.

Abstract:

Background: In laparoscopy, the digital camera offers surgeons the opportunity to receive support from image-guided surgery systems. Such systems require image understanding, the ability for a computer to understand what the laparoscope sees. Image understanding has recently progressed owing to the emergence of artificial intelligence and especially deep learning techniques. However, the state of the art of deep learning in gynaecology only offers image-based detection, reporting the presence or absence of an anatomical structure, without finding its location. A solution to the localisation problem is given by the concept of semantic segmentation, giving the detection and pixel-level location of a structure in an image. The state of the art results in semantic segmentation are achieved by deep learning, whose usage requires a massive amount of annotated data. We propose the first dataset dedicated to this task and the first evaluation of deep learning based semantic segmentation in gynaecology.

Methods: We used the deep learning method called Mask R-CNN. Our dataset has 461 laparoscopic images manually annotated with three classes: uterus, ovaries and surgical tools. We split our dataset in 361 images to train Mask R-CNN and 100 images to evaluate its performance.

Results: The segmentation accuracy is reported in terms of percentage of overlap between the segmented regions from Mask R-CNN and the manually annotated ones. The accuracy is 84.5%, 29.6% and 54.5% for uterus, ovaries and surgical tools respectively. An automatic detection of these structures was then inferred from the semantic segmentation results which led to state of the art detection performance, except for the ovaries. Specifically, the detection accuracy is 97%, 24% and 86% for uterus, ovaries and surgical tools respectively.

Conclusion: Our preliminary results are very promising, given the relatively small size of our initial dataset. The creation of an international surgical database seems essential.

Keywords: *laparoscopic surgery - artificial intelligence - deep learning - gynaecological surgery*

Main text:

Introduction

Laparoscopic surgery has revolutionised surgery. In particular, it has brought a digital camera to the operating room. The acquired laparoscopic images provide a wealth of information, yet substantially underused, as processing this amount of information in real-time exceeds the human brain's ability. The increased computational capabilities and the recent advances in artificial intelligence (AI), specifically in machine learning, now make a standard computer able *to understand* the content of an image or a video stream in real-time.

Machine learning is a scientific field which develops algorithms to train computers at performing a task. In machine learning, the computers are not specifically programmed to achieve the task at hand: they use data from past experience along with the expected results

to teach the computers what should be done or solved. Among the machine learning techniques, deep learning, based on the use of artificial neural networks, has shown to perform better than the human perception on a wide spectrum of visual tasks [1], such as face recognition. This is also the case in the medical domain with the development of computer-aided diagnosis systems to detect melanoma [2,3] or diabetic retinopathies [4].

The state of the art in deep learning for laparoscopic surgery in gynaecology only explored detection tasks. In [5,6], the authors propose to use detection to automatically solve for the presence or absence of some anatomical structures in a laparoscopic image and recognise the action being performed. The state of the art is thus very limited. However, it is now clear that the use of the recent scientific advances in deep learning would be beneficial to a large number of image-guided surgery applications. For example, deep learning techniques could be used to fully automate image-guided surgery systems such as the one developed by our team for gesture guidance in myomectomy [7-10]. Automatically locating and highlighting critical anatomical structures during surgery, such as the ureters or major vessels, on the surgeon's monitor, could improve gesture safety. This would allow the surgeon to know the exact position of critical anatomical structures and better respect these during the procedure. A major reason currently preventing the use of deep learning techniques in image-guided surgery systems for gynaecology is the lack of annotated data, required to train artificial neural networks.

We provide a dataset of laparoscopic gynaecological images for which the image extent of anatomical structures and tools was precisely annotated. It is the first dataset of its kind, allowing one to train an artificial neural network to achieve semantic segmentation on laparoscopic images in gynaecology. The semantic segmentation task consists in classifying every pixel of an image, revealing the anatomical structure it belongs to. This is a much richer piece of information than the state of the art datasets dedicated to detection only, which simply indicate the presence or absence of each class. We report the results delivered by Mask R-CNN [11], a state of the art deep learning based semantic segmentation method in the scientific field of deep learning. The artificial neural network is trained on the proposed image dataset

with pre-training on ImageNet. Our dataset, named *SurgAI*, will be made publicly available upon acceptance of the paper. Image-guided surgery applications that would possibly benefit from semantic segmentation are then discussed.

Materials and Methods

Overview of deep learning and the semantic segmentation task

Deep learning techniques follow two stages. The first stage is the training stage, during which the computer learns the task from a large dataset of examples including the information to be automatically predicted, namely image annotations defined manually by an expert. The second stage is the prediction stage. The computer predicts the desired information on new, previously unseen images, thanks to its "experience" accumulated at the training stage [1]. In this work, this approach has been applied to laparoscopic images in order to perform a semantic segmentation of anatomical structures and surgical tools. Our system performs semantic segmentation by attributing a class per pixel. The pixel classes are uterus, ovary and surgical tools composed of scissors, grasper, bipolar and atraumatic forceps.

Patients and images

A total of eight laparoscopic hysterectomies recorded in our tertiary center in Clermont-Ferrand from February to March 2018 were used. The patients underwent a classical laparoscopic hysterectomy, whose videos were recorded. The indication for each hysterectomy was endometriosis or adenomyosis. A total of 461 images were then manually chosen to provide the dataset with a diversity of points of view and surgical steps permitting to ensure an intra-patient variability. The images were extracted from eight different procedures, to also ensure an inter-patient variability. The videos were collected in a study protocol which was approved by the French Ethics committee "Comité de Protection des Personnes Sud-Est VI, France", 2016-002773-35. The protocol was registered in clinicaltrials.gov (NCT03080558). A signed consent form was obtained from each patient. This protocol permits the use of images

in an anonymous way for other studies in the same tertiary center. All the images are strictly anonymous.

Image annotation and machine training

The laparoscopic images were manually annotated with the organ regions representing the anatomical structures and the surgical tools to be segmented. The annotations were provided by a junior surgeon (SMZ) supervised by an expert surgeon (NB). Figure 1 shows an excerpt of the proposed dataset with examples of annotations for uterus, ovaries and surgical tools. The annotations were performed with a software called “Supervisely” available online [12]. Table 1 reports the statistics of the annotated structures.

We used Mask Regional Convolutional Neuronal Network (Mask R-CNN) from Facebook Artificial Intelligence Research [13], which is known as a competitive deep learning algorithm for automatic semantic segmentation. Its computer code is publicly available. Transfer learning generally consists in updating a pretrained model, namely updating the neural network weights, in order to specialize the network function either to another task than the one it has been trained for or to make it deal with a different input image domain. In our case, the pretraining is performed through the same task while the input image domain differs. ImageNet [14] was used for the original training. It is composed of natural images containing daily life objects (e.g. tables, cars and bicycles). More details on this dataset can be found on http://www.image-net.org. Pretraining a network with a very large but unrelated to the actual task dataset such as ImageNet is a common practice in deep learning, especially if the actual dataset is limited in size.

In our context the intra-operative images form the input data. The images were stored in PNG format at a resolution of 1920x1080 pixels. The images were resized during training. Images of same resolution are particularly desirable to deal with GPU memory issues but is not a strong requirement. Images of different resolutions are handled by Mask R-CNN whose

preprocessing includes an automatic image resizing so that the shorter image side is in our case set to 600 pixels. A total of 886 annotated regions were marked over the 461 images of the dataset. About 90% of the images actually contain annotated regions. The remaining images do not contain any structure of interest: they are called “negative images” (external view of the OR, images inside the trocar). The state of the art shows that in a limited-size dataset, these negative images improve the knowledge of the network about the studied environment, here the medical one. The output data are a class label per image pixel. Mask R-CNN was trained with images from 6 surgeries and tested with laparoscopic images from 2 surgeries not included in the training set. The training took about 5 to 6 hours. A single GPU GeForce GTX 1080 was used.

Results

An evaluation was performed for the semantic segmentation task and for a detection task of uterus, ovaries and surgical tools. The detection was inferred from the semantic segmentation results. Two examples from the obtained segmentation results we obtained are shown in Figure 2.

Semantic segmentation task. The proposed semantic segmentation is assessed by measuring the Intersection over Union (IoU). IoU represents the percentage of overlap between the manually annotated region considered as *ground truth*, namely the ideal segmentation for a given structure, and the predicted region, segmented by Mask R-CNN. This notion is illustrated in Figure 3 and in Figure 4. The obtained segmentation accuracy in terms of mean IoU over the entire evaluation image dataset is of 84.5%, 29.6% and 54.5% for uterus, ovaries and surgical tools respectively.

Detection task. To evaluate the accuracy of detection, a predicted region is considered a *true positive* (TP) if its IoU with the annotated region is greater than a threshold *t* and its associated predicted class corresponds to the annotated region class. It is considered a *false positive* (FP) otherwise. Precision is the test's positive predictive value: it measures the proportion of actual positives which are correctly identified depending on the prevalence of the to be detected structure (a class) in the dataset. Recall reports the TP rate, namely the test's sensitivity: it measures the proportion of actual positives that are correctly identified. The results can be graphically shown using ROC curves or precision-recall curves according to the class balance of the dataset. In case of a dataset with a class imbalance, a ROC curve presents an optimistic view of the results [15]. In our case, the dataset is composed of slightly imbalanced classes, and we thus decided to provide precision-recall curves. A perfect test is depicted as a point at (1,1). The results are shown in Figure 5 We observe that very good detection results are obtained with an IoU threshold of 50% for the uterus and the surgical tools, for which both precision and recall then to 1, specifically converging to (0.99, 0.97) and to (0.88, 0.86) respectively. For the ovaries however, the detection results are poorer, with precision and recall stalling at (0.28, 0.24).

Discussion

The proposed preliminary trial of a deep learning based semantic segmentation method applied to laparoscopic images in gynaecology shows very encouraging results. Among the segmented structures, similar performance results to those reported in the literature for the semantic segmentation of everyday life objects [16] are obtained for the uterus and surgical tools. The case of the uterus is explained by the highest number of annotated instances in the proposed dataset compared to the other structures. The case of the surgical tools, for which the number of annotated images is significantly lower than for the uterus, is explained as follows. First, the appearance of surgical tools highly differs from the tissues of the abdominal cavity. Second, the deep learning system has been pretrained with a dataset including basic

pairs of scissors, which have a similar appearance to surgical tools. Some recent papers provided better performance in semantic segmentation of surgical tools such as Islam et al [17]. The difference in segmentation performance between Islam et al.'s and ours can be essentially explained by a higher number of training images of 1350 against 361 in our case, and a lower number of classes associated to their dice score of 0.916 (mIoU of 84%), namely two for “surgical tool” and “background”, compared to our three classes. We recall that the surgical tool class is the easiest to segment, given its colour distinctiveness to the rest of the image colours.

The segmentation results for the ovaries were not as good as for uterus and tool. This can be explained by the following two main reasons. First, the training dataset contains a lower number of ovary instances as this structure is often hidden by other structures (the fallopian tube or uterus). Second, the ovaries present a highly varying appearance across patients. For the detection task, the results are similar to the state of art, such as those reported in Liebetseder et al. [6] and Choi et al. [17]. The threshold of the IoU must be chosen tighter than 50% to increase the accuracy of the semantic segmentation for an application in surgery. The indication for each hysterectomy was endometriosis or adenomyosis. It is a limitation of the proposed dataset in its current state as covering a single pathology. However, we consider that images and anatomical structures appearance overall present a similar variability as for other uterine pathologies such as endometrial cancers. A good generalization to other pathologies is therefore expected. Future contributions to the proposed dataset will support the validation of this assumption through tests run on other pathologies.

Our segmentation results are provided along with the first dataset for laparoscopic surgery, allowing one to train a computer for a semantic segmentation task for laparoscopic images from surgical procedures in gynaecology. The current public datasets including laparoscopic images from surgery in gynaecology are either limited to surgical tools annotations [18-20] or are limited to classification [6], namely indicating whether or not an

anatomical structure is visible in the image or the surgical action being performed [21]. The proposed dataset provides a much richer information as it consists of annotations per image pixel.

Such a dataset may pave the way to the development of highly impactful deep learning based computer-aided applications. It will enable the use of deep learning methods for the segmentation of pelvic anatomical structures segmentation, allowing one to automate the gesture guidance in myomectomy developed by our team [7-10]. Our augmented reality software requires only the detection of surgical tools regardless of their specific type. It is the reason why Mask R-CNN was trained to perform a category-level semantic segmentation. We plan to extend the proposed dataset with images including ureter annotations. Ureteral wounds remain one of the most critical complications in laparoscopic gynaecological surgery. Being able to automatically segment the ureters in real time during surgery is therefore of major interest. Other tasks that could be possibly achieved through deep learning methods are the localisation of the origin of bleeding, which presents significant clinical benefits when not directly visible by the human eye. An application for automatic and real-time segmentation of peritoneal carcinomatosis or endometriosis nodules would also be of great interest as remaining a task difficult to achieve for non-experienced surgeons.

Our approach is part of the trend towards the development of AI applications. AI development has been largely accelerated by massive image datasets available in open source such as “ImageNet” or “Microsoft coco” [22] for the automatic detection of everyday life objects. In other medical fields, computer-aided diagnosis based on machine learning techniques have been developed and is used in daily-medicine like the computer-aided diagnosis in breast cancer screening in the USA [23]. The AI has shown better capacity than the human brain for tasks such as melanoma detection in dermatology [3], even compared to experimented dermatologists [2]. Another study showed that a deep learning system obtained similar results to the ophthalmologist experts in the analysis of diabetic retinopathy [4]. A deep learning based computer system has been used by Tinwanda et al. to predict surgery duration

on the basis of the first minutes of laparoscopic surgery [24]. It could help the surgeon to adjust precisely the operative theatre schedule.

From a more general perspective, this approach and discussed applications take part to the introduction of surgery in the current digital transformation already observed in many industrial fields. It could be very helpful for risk assessment, complication anticipation, or surgeon gesture quality rating and follow-up.

Conclusion

Our preliminary results of applying a deep learning based semantic segmentation system in gynaecology are very promising, even with a limited amount of training data. More images are necessary to train the deep learning system to improve the results obtained for some challenging anatomical structures. Similarly to the datasets dedicated to everyday life object detection, the creation of an international surgical database seems essential.

Acknowledgments: None.

Disclosures: S. Madad Zadeh, T. François, L. Calvet, P. Chauvet, M. Canis, A. Bartoli and N. Bourdel report no conflict of interest.

References

1. Ian Goodfellow, Yoshua Bengio and Aaron Courville, Deep Learning, MIT 2016

2. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, Kalloo A, Hassen ABH, Thomas L, Enk A, Uhlmann L, Reader study level-I and level-II Groups (2018) Man against machine: diagnostic performance of a deep learning convolutional neural network for

dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 29:1836–1842 . doi: 10.1093/annonc/mdy166

3. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118

4. Ting DSW, Cheung CY-L, Lim G, Tan GSW, Quang ND, Gan A, Hamzah H, Garcia-Franco R, San Yeo IY, Lee SY, Wong EYM, Sabanayagam C, Baskaran M, Ibrahim F, Tan NC, Finkelstein EA, Lamoureux EL, Wong IY, Bressler NM, Sivaprasad S, Varma R, Jonas JB, He MG, Cheng C-Y, Cheung GCM, Aung T, Hsu W, Lee ML, Wong TY (2017) Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. JAMA 318:2211–2223

5. Petscharnig S, Schöffmann K (2018) Learning laparoscopic video shot classification for gynecological surgery. Multimed Tools Appl 77:8061–8079 .

6. Leibetseder A, Petscharnig S, Primus MJ, Kletz S, Münzer B, Schoeffmann K, Keckstein J (2018) Lapgyn4: A Dataset for 4 Automatic Content Analysis Problems in the Domain of Laparoscopic Gynecology. In: Proceedings of the 9th ACM Multimedia Systems Conference. ACM, New York, NY, USA, pp 357–362

7. Bourdel N, Chauvet P, Calvet L, Magnin B, Bartoli A, Michel C (2019) Use of augmented reality in Gynecologic surgery to visualize adenomyomas. J Minim Invasive Gynecol.

8. Chauvet P, Collins T, Debize C, Novais-Gameiro L, Pereira B, Bartoli A, Canis M, Bourdel N (2018) Augmented reality in a tumor resection model. Surg Endosc 32:1192–1201

9. Bourdel N, Collins T, Pizarro D, Debize C, Grémeau A-S, Bartoli A, Canis M (2017) Use of augmented reality in laparoscopic gynecology to visualize myomas. Fertil Steril 107:737–739

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

11. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp 2980–2988

12. Supervisely - Web platform for computer vision. Annotation, training and deploy. https://supervise.ly

13. (2018) FAIR's research platform for object detection research, implementing popular algorithms like Mask R-CNN and RetinaNet.: facebookresearch/Detectron. Facebook Research

14. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L ImageNet: A Large-Scale Hierarchical Image Database

15. Jesse Davis, Mark Goadrich, The relationship between Precision-Recall and ROC curves in International Conference on Machine Learning (ICML) 2006 Proceedings of the 23rd ICML pages 233-240.

15. Chen L, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence 40:834–848

16. Islam, M., Atputharuban, D.A., Ramesh, R., Ren, H.: Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning. IEEE Robot. Autom. Lett. 4, 2188–2195 (2019)

18. Choi B, Jo K, Choi S, Choi J (2017) Surgical-tools detection based on Convolutional Neural Network in laparoscopic robot-assisted surgery. In: 2017 39th Annual International Conference

of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, Seogwipo, pp 1756–1759

19. EndoVisSub-Instrument-Results.https://endovissub-instrument.grand challenge.org/Results

20. García-Peraza-Herrera LC, Li W, Fidon L, Gruijthuijsen C, Devreker A, Attilakos G, Deprest J, Poorten EV, Stoyanov D, Vercauteren T, Ourselin S (2017) ToolNet: Holistically-nested real-time segmentation of robotic surgical tools. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp 5717–5722

21. Stauder R, Ostler D, Kranzfelder M, Koller S, Feußner H, Navab N (2016) The TUM LapChole dataset for the M2CAI 2016 workflow challenge. arXiv:161009278 [cs]

22. Lin T-Y, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollár P (2014) Microsoft COCO: Common Objects in Context. arXiv:14050312 [cs]

23. Fazal MI, Patel ME, Tye J, Gupta Y (2018) The past, present and future role of artificial intelligence in imaging. Eur J Radiol 105:246–250

24. P. Twinanda, G. Yengera, D. Mutter, J. Marescaux and N. Padoy. RSDNet: Learning to Predict Remaining Surgery Duration from Laparoscopic Videos Without Manual Annotations. In 2019 IEEE Transactions on Medical Imaging, vol. 38, no. 4, pp. 1069-1078.

Table legends:

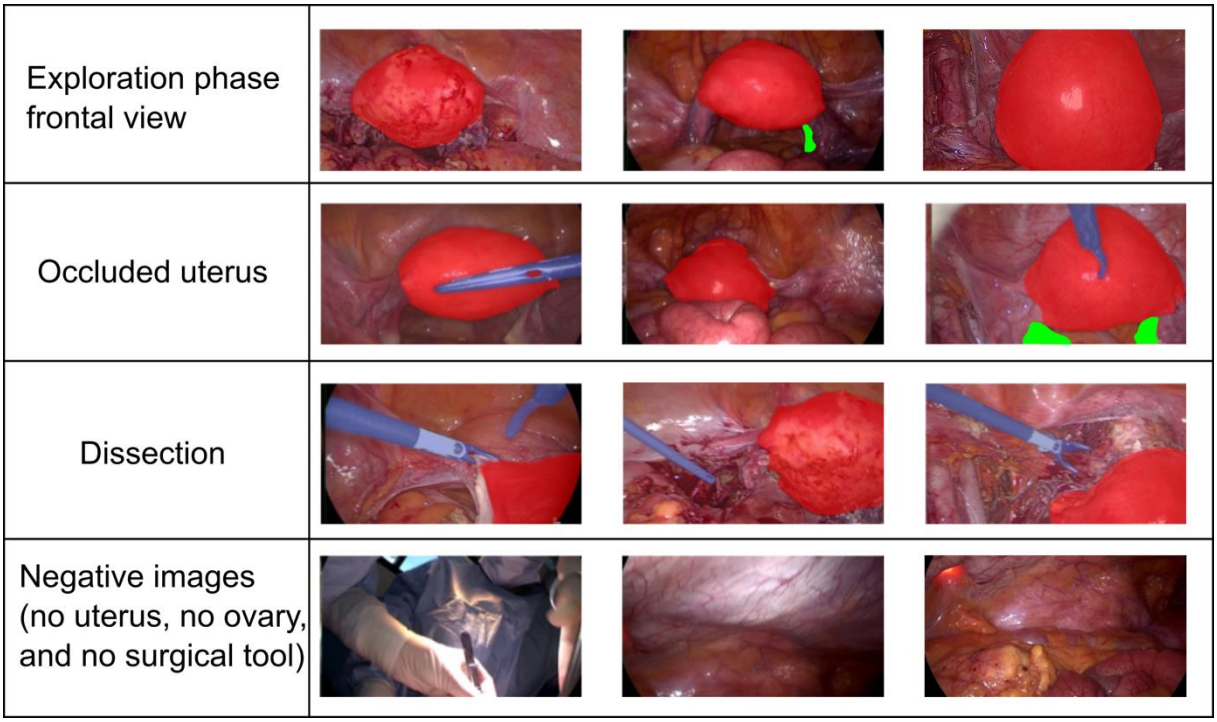
Legend Table 1:

Table 1. Overview of the provided image annotations.

Figure legends:

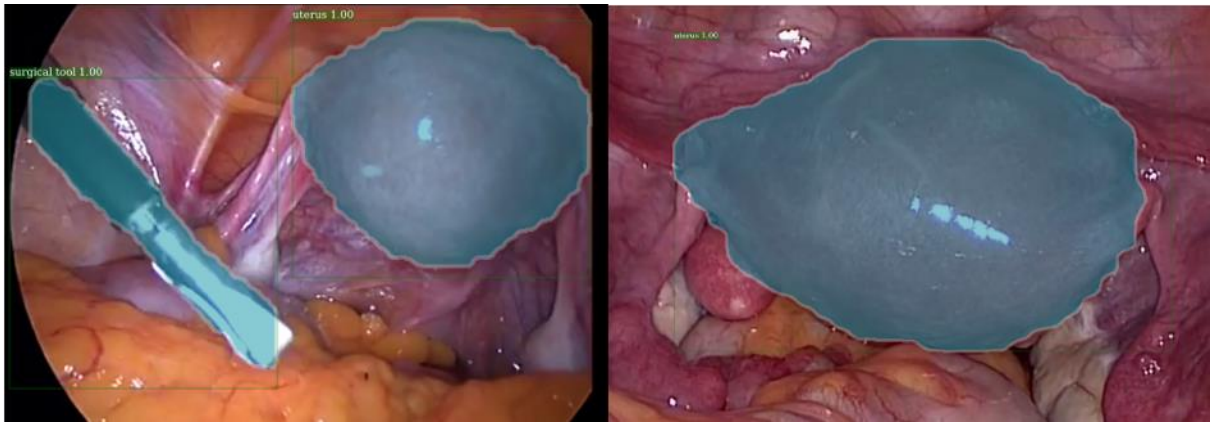
Legend Figure 1:

Figure 1: Overview of SurgAI, the proposed dataset of annotated laparoscopic gynaecological images. The dataset includes images representative of the various stages of surgery. The annotated regions in red, green and blue are associated to the uterus, ovaries and surgical tools classes respectively.



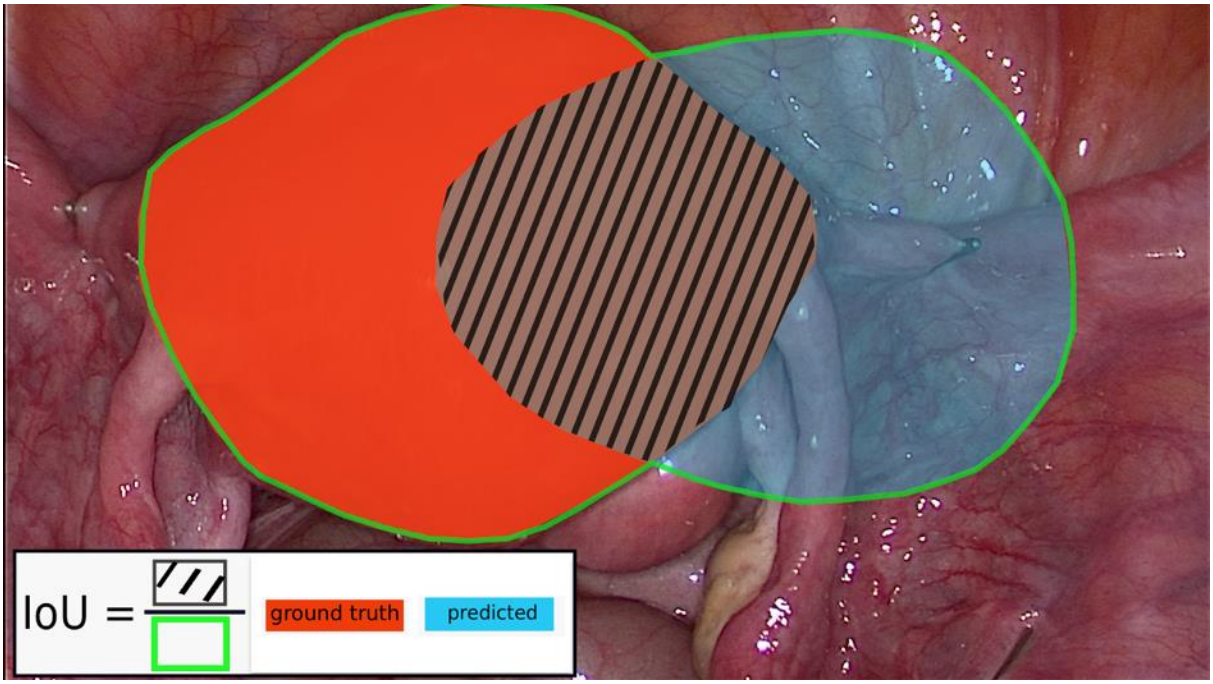
Legend Figure 2:

Figure 2: Illustration of the semantic segmentation results obtained for (a) the surgical tool with an IoU of 87% and uterus with an IoU of 92% and (b) the uterus with an IoU of 91%.



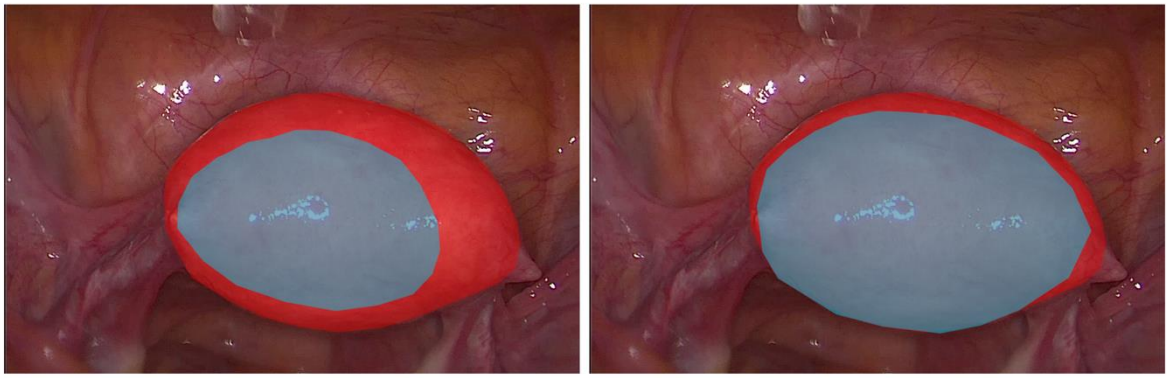
Legend Figure 3:

Figure 3: Illustration of the Intersection over Union (IoU) metric.



Legend Figure 4:

Figure 4: Illustration of the quality of the prediction. In these examples, the IoU is 50% in image (a) and 70% in image (b).



(a) (b)

Legend Figure 5:

Figure 5. Automatic detection results obtained for the uterus, the ovaries and the surgical tools. The results are reported in terms of precision and recall (%).