Automatically Recognising the Kidney in Robot-Assisted Partial Nephrectomy using Deep Learning: A Comprehensive Dataset with Inter-Annotator Variability and Neural Network Performance Analysis (UroCCR 122)

G. Margue, K. Chandelon, M. Pattou, A. Pitout, A. Khaddad, F. Rubat-Baleuri, J. Desternes, J. Peyras, L. Richert, N. Bourdel, A. Bartoli, JC. Bernhard

Introduction

Partial nephrectomy (PN) is the gold standard for treating localised kidney tumors, preserving renal function while achieving excellent oncological outcomes[1,2]. The advent of minimally invasive techniques, particularly the robot-assisted approach, has transformed surgical oncology by offering benefits including reduced blood loss, shorter hospital stays, and fewer complications[3,4]. However, robot assisted partial nephrectomy (RAPN) faces challenges in managing complex renal masses, which can lead to significant morbidity[5,6]. Achieving optimal outcomes require meticulous preoperative planning and precise intraoperative guidance. This may be achieved by using high-resolution contrast-enhanced CT, which enables the creation of personalized virtual 3D models[7–9]. The 3D image-guided RAPN (3D-IGRAPN) technique[10] enhances surgery by displaying these models into the robotic interface. Augmented Reality (AR) goes a step further by overlaying digital information onto the surgical field directly,[11] providing real-time visualizations aiding in precise tumor resection[12]. Unlike Virtual Reality (VR), AR enhances the real-world perception with digital overlays, relying on registration and real-time organ tracking[13,14]. These processes are crucial for overcoming challenges such as organ motion and intraoperative perturbations.

The most difficult problem in AR is probably registration, which is to align the virtual 3D models with the surgical field. Solving the registration problem requires a number of steps, including the recognition and localisation of the kidney in the surgical images, called the image segmentation problem, from which an intraoperative 3D surface may be reconstructed by means of stereovision. Recent advances in computer vision and machine learning have established that deep neural networks could be specifically trained to solve image segmentation tasks, and could thus form an important tool to solve the kidney recognition and localisation step in an automatic AR system.

This study aims to describe the development of a kidney segmentation system for RAPN. It focuses on the collection of the huge amount of data required to train neural network. It shows how to construct a comprehensive and diverse dataset with quality annotations and on evaluating inter-annotator variability.

Method

To train a neural network for accurately segmenting the renal parenchyma in intraoperative images. A high-quality dataset was constructed through the following steps: image collection, anonymization, pre-processing, annotation, and validation.

RAPN surgical videos were collected at one expert centre using the da Vinci® Surgical System (Intuitive Surgical®). Videos were recorded through the MVR Pro system (MediCapture®), ensuring high-resolution video streams (1280 x 1024 px) were maintained. Ethical compliance was ensured under UroCCR protocol N°122 (Clinical Trials: NCT03293563; CNIL DR-2013-206) and patient confidentiality was preserved through rigorous anonymization techniques employing Spyder (Scientific PYthon Development EnviRonment; Version 5) and the FFMpeg library. This involved removing identifiable information and employing pseudonymization, with the resulting encrypted dataset stored securely with restricted access.

Image extraction from the video footage was automated based on a predefined set of selection criteria to guarantee optimal quality and diversity. Image sharpness was assessed using Canny edge detection[16], founded on the principle that sharp images exhibit a high density of well-defined contours. The sharpness score was computed as the ratio of detected edge pixels to the total image size, filtering out blurry or low-detail frames. A temporal spacing constraint was implemented to minimize redundancy by restricting frame extractions. This involved ensuring that frames were only extracted when the time interval between them was greater than five seconds. To achieve frame diversity, a Marr-Hildreth image hashing[17] algorithm was employed. This technique computed the perceptual similarity between frames, discarding those with with high redundancy based on a predefined Hamming distance threshold (value set to 200). Extracted images were carefully reviewed to exclude artifacts, subsequently annotated with metadata, including tumor location, size, and RENAL score[18]. The surgical procedures were divided into five critical phases: renal dissection, tumor preparation, hilar dissection, tumor resection, and tumor bed reconstruction, each phase chosen for its relevance to AR-guided improvements.

To achieve high-quality annotation, diverse annotator groups were defined: two junior urologists (regularly assisting in RAPN procedures); two physicians from other specialties (medical professionals without expertise in RAPN); two medical engineers (with technical backgrounds in segmentation process); two external annotators (individuals without surgical or anatomical expertise) and an expert surgeon (professor of urology with extensive experience in performing RAPN). All annotators received uniform structured training, leveraging a detailed guidebook and the Supervisely® software, focusing initially on basic annotation principles before advancing to the nuances of renal surgery images. Annotators were required to demonstrate proficiency by achieving scores of 80% in general training and 90% in renal-specific tasks, ensuring consistency and accuracy.

The reliability of annotated images was evaluated using sensitivity and the Dice Similarity Coefficient (DSC)[19] (Fig. 1), compared to the ground truth provided by the expert surgeon. Consistency across annotators was ensured through an evaluative dataset featuring images of varying complexity: simple, intermediate, and complex. The Structural Similarity Index Measure (SSIM) was used to confirm diversity, focusing on luminance, contrast, and structural features. SSIM scores, ranging from 0 to 100%, are higher with greater similarity. Our dataset's mean SSIM score was 35.9%, indicating sufficient diversity to test the segmentation process while

maintaining structural integrity and consistency. Heatmaps were generated to visually represent areas of consensus and discrepancies across annotators.

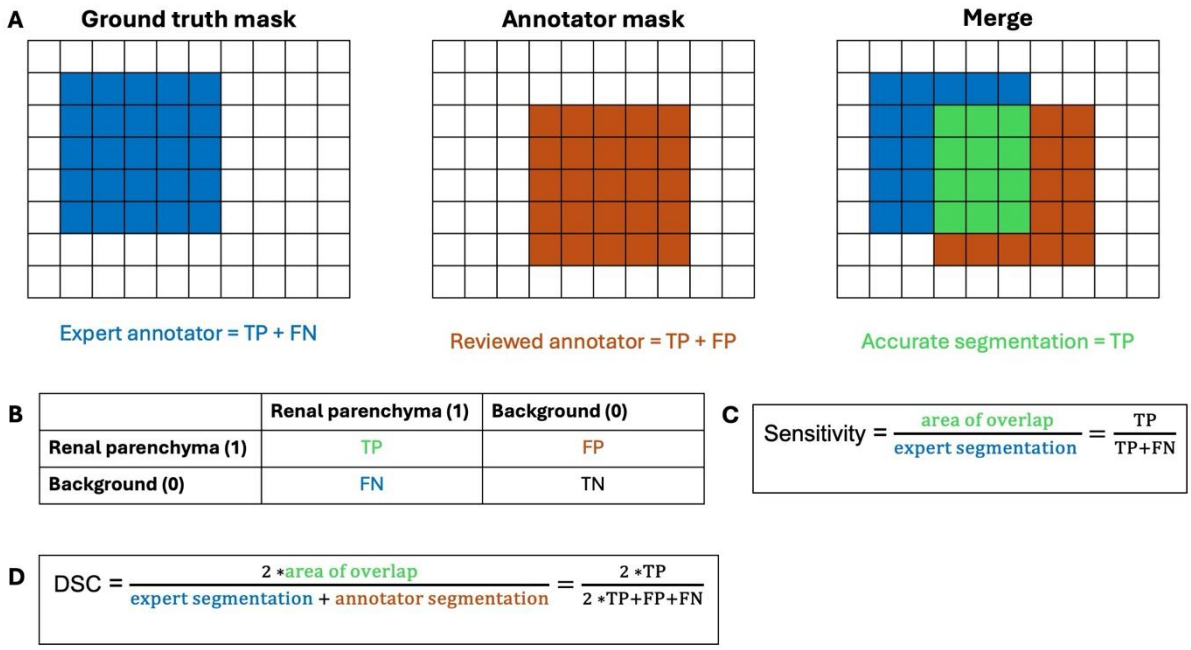

Figure 1 – Diagram of segmentation metrics derived from a confusion matrix for renal parenchyma segmentation. (A) Example of a confusion matrix; (B) Table defining true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN); (C) Equation for sensitivity; and (D) Equations for the Dice Similarity Coefficient (DSC).

Annotations from external contributors underwent statistical analysis to gauge accuracy before inclusion in the neural network dataset. This approach determined the minimum number of images required to validate the entire dataset's accuracy. We calculated confidence intervals for the proportion of correctly annotated images, using the formula:

$$IC = p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

where IC represents the confidence interval, p is the proportion of correctly segmented images, n is the total number of images, and 1.96 is the coefficient for a 95% confidence level.

AlbuNet-34[21], a U-Net architecture-based[20] deep convolutional neural network was implemented to automate renal parenchyma segmentation (Figure 2). The annotated dataset was divided into training and validation sets, with all images standardized to a resolution of 512 × 512 pixels. To address class imbalance inherent in the dataset, Weighted Dice Loss was applied to emphasize the minority class (renal parenchyma). This loss function assigned greater weight to the parenchyma, ensuring the model prioritized its accurate segmentation. The training process spanned 30 epochs, with a batch size of 8 to balance computational efficiency and convergence speed. Stochastic Gradient Descent (SGD) was used for optimization, starting with an initial

learning rate of 0.001. To enhance convergence, an automatic learning rate scheduler reduced the rate by a factor of 0.1 every 5 epochs, preventing overshooting as training advanced. Model performance was rigorously assessed through sensitivity and DSC metrics, comparing its segmentation outputs to those of expert annotations.

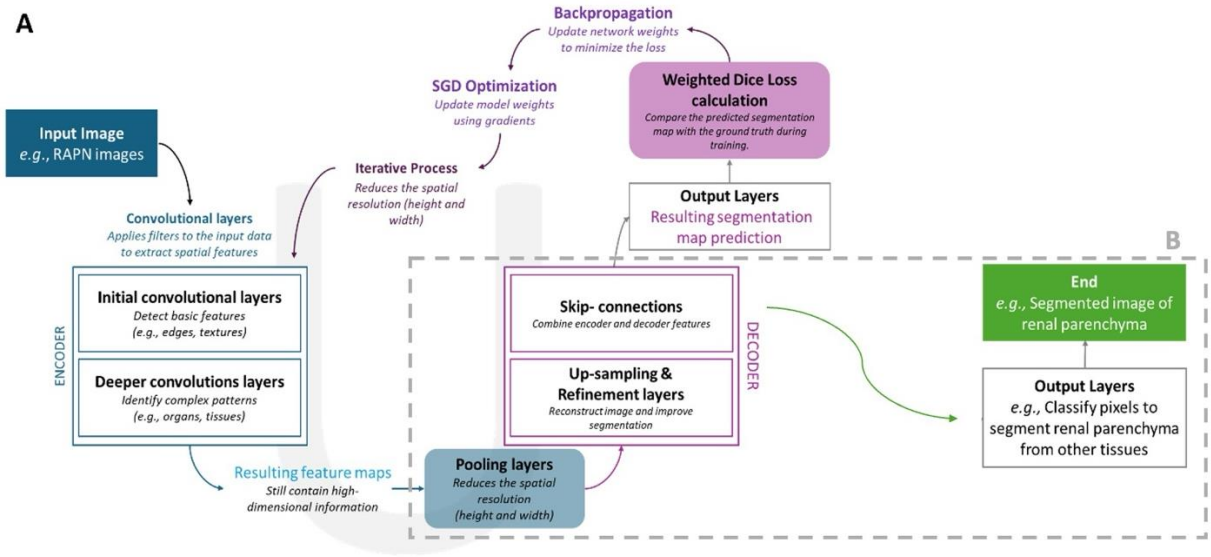

Figure 2 – Diagram of a typical deep convolutional neural network (CNN) combined with a segmentation algorithm based on the U-Net architecture (AlbuNet-34), illustrating (A) the training phase and (B) the inference phase. SGD, Stochastic Gradient Descent.

Statistical method.

Statistical analyses were conducted using GraphPad Prism® 10.4.1. software, applying non-parametric tests due to the non-normal distribution of data. The Kruskal-Wallis test was used to compare differences between multiple groups. Dunn's multiple comparisons test was then conducted to determine pairwise differences between groups. Statistical significance was considered at $p < 0.05$ for all tests. Results are presented as mean ± standard deviation (SD).

Results

To establish a comprehensive dataset for training an automated renal parenchyma recognition system, we gathered 131 RAPN surgical videos over three years, yielding approximately 48,000 images. We meticulously recorded patient information (Supplementary Table 1) and detailed tumor characteristics such as location, size, and proximity to critical structures, along with complexity metrics from the RENAL score (Table 1). This dataset faithfully represents modern urologic practice[22] and while some clinical data points were missing, they were randomly distributed, ensuring the sample's representativeness.

	Tumors n = 85
Number of tumors, mean ± SD	1.1 ± 0.3
Tumor size (cm), mean ± SD	4.7 ± 2.2
Kidney, n (%)	
Left	44 (51.8)
Right	41 (48.2)
Polar location, n (%)	
Superior	24 (28.2)
Equatorial	29 (34.1)
Inferior	32 (37.6)
Side of the kidney, n (%)	
Posterior	42 (50.6)
Anterior	43 (49.4)
Border, n (%)	
Median	39 (45.9)
Lateral	46 (54.1)
Hilar location, n (%)	
Yes	35 (41.2)
Endophytic component, n (%)	
Entirely endophytic	15 (17.6)
>50%	38 (44.7)
≤50%	32 (37.6)
Nearness to collecting system, n (%)	
> 7mm	14 (16.5)
4-7mm	8 (9.4)
<4mm	63 (74.1)
Complexity according to RENAL score, n (%)	
Low	19 (22.4)
Intermediate	39 (45.9)
High	27 (31.8)

Table 1 - Characteristics of kidney tumors in RAPN patients from the dataset (n =85 tumors in 78 patients).

We assessed segmentation performance across annotator groups on a subset of 454 manually annotated images used for training a neural network. Figure 3 illustrates DSC and sensitivity, revealing significant differences between groups. Junior urologists achieved the highest DSC (mean 0.86 ±0.23), contrasting with the external annotators' lower DSC (mean 0.81 ±0.24). Sensitivity also highlighted the superior performance of junior urologists. Heatmaps (Fig. 4) provided visual overlays of segmentation overlap and discrepancies among the nine annotators. While DSC scores over 0.8 across all groups reflect strong alignment with expert segmentations, a phase-based analysis (Fig. 5) indicated renal dissection, tumor resection, and bed reconstruction as challenging, suggesting areas for training refinement and guidebook

improvement. Despite variability, annotation reliability was robust (DSC > 0.85 and sensitivity > 0.89), underscoring the annotation accuracy.

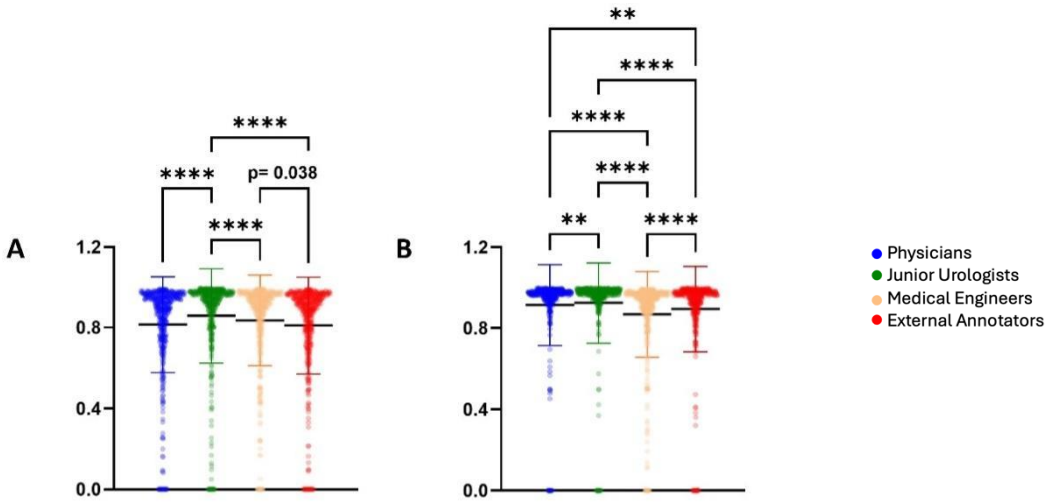

*Figure 3 - Comparison of Dice Similarity Coefficient (DSC) (A) and sensitivity (B) between annotators compared to an expert urologist's ground truth segmentation. Data are presented as mean ±SD. ** p < 0.01, **** p < 0.0001.*

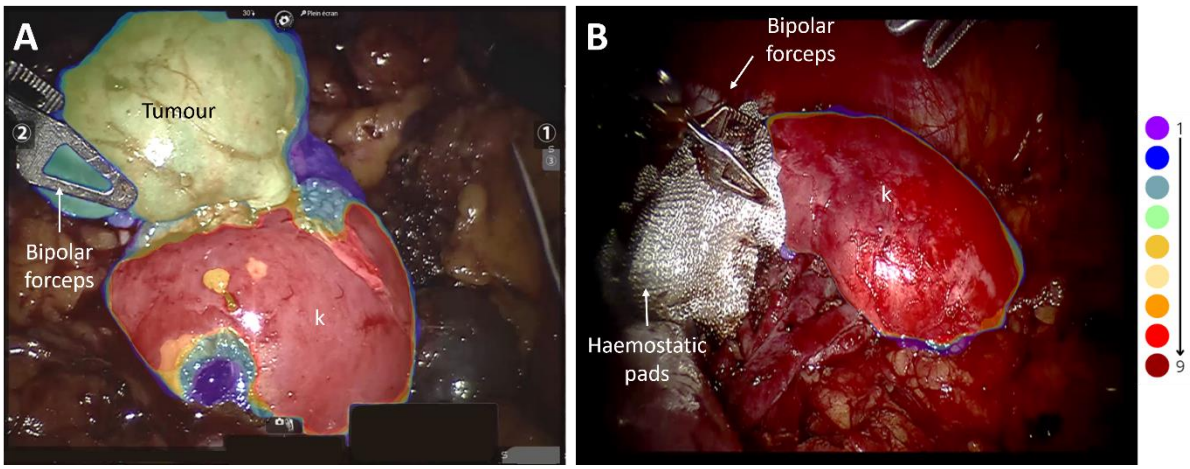

Figure 4 – Variability in annotations of renal parenchyma in images from robot-assisted laparoscopic partial nephrectomy at two surgical stages (A: Renal dissection; B: Tumor bed reconstruction). The purple colour corresponds to pixels annotated by a single annotator, while the burgundy red corresponds to pixels annotated by all 9 annotators.

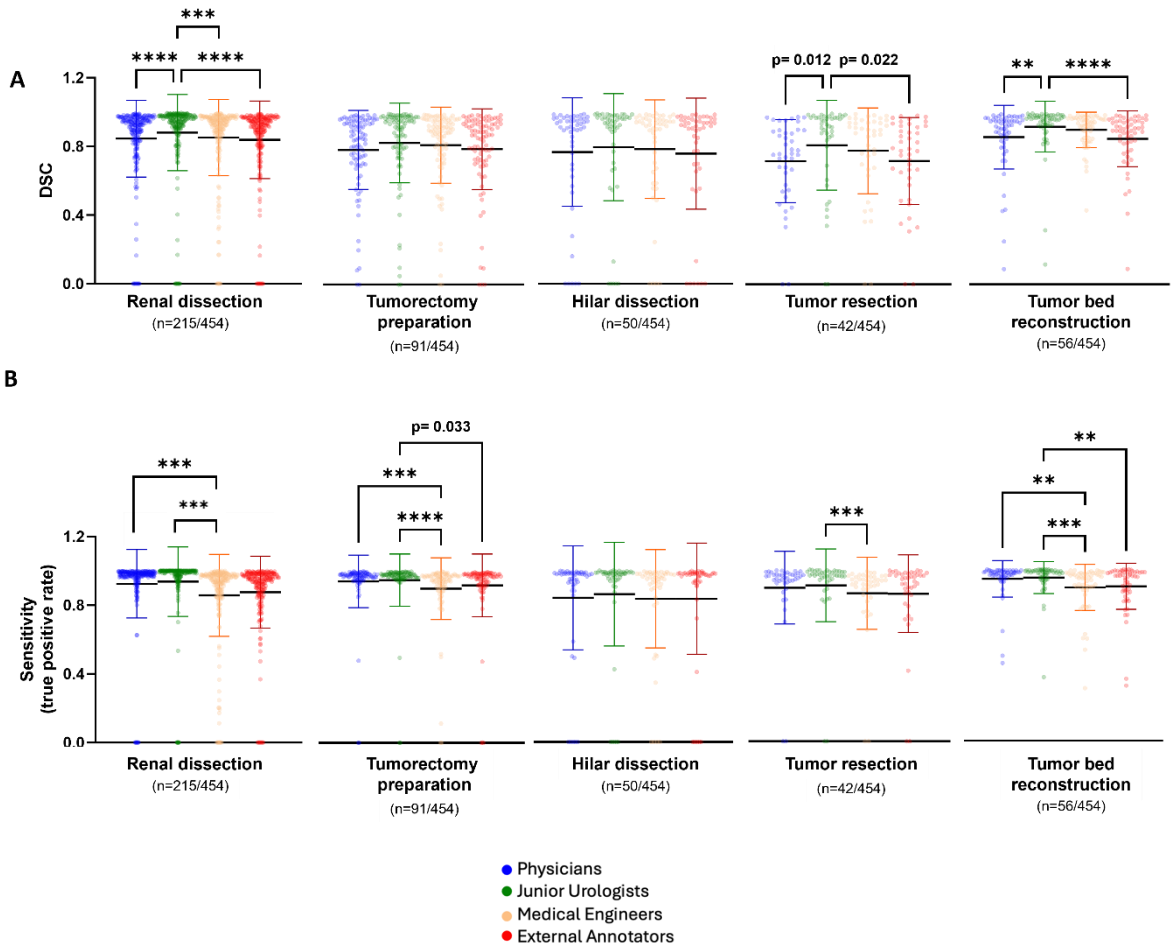

*Figure 5 - Comparison of (A) Dice Similarity Coefficient (DSC) and (B) sensitivity between annotators compared to an expert urologist's ground truth segmentation, categorized by surgical phases. Data are presented as mean ±SD. ** p < 0.01, *** p < 0.001, **** p < 0.0001.*

Time analysis showed significant variability, with external annotators taking the longest (average 4 min 24 sec ±3 min 21 sec) and medical engineers the shortest (average 40 sec ±23 sec). The high standard deviations in Supplementary Table 4 reveal significant variability in participant time spent, explaining the absence of statistical differences across groups. These findings underscore the laborious and demanding nature of the segmentation task, especially given the volume of images analyzed.

For quality control, we reviewed the external annotators' contributions. Achieving an average DSC above 0.81, their efforts facilitated a dataset spanning 131 surgeries and 48,000 images, while ensuring compliance with data security protocols. Due to the volume, full expert re-validation was unfeasible; thus, a statistically representative sample of nine surgeries (n=5,459 images) was analyzed (Table 2). We found a 68.85 ±17.37% correct segmentation rate, identifying judgment (misclassification of pixels) and selection (under- or over-segmentation) errors as main inaccuracies. A confidence-intervened threshold of 30% error was established for data return. Phase-specific success varied significantly, with rates of 95.3 ±5.96% for fat removal, 61.61 ±28.61% for tumor resection, 64.66 ±31.60% for tumor excision, and 78.14±25.86% for

reconstruction. The hilar dissection phase (499 images across three surgeries), not routinely performed, showed the lowest success rate (48.32 ±38.84%). No correlation was found between success rates and tumors' characteristics. Errors were primarily due to repeated judgment mistakes across multiple images, emphasizing the need for ongoing error correction and refinement.

The AlbuNet-34 model achieved an average DSC of 0.75 ±0.23 and sensitivity of 0.71 ±0.24 against expert standards. Figure 6 illustrates these comparisons across images of varying complexity. Utilizing a training set of 12,546 images and a validation set of 3,137 images, the model required approximately 10 hours and 24 minutes for training.

	Total number of images per surgery, n	**Segmentations judged correct by the expert reviewer, n (%)**
RAPN (n=9)	294	147 (50.0)
	608	357 (58.72)
	281	274 (97.51)
	646	336 (52.01)
	1005	910 (90.55)
	759	530 (69.83)
	618	437 (70.71)
	864	449 (51.97)
	384	301 (78.39)
Mean ±SD	607 ±251	416 ±216

Table 2 – Segmentation success rates by surgery and confidence intervals.

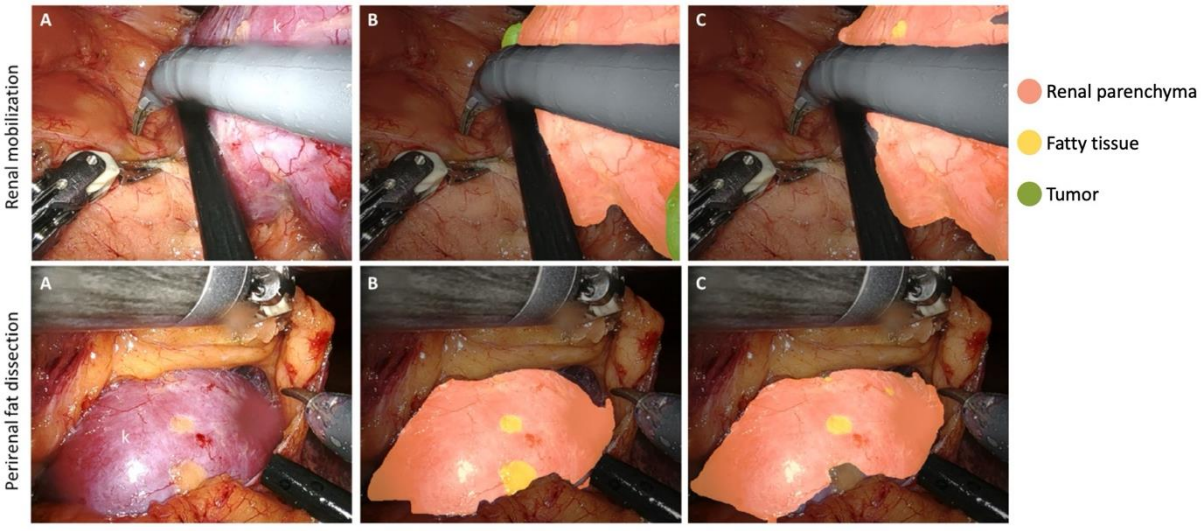

Figure 6 – Examples of parenchyma segmentation in images from robot-assisted laparoscopic partial nephrectomy, at two different stages of the surgery. A) Surgical view; B) Ground truth (expert segmentation); and C) Segmentation by the AlbuNet neural network trained from the proposed dataset.

Discussion

In this study, we described the initial stages of developing an AR system tailored for RAPN, focusing on technical workflow, dataset creation, and non-expert annotator evaluation.

Integrating AR into RAPN presents a revolutionary shift from traditional imaging, overcoming the limitations of two-dimensional modalities like ultrasound, CT, and MRI, which fall short in dynamic surgical settings. AR offers real-time, three-dimensional visualization, enhancing spatial understanding and decision-making[23]. This innovation is particularly critical in complex tasks such as endophytic tumor localization and preserving critical structures, showing AR's potential to improve surgical precision, safety, and patient outcomes.

Accurate real-time segmentation is crucial for the seamless overlay and alignment of digital models onto the surgical field. Surface-based registration plays a key role in enhancing the localization of renal surfaces and enabling image fusion without the need for physical markers. This capability is achieved through advanced stereoscopic vision and 3D camera technology, which facilitate precise spatial triangulation, and the creation of an intraoperative 3D model derived from the renal surface topography. This intraoperative model is aligned with the virtual 3D model generated from preoperative CT scans. Surgical guidance using AR benefits from the combination of detecting and tracking the renal parenchyma, and the successful implementation of these processes relies on the initial recognition of renal surfaces within intraoperative images using a trained neural network.

The development of a reliable parenchyma segmentation system relies on extensive datasets critical for accurate real-time organ segmentation. Our work highlights the time-consuming nature of manual segmentation, corroborating De Backer et al.'s findings of 1,248 hours needed to segment 15,100 frames during a study on real-time instrument delineation[24]. Utilizing non-expert annotations effectively expands the dataset, alleviating the burden on medical professionals. Inter-annotator variability remains a challenge, with DSC scores often below 0.70 for complex organs[25,26]. Additionally, recent studies underscore reduced identifiability for retroperitoneal organs, such as kidneys, versus intraperitoneal ones[27]. Despite these challenges, our study achieved acceptable DSC scores and variability, affirming the reliability of external contributions and demonstrating the feasibility of involving trained non-experts to enrich data diversity.

To optimize non-expert contributions, structured training and advanced annotation tools are critical. Providing clear guidelines and feedback loops with expert reviews enhances annotation accuracy, offering valuable learning opportunities for non-experts and improving dataset quality. Expanding the image source beyond a single expert center will increase dataset variability, making it more representative of different clinical conditions and enhancing the AR system's utility and reliability in real-world surgical environments.

Real-time performance remains challenging, as rapid data processing is crucial for continuous updates during surgery. Addressing latency[28], reducing equipment bulk, and improving user interfaces are essential steps for broader adoption and practical application in surgical environments. Precise initial model positioning on organs is crucial for accurate alignment and tumor localization but remains a challenge due to difficult-to-identify anatomical landmarks on radiological and laparoscopic images[29]. Kidney presentations vary, affecting landmark visibility. Currently, manual registration is required; however, automating this process would streamline surgical workflows, allowing surgeons to focus entirely on procedures.

Our rigid registration framework highlights the need for algorithms accommodating soft tissue deformation. Addressing this can significantly improve real-time modeling of organ dynamics, refining AR tools' precision. Additionally, our findings reveal variability in segmentation accuracy across surgical phases, highlighting the dynamic challenges of maintaining consistent AR performance. Nevertheless, similar to ultrasound, AR could be applied on-demand at critical moments to support surgical decision-making. This strategic use would enhance visualization precisely when needed, improving precision and outcomes while minimizing unnecessary cognitive or technical burdens.

Few robust studies have successfully segmented kidney structures from laparoscopic videos, underscoring the complexity of such imagery. The relatively low DSC score of our neural network, compared to CT or MRI studies[30], arises from the unique challenges of laparoscopic images, such as lower resolution, variable lighting, and motion artifacts. These challenges present opportunities for innovation in improving segmentation accuracy. Although our algorithm currently underperforms compared to human annotators, we anticipate that AlbuNet-34 will improve with time. Enhanced by data from the UroCCR network and our guidebook, which highlights annotator challenges, continuous training and additional data are expected to elevate the algorithm's performance to approach human-level accuracy.

The development of AR systems for RAPN exemplifies interdisciplinary collaboration, merging medical, engineering, and computer science expertise to enhance clinical viability and technological sophistication. Real-life trials are crucial to validate AR systems' effectiveness and ensure they translate into tangible improvements in renal oncology surgeries, culminating in better patient outcomes.

Conclusion

This study presents the initial steps in developing an AR system for RAPN, emphasizing interdisciplinary collaboration to enhance surgical precision. Despite challenges in segmentation variability and real-time processing, leveraging deep learning with structured training of non-expert annotators demonstrates promising potential. Future efforts should address model refinement through diverse datasets and validate clinical efficacy in renal oncology, with real-world trials essential for translating these innovations into improved patient outcomes.

Supplementary figures

	RAPN (*N* = 78)
Age (years), mean ±SD	61.6 ± 14.3
Gender, *N* (%)	
Male	49 (62.8%)
Female	29 (37.2%)
BMI (kg.m^2), mean ±SD	27.3 ± 5.9
Indication for NSS, *N* (%)	
Elective	54 (69.2%)
Relative	10 (12.8%)
Imperative	14 (18%)

Supplementary table 1 - Clinical characteristics of RAPN patients from our dataset.

Group	Sub-group	DSC	Sensitivity
Physicians (other than urologist)	Physician 1	0.82 ±0.24	0.92 ±0.20
	Physician 2	0.81 ±0.25	0.91 ±0.22
Junior Urologists	Junior Urologist 1	0.88 ±0.24	0.93 ±0.21
	Junior Urologist 2	0.84 ±0.23	0.92 ±0.20
Medical Engineers	Medical Engineer 1	0.84 ±0.23	0.87 ±0.22
	Medical Engineer 2	0.83 ±0.24	0.86 ±0.23
External Annotators	External Annotator 1	0.81 ±0.24	0.87 ±0.22
	External Annotator 2	0.81 ±0.25	0.92 ±0.22

Supplementary table 2 - Comparison of Dice Similarity Coefficient (DSC) and sensitivity between annotators compared to an expert urologist's ground truth segmentation. Data are presented as mean ±SD.

Surgical phase	Group	DSC	Sensitivity
Renal dissection (n=215/454)	Physicians	0.85 ±0.22	0.94 ±0.15
	Junior Urologists	0.88 ±0.22	0.95 ±0.15
	Medical Engineers	0.85 ±0.22	0.92 ±0.18
	External Annotators	0.84 ±0.23	0.90 ±0.18
	p value	*<0.0001*	*<0.001*
Tumorectomy preparation (n=91/454)	Physicians	0.79 ±0.23	0.94 ±0.11
	Junior Urologists	0.83 ±0.22	0.95 ±0.09
	Medical Engineers	0.81 ±0.22	0.89 ±0.13
	External Annotators	0.79 ±0.24	0.90 ±0.13
	p value	*0.0703*	*<0.0001*
Hilar dissection (n=50/454)	Physicians	0.77 ±0.31	0.84 ±0.30
	Junior Urologists	0.80 ±0.31	0.86 ±0.30
	Medical Engineers	0.79 ±0.29	0.83 ±0.30
	External Annotators	0.76 ±0.32	0.83 ±0.32
	p value	*0.4143*	*0.1011*
Tumor resection (n=42/454)	Physicians	0.72 ±0.24	0.94 ±0.15
	Junior Urologists	0.81 ±0.26	0.95 ±0.15
	Medical Engineers	0.77 ±0.25	0.90 ±0.18
	External Annotators	0.72 ±0.72	0.92 ±0.18
	p value	*0.0058*	*<0.0001*
Tumor bed reconstruction (n=56/454)	Physicians	0.86 ±0.19	0.89 ±0.21
	Junior Urologists	0.92 ±0.15	0.91 ±0.21
	Medical Engineers	0.90 ±0.10	0.86 ±0.21
	External Annotators	0.85 ±0.16	0.86 ±0.23
	p value	*<0.0001*	*0.0015*

Supplementary table 3 - Comparison of Dice Similarity Coefficient (DSC) and sensitivity between annotators compared to an expert urologist's ground truth segmentation, categorized by surgical phases. Data are presented as mean ±SD.

Group	Sub-group	Time (min:sec)	Time (min:sec)
Physicians (other than urologist)	Physician 1	01:18 ±00:58	01:05 ±00:57
	Physician 2	00:53 ±00:56	
Junior Urologists	Junior Urologist 1	04:24 ±03:21	03:23 ±02:25
	Junior Urologist 2	02:22 ±01:29	
Medical Engineer	Medical Engineer 1	00:31 ±00:17	00:40 ±00:23
	Medical Engineer 2	00:50 ±00:30	
External Annotators	Medical Engineer 1	03:56 ±02:07	04:16 ±02:14
	Medical Engineer 2	04:37 ±02:21	

Supplementary table 4 – Time spent for segmenting one image between annotators compared to an expert urologist's ground truth segmentation. Data are presented as mean ±SD. N=454.

Bibliography

1. Scosyrev E, Messing EM, Sylvester R, Campbell S, Van Poppel H. Renal function after nephron-sparing surgery versus radical nephrectomy: results from EORTC randomized trial 30904. *Eur Urol*. 2014;65(2):372-377. doi:10.1016/j.eururo.2013.06.044

2. Touijer K, Jacqmin D, Kavoussi LR, et al. The expanding role of partial nephrectomy: a critical analysis of indications, results, and complications. *Eur Urol*. 2010;57(2):214-222. doi:10.1016/j.eururo.2009.10.019

3. Ingels A, Bensalah K, Beauval JB, et al. Comparison of open and robotic-assisted partial nephrectomy approaches using multicentric data (UroCCR-47 study). *Sci Rep*. 2022;12(1):18981. doi:10.1038/s41598-022-22912-8

4. Peyronnet B, Seisen T, Oger E, et al. Comparison of 1800 Robotic and Open Partial Nephrectomies for Renal Tumors. *Ann Surg Oncol*. 2016;23(13):4277-4283. doi:10.1245/s10434-016-5411-0

5. Simhan J, Smaldone MC, Tsai KJ, et al. Objective measures of renal mass anatomic complexity predict rates of major complications following partial nephrectomy. *Eur Urol*. 2011;60(4):724-730. doi:10.1016/j.eururo.2011.05.030

6. Margue G, Ingels A, Bensalah K, et al. Late complications and 5 years outcomes of robotic partial nephrectomy in France: prospective assessment in the French Kidney Cancer Research Network (UroCCR 10). *World J Urol*. Published online July 6, 2023. doi:10.1007/s00345-023-04491-z

7. Klatte T, Ficarra V, Gratzke C, et al. A Literature Review of Renal Surgical Anatomy and Surgical Strategies for Partial Nephrectomy. *Eur Urol*. 2015;68(6):980-992. doi:10.1016/j.eururo.2015.04.010

8. Porpiglia F, Amparore D, Checcucci E, et al. Three-dimensional virtual imaging of renal tumours: a new tool to improve the accuracy of nephrometry scores. *BJU Int*. 2019;124(6):945-954. doi:10.1111/bju.14894

9. Shirk JD, Thiel DD, Wallen EM, et al. Effect of 3-Dimensional Virtual Reality Models for Surgical Planning of Robotic-Assisted Partial Nephrectomy on Surgical Outcomes: A Randomized Clinical Trial. *JAMA Netw Open*. 2019;2(9):e1911598. doi:10.1001/jamanetworkopen.2019.11598

10. Michiels C, Khene ZE, Prudhomme T, et al. 3D-Image guided robotic-assisted partial nephrectomy: a multi-institutional propensity score-matched analysis (UroCCR study 51). *World J Urol*. Published online April 2, 2021. doi:10.1007/s00345-021-03645-1

11. Vávra P, Roman J, Zonča P, et al. Recent Development of Augmented Reality in Surgery: A Review. *J Healthc Eng*. 2017;2017:4574172. doi:10.1155/2017/4574172

12. Porpiglia F, Checcucci E, Amparore D, et al. Three-dimensional Augmented Reality Robot-assisted Partial Nephrectomy in Case of Complex Tumours (PADUA ≥10): A New

Intraoperative Tool Overcoming the Ultrasound Guidance. *Eur Urol*. 2020;78(2):229-238. doi:10.1016/j.eururo.2019.11.024

13. Khaddad A, Bernhard JC, Margue G, et al. A survey of augmented reality methods to guide minimally invasive partial nephrectomy. *World J Urol*. Published online July 1, 2022. doi:10.1007/s00345-022-04078-0

14. Schiavina R, Bianchi L, Chessa F, et al. Augmented Reality to Guide Selective Clamping and Tumor Dissection During Robot-assisted Partial Nephrectomy: A Preliminary Experience. *Clin Genitourin Cancer*. 2021;19(3):e149-e155. doi:10.1016/j.clgc.2020.09.005

15. Chandelon K, Sharifian R, Marchand S, et al. Kidney tracking for live augmented reality in stereoscopic mini-invasive partial nephrectomy. *Comput Methods Biomech Biomed Eng Imaging Vis*. 2023;11(4):1251-1260. doi:10.1080/21681163.2022.2157750

16. Rong W, Li Z, Zhang W, Sun L. An improved Canny edge detection algorithm. In: *2014 IEEE International Conference on Mechatronics and Automation*. ; 2014:577-582. doi:10.1109/ICMA.2014.6885761

17. Lamdan Y, Wolfson HJ. Geometric Hashing: A General And Efficient Model-based Recognition Scheme. In: *[1988 Proceedings] Second International Conference on Computer Vision*. ; 1988:238-249. doi:10.1109/CCV.1988.589995

18. Kutikov A, Uzzo RG. The R.E.N.A.L. nephrometry score: a comprehensive standardized system for quantitating renal tumor size, location and depth. *J Urol*. 2009;182(3):844-853. doi:10.1016/j.juro.2009.05.035

19. Müller D, Soto-Rey I, Kramer F. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Res Notes*. 2022;15(1):210. doi:10.1186/s13104-022-06096-y

20. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Published online May 18, 2015. doi:10.48550/arXiv.1505.04597

21. Shvets AA, Iglovikov VI, Rakhlin A, Kalinin AA. Angiodysplasia Detection and Localization Using Deep Convolutional Neural Networks. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. ; 2018:612-617. doi:10.1109/ICMLA.2018.00098

22. Bukavina L, Bensalah K, Bray F, et al. Epidemiology of Renal Cell Carcinoma: 2022 Update. *Eur Urol*. 2022;82(5):529-542. doi:10.1016/j.eururo.2022.08.019

23. Madad Zadeh S, François T, Comptour A, Canis M, Bourdel N, Bartoli A. SurgAI3.8K: A Labeled Dataset of Gynecologic Organs in Laparoscopy with Application to Automatic Augmented Reality Surgical Guidance. *J Minim Invasive Gynecol*. 2023;30(5):397-405. doi:10.1016/j.jmig.2023.01.012

24. Backer PD, Praet CV, Simoens J, et al. Improving Augmented Reality Through Deep Learning: Real-time Instrument Delineation in Robotic Renal Surgery. *Eur Urol*. 2023;84(1):86-91. doi:10.1016/j.eururo.2023.02.024

25. YANG F, ZAMZMI G, ANGARA S, et al. Assessing Inter-Annotator Agreement for Medical Image Segmentation. *IEEE Access Pract Innov Open Solut*. 2023;11:21300-21312. doi:10.1109/access.2023.3249759

26. Owen D, Grammatikopoulou M, Luengo I, Stoyanov D. Automated identification of critical structures in laparoscopic cholecystectomy. *Int J Comput Assist Radiol Surg*. 2022;17(12):2173-2181. doi:10.1007/s11548-022-02771-4

27. Kolbinger FR, Rinner FM, Jenke AC, et al. Anatomy segmentation in laparoscopic surgery: comparison of machine learning and human expertise - an experimental study. *Int J Surg Lond Engl*. 2023;109(10):2962-2974. doi:10.1097/JS9.0000000000000595

28. Kumcu A, Vermeulen L, Elprama SA, et al. Effect of video lag on laparoscopic surgery: correlation between performance and usability at low latencies. *Int J Med Robot Comput Assist Surg MRCAS*. 2017;13(2). doi:10.1002/rcs.1758

29. Hattab G, Arnold M, Strenger L, et al. Kidney edge detection in laparoscopic image data for computer-assisted surgery : Kidney edge detection. *Int J Comput Assist Radiol Surg*. 2020;15(3):379-387. doi:10.1007/s11548-019-02102-0

30. Pandey M, Gupta A. A systematic review of the automatic kidney segmentation methods in abdominal images. *Biocybern Biomed Eng*. 2021;41(4):1601-1628. doi:10.1016/j.bbe.2021.10.006