APPENDIX

Supplementary Table 1. Breakdown of all the included studies by AI application (n=59)

Title	Author (year)	Type of procedure	Reported sample size	Key outcomes/endpoints	Summary
Use of AI: Annotation (n=22)					
Computer Vision and Machine-Learning Techniques for Automatic 3D Virtual Images Overlapping During Augmented Reality Guided Robotic Partial Nephrectomy	Amparore (2024)[8]	RAPN	20 patients (Group A(Computer vision-based)=12); Group B (CNN-based)=8).	1) Both AI technologies showed comparable preoperative and post-operative characteristics. 2) Average co-registration time for Computer vision was 7 (3–11) seconds; CNN was 11 (6–13) seconds. 3) No major intraoperative or postoperative complications were recorded in either technology.	This study explored two machine-learning techniques (CNN & CV) for automatically overlaying 3D virtual kidney models onto real surgical fields during RAPN to enhance precision. Both methods demonstrated feasibility, with no major complications and comparable functional outcomes.
Estimating Surgical Urethral Length on Intraoperative Robot-	Bakker (2024)[9]	RARP	84 patients were used for AI training; 30	1) The DSC of the best performing model were 0.735 and 0.755 for the catheter and	The study developed a CNN (U-Net) model to segment anatomical structures in

Assisted Prostatectomy Images Using Artificial Intelligence Anatomy Recognition			videos of different patients were used for model testing.	urethra classes, respectively, with a Hd95 of 29.27 and 72.62, respectively. 2) The predicted SUL showed a mean difference of 0.64 to 1.86 mm difference vs human annotators, but with significant deviation (standard deviation = 3.28–3.56). 3) Notably, the results also show significant interobserver variability between human annotators.	intraoperative RARP videos and predict surgical urethral length (SUL), a metric linked to postoperative urinary incontinence. The model achieved SUL predictions with small mean differences (<2 mm) compared to manual annotators.
Real-time feature tracking and segmentation in urologic robotic-assisted surgery: An artificial intelligence-based platform	Canneto (2024)[10]	RAPN and total nephrectomy (Singleport (SP) retroperitoneal partial and total nephrectomies)	15 videos of Single-Port retroperitoneal procedures.	1) Various segmentation and detection methods were assessed, calculating the area under the precision-recall curve (AUPR), see abstract for exact values. Notably, psoas muscle had the highest detection (AUPR = 0.901±0.08) and highest segmentation (AUPR = 0.741±0.091)	This study developed an artificial intelligence (AI) model to provide real-time surgical guidance and annotation during retroperitoneal SP nephrectomy videos.

NephCNN: A deep-learning framework for vessel segmentation in nephrectomy laparoscopic videos	Casella (2021)[11]	RAPN	8 RAPN patient videos (1,871 annotated frames from the Nephrec9 dataset).	1) The NephCNN model segmented renal vessels in RAPN videos with significantly better accuracy than standard U-Net baselines (DSC 71.76% vs 59.7% for 2D and 66.3% for 3D, $p < 0.05$). Precision was 0.898 and recall was 0.507 for the NephCNN model. 2) Overall performance showed the advantage of adding spatio-temporal and adversarial training.	This study proposed an AI model to kidney vessel segmentation from nephrectomy laparoscopic vision to minimize unwanted vessel resection. The proposed approach could be a valuable solution with a view to assisting surgeons during RAPN.
Three-dimensional automatic artificial intelligence-driven augmented-reality selective biopsy during nerve-sparing robot-assisted radical prostatectomy: A feasibility and accuracy study	Checcucci (2023)[12]	RARP	The data consists of 34 patients (pT2: 18; pT3: 16).	1) Tumour was correctly identified in 87.5% of pT3 patients. 2) Positive surgical margins were low (0% in pT2, 7.1% in pT3, <3 mm). 3) Potency recovery at 12 months was 70.6% and continence 94.1%. No biochemical recurrence at 12	The study aimed to develop an AI-driven 3D automatic augmented reality (AAR) system for real-time projection of prostate and tumour models during nerve-sparing RARP. The group trained two CNNs to localize a catheter and align 3D models automatically onto intraoperative views. The system guided the selective

				months.	biopsy of the NVB.
Improving Augmented Reality Through Deep Learning: Real-time Instrument Delineation in Robotic Renal Surgery	DeBacker (2023)[13]	RAPN, kidney transplantation (RAKT), and kidney autotransplantation (RAKAT)	10 patients, categorized as 8 RAPN, 1 RAKT, 1 RAKAT (15100 frames).	1) The model was able to distinguish instruments from soft tissue background across unseen test videos (IoU:94.4%; Dice:97.1%). 2) The model successfully enhanced visualization during RAPN and transplantation cases, although a 0.5s latency in time was noted.	This study used a deep learning-based system for real-time detection and segmentation of surgical instruments during AR-guided robotic renal procedures. The setup runs on a standalone laptop and was deployed in three different hospitals, used by four different surgeons. The group acknowledged a 0.5-second latency and discussed that instrument detection is a simple and feasible way to enhance the safety of AR-guided surgery.
Deep Learning Model for Real time Semantic Segmentation During Intraoperative Robotic Prostatectomy	GonPark (2024)[14]	RARP	120 RARP videos (2400 images).	1) The mean Dice scores for the identification of the instruments, bladder, prostate, and seminal vesicle–vas deferens were 0.96, 0.74, 0.85, and	This study developed a deep learning model (Reinforcement U-Net) for real-time semantic segmentation during RARP. Specifically, the instruments, bladder, prostate, and seminal

				0.84, respectively. 2) Overall, when applied to the test data, the model had a mean Dice coefficient value of 0.85, IoU of 0.77, and accuracy of 0.85.	vesicle-vas deferens.
A spatio-temporal network for video semantic segmentation in surgical videos	Grammatikopo-ulou (2024)[15]	RAPN	137 RAPN videos (53,000 images).	1) 1.04–1.3% increase of the mean Intersection over Union (IoU), when adding a temporal encoder for RAPN	This study aims to use a spatial-temporal model to improve the state-of-the-art static models in segmentation. The researchers presented a model that improved per-frame segmentation output and video temporal consistency in the context of RAPN and cholecystectomy.
Using spatial-temporal ensembles of convolutional neural networks for lumen segmentation in ureteroscopy	Lazo (2021)[16]	Urinary endoscopy (Ureteroscopy)	11 ureteroscopy videos from 6 patients (2673 annotated frames).	1) The final model achieved a dice similarity coefficient of 0.80 for lumen segmentation, which outperformed framed-based U-NET and Mask RCNN models.	This study aimed to annotate the hollow lumen in ureteroscopy videos to indicate the passage the instrument should follow. While inference time is longer, the model holds promise for intraoperative guidance and safer navigation during endoscopic procedures.

UO-YOLO: Ureteral Orifice Detection Network Based on YOLO and Biformer Attention Mechanism	Liang (2024)[17]	Urinary endoscopy(Endoscopy procedures requiring ureteral orifice localization)	104 cases total (84 training with 820 images; 20 validations with 223 images; total 1043 images).	1) The UO-YOLO model detected ureteral orifices in cystoscopy images with high accuracy, reaching a mean precision of 0.928, a recall of 0.756, and an average precision, mAP (overlap threshold of 0.5), of 0.896. 2) Achieved a single-frame detection speed of 5.7 ms on the platform, with a frame rate of 175FPS.	This paper aimed to enhance the adaptability of target detection networks in the clinical conditions of ureteroscopes. The group developed an innovative deep learning-based ureteric orifice detection system that offers good real-time performance by analyzing and optimizing the existing YOLO series networks.
A real-time system using deep learning to detect and track ureteral orifices during urinary endoscopy	Liu (2021)[18]	Urinary endoscopy (ureteroscopy and resectoscopy)	51 patients (447 images from resectoscopy for training; 818 images from ureteroscopy for testing; also contained video frames from 19 resectoscopy + 32 ureteroscopy sequences).	1) When tested on 818 ureteroscopy images, the AI achieved an accuracy of 0.902. 2) In the video sequence analysis, the AI yielded test sensitivities of 0.840 and 0.922 on resectoscopy and ureteroscopy video frames, respectively.	This study aimed to develop an AI model to identify the UO as a key landmark in endoscopic procedures. Researchers noted it has great potential to serve as an excellent learning and feedback system for trainees and new urologists in clinical settings.

Automated Upper Tract Urothelial Carcinoma Tumor Segmentation During Ureteroscopy Using Computer Vision Techniques	Lu (2024)[19]	Urinary endoscopy (Endoscopic ureteroscopy with tumor identification)	20 videos (3387 annotated frames).	1) The model U-Net achieved strong tumour segmentation (AUC-ROC: 0.96, DSC: 0.78, pixel accuracy: 0.98, sensitivity: 0.76, specificity: 0.99). 2) For ablated tissue, performance remained robust (AUC-ROC: 0.90, DSC: 0.50, accuracy: 0.87).	The study aimed to develop a computer vision model for real-time, automated segmentation of UTUC tumours to augment visualization during treatment. The group's U-Net model had the best performance overall. The group also created a model that can process real-time video feeds and overlay model predictions intraoperatively.
Artificial Intelligence-Based Intraoperative Guidance: Demonstration of Real-Time Segmentation and Feature Tracking in Robotic Surgery	Morgantini (2024)[20]	RAPN and total nephrectomy (Singleport (SP) retroperitoneal partial and total nephrectomies)	Not reported.	1) The AI system showed real-time segmentation and tracking of key anatomical structures during single-port retroperitoneal robotic surgery (only qualitative outcomes).	This study provides an approach for real-time tracking and segmenting key anatomical structures during single-port retroperitoneal robotic surgery. Only qualitative outcomes are provided.

A deep learning framework for real-time 3D model registration in robot-assisted laparoscopic surgery	Padovan (2022)[21]	RARP and RAPN	9 endoscopic videos (5 prostate and 4 kidney).	1) The AI system segmented catheters, prostates, and kidneys with relatively good accuracy (IoU: 0.95 for catheters, 0.73 for prostates, 0.86 for kidneys). 2) Next, the rotation estimation was highly accurate on synthetic datasets (≥0.99 within ±5°), indicating that the network learnt.	This study developed an AI model for segmentation of RARP and RAPN and superimposed 3D models for AR surgical guidance. Case studies include catheter-guided biopsy and nerve-sparing during RARP, and kidney tumor demarcation in RAPN.
Application of deep learning for semantic segmentation in robotic prostatectomy: Comparison of convolutional neural networks and visual transformers	Pak (2024)[22]	RARP	150 patients with prostate cancer who underwent RARP (3,000 ground truth images from videos. 2,400 images were used for training and 600 for validation test data).	1) The CNNs achieved a higher Dice scores (0.930-0.44) compared to visual transformers (0.916-0.940) which shows that CNNs were better in terms of prostate, vas deferens, and seminal vesicle segmentation. 2) Surgical instruments and bladder structure were both segmented with ~97-98% accuracy across the different models. However, authors noted the CNNs	This study aimed to compare CNN and vision transformer segmentation models for RARP. They found that both CNN and transformer models showed reliable predictions in the segmentation task. CNN models are deemed more suitable than transformer models in unusual cases.

				outperformed transformers in small or complex structure segmentation.	
Real-Time Detection of Ureteral Orifice in Urinary Endoscopy Videos Based on Deep Learning	Peng (2019)[23]	Urinary endoscopy (ureteroscopy and resectoscopy)	129 patients (training: 447 resectoscopy images from 51 patients; testing: 418 ureteroscopy images + 400 no ureteral orfice images).	1) The annotation on still images achieved precision of 0.876, recall of 0.813, specificity of 0.891 (F1=0.844, F2=0.825). 2) On ureteroscopy video datasets, the recall outcome was 0.845 and the precision was 0.851. 3) On resectoscopy datasets, the recall was 0.78 and the specificity was 0.83.	This study aimed to assess AI models ability to identify urinary orifice in urological endoscopic procedures. The UO detection system in this study can identify and locate UOs of ureteroscopy and resectoscopy in real time with average processing time equal to 25 ms per frame and simultaneously achieve satisfactory recall and specificity.
Artificial Intelligence in Minimally Invasive Adrenalectomy: Using Deep Learning to Identify the Left Adrenal Vein	Sengun (2023)[24]	Minimally invasive adrenalectomy	40 videos (2000 frames were annotated).	1) The AI model segmented the left adrenal vein with strong accuracy, achieving a mean Dice score of 0.77 (±0.16) and sensitivity of 0.82 (±0.15), with maximum Dice reaching 0.93. 2) Across patient datasets, performance was consistent	This study explored the use of artificial intelligence (AI) and deep learning to identify the left adrenal vein (LAV) during minimally invasive laparoscopic adrenalectomy. The group highlighted the model's high performance and the potential to be utilized to

				(IoU up to 0.66, PPV 0.79), showing that deep learning can reliably aid recognition of this critical vessel during adrenalectomy.	identify critical anatomy during adrenal surgery intraoperatively.
Utilization of artificial intelligence in minimally invasive right adrenalectomy: recognition of anatomical landmarks with deep learning	Sengun (2024)[25]	Minimally invasive adrenalectomy	20 intraoperative videos (1000 extracted frames; 800 for training, 200 for validation using 5-fold cross-validation).	1) Incorporation of Dice-Cross Entropy Loss with SwinUNETR(transformer-based model) achieved 78.37%, whereas the MedNeXt(CNN-based model) reached a 77.09% mDSC score. Conversely, MedNeXt reaches a higher mIoU score of 63.71% than SwinUNETR by 62.10% on a three-region prediction task. 2) For pixelwise classification, MedNeXt with Dice-Cross Entropy Loss showed higher accuracy (88.54%), sensitivity (77.08%), and specificity (92.36%).	This study developed and compared deep learning models that can identify critical anatomical structures during minimally invasive right adrenalectomy. The group highlighted the potential to be utilized to identify critical anatomy during adrenal surgery intraoperatively.

Deep Learning-Based Seminal Vesicle and Vas Deferens Recognition in the Posterior Approach of Robot-Assisted Radical Prostatectomy	Takeshita (2023)[26]	RARP	26 patient videos which provided 1040 annotated images (520 seminal vesicles and vas deferens annotated image and 520 images not containing seminal vesicles and vas deferens).	1) The CNN had overall moderate segmentation accuracy (Dice Similarity Coefficient: 0.73; precision: 0.74, recall: 0.73). 2) The six urological trainees, the videos with CNN-generated masks allowed for faster recognition of seminal vesicle and vas deferens compared to original videos alone ($p<0.001$).	The study aimed to use AI to recognize the seminal vesicle and vas deferens (SV-VD) in the posterior approach of RARP and assess the performance of the convolutional neural network model under clinically relevant conditions. Researchers showed that the AI model helped the inexperienced urologists recognize the SV-VD faster.
Real-time deep learning semantic segmentation during intra-operative surgery for 3D augmented reality assistance	Tanzi (2021)[27]	RARP	5 intraoperative endoscopic videos (~15,570 frames).	1) Segmented the catheter in real time during RARP with high accuracy (IoU 0.894 ± 0.076), outperforming a prior computer vision method (0.339 ± 0.195). 2) It reduced overlay errors, cutting anchor point distance from 12.6 mm to 4.16 mm and rotation error from 0.266 to 0.169.	This study developed an AI model for the semantic segmentation and prostate overlay (AR) during the targeted biopsy phase of RARP. The authors compared multiple encoder-decoder CNN architectures (U-Net, SegNet, PSPNet) paired with different backbones (MobileNet, ResNet, VGG) and selected U-Net + MobileNet for its

					superior real-time performance.
A markerless automatic deformable registration framework for augmented reality navigation of laparoscopy partial nephrectomy	Zhang (2019)[28]	Laparoscopic partial nephrectomy	Phantom kidney models with CT imaging + in vivo validation on 1 patient video.	1) The Mask R-CNN model automatically aligned preoperative CT kidney and tumor models with intraoperative laparoscopic views, achieving high segmentation accuracy (94.9%) and low registration error (TRE 1.28 ± 0.68 mm). 2) In phantom and one in vivo case, the augmented reality overlay showed precise alignment, demonstrating feasibility of real-time AR navigation during partial nephrectomy.	This study proposed a markerless augmented reality (AR) framework to guide laparoscopic partial nephrectomy by automatically aligning preoperative 3D models with intraoperative video. Experimental results have demonstrated that the proposed framework could accurately overlay comprehensive preoperative models on deformable soft organs automatically.
A novel 3D deep learning model to automatically demonstrate renal artery segmentation and its validation in	Zhang (2022)[29]	Laparoscopic partial nephrectomy	6 pigs (11 kidneys) and 131 patients with cT1a tumors.	1) The AI-generated Tridimensional Kidney Perfusion model predicted renal artery territories with high accuracy, achieving Dice coefficients of 0.82 in animal	This study aimed to demonstrate renal artery segmentation with an AI model and verified the precision and feasibility during laparoscopic partial nephrectomy. They

nephron-sparing surgery				experiments and 0.81 (0.72–0.94) in clinical validation against ICG fluorescence. 2) In 131 laparoscopic partial nephrectomy cases, predicted perfusion regions closely matched intraoperative findings, with safe operative outcomes including mean warm ischemia time of 27.0 minutes and blood loss of 110 ml.	found that the AI can automatically present the renal artery trees and precisely delineate the perfusion regions of different segmental arteries.
Use of AI: Surgical Feedback (n=20)					
Spectral imaging enables contrast agent–free real-time ischemia monitoring in laparoscopic surgery	Ayala (2023)[30]	Laparoscopic partial nephrectomy	10 human patients undergoing partial nephrectomy.	1) In almost all patients, the data corresponding to ischemic tissue could be perfectly separated from those corresponding to perfused tissue. This led to a median/mean area under the receiver operating curve (AU-ROC) of 1.0/0.9 obtained for patients. 2) However, the model failed	This study introduced a real-time deep learning framework using multispectral imaging (MSI) and invertible neural networks (INNs) for intraoperative ischemia monitoring during laparoscopic partial nephrectomy. Unlike conventional methods using contrast agents like ICG, the authors suggested this

				in 1 case due to unusual tissue properties.	technique enables contrast-free, repeatable, and continuous monitoring by treating ischemia detection as an out-of-distribution (OoD) problem.
Is AI Seeing What I'm Seeing? Validation of Touch Surgery's Enterprise System in Calculating Warm Ischemia Time during Robotic-Assisted Partial Nephrectomy	Bobrowski (2024)[31]	RAPN	32 patients who underwent RAPN.	1) Compared to the reference standard, the mean difference in WIT computed by TSE and as dictated in operative note was 68.6 sec (SD=159.3 sec) and 120.0 sec (SD=213.3 sec), respectively. 2) There were 4 cases where TSE computed WIT and 3 cases where the dictated WIT was>5 minutes off the reference standard.	The study used a commercial platform that estimated warm ischemic time in RAPN. They demonstrated that the AI can provide WIT more accurately than surgeon documented times during RAPN, though variant cases exist.
Artificial Intelligence alert systems during robotic surgery: a new potential tool to improve the safety of the intervention	Checcucci (2023)[32]	RARP	10 RARP videos.	1) true positive 98% (identification of bleeding); 2) false positive 3% (false alert in the absence of bleeding) 3) anticipation, represented	This study describes an AI model that predicted the occurrence of intraoperative bleeding in RARP and warned the surgeon of bleeding risk. This helps surgeons/trainees identify bleeding events.

				by increasing % before the possible occurrence of the event	
Development of Bleeding Artificial Intelligence Detector (BLAIR) System for Robotic Radical Prostatectomy	Checcucci (2023)[33]	RARP	10 RARP videos.	1) The MTL-CNN demonstrated event recognition accuracy of 90.63%. 2) The density plot described a similarity between BLAIR and human assessments.	This study is a prospective investigation involving the development and validation of the Bleeding Artificial Intelligence Detector (BLAIR) system. This helps surgeons/trainees identify bleeding events.
Automated intra-operative video analysis in robot-assisted partial nephrectomy: Paving the road for surgical automation	DeBacker (2023)[34]	RAPN	79 RAPN videos.	1) Tissue detection: 97.0% specificity, 59.9% sensitivity and 66.1% precision. 2) Phase detection: 97.7% specificity, 58.8% sensitivity and 77.4% precision.	This study presents an AI system for automated tissue and phase detection in robot-assisted partial nephrectomy (RAPN), using 79 videos annotated at pixel-level (48,371 organs) and phase-labelled (13 phases). The authors saw the benefit of such AI training/teaching.
Automated surgical step recognition in transurethral bladder tumor resection using	Deol (2024)[35]	TURBT	A total of 300 full-length TURBT videos (179 videos	1) AI video analysis achieved an accuracy of 89.6%. 2) Model accuracy was	This study describes using AI to facilitate surgical step recognition in TURBT

artificial intelligence: transfer learning across surgical modalities			served as a training dataset for algorithm development, 44 for internal validation, and 77 as a separate test cohort).	highest for the primary endoscopic evaluation step (98.2%) and lowest for the surface coagulation step (82.7%).	procedures.
Artificial intelligence model for automated surgical instrument detection and counting: an experimental proof-of-concept study	Deol (2024)[36]	General urology	1,004 images containing 13,213 surgical tools across 11 categories.	1) The model demonstrated high precision (98.5%) and recall (99.9%) in distinguishing surgical tools from the background. 2) It also exhibited excellent performance in differentiating between various surgical tools (precision ranging from 94.0 to 100% and recall ranging from 97.1 to 100% across 11 tool categories).	This study describes the feasibility of using a deep learning-based computer vision model for automated surgical tool detection and counting.
The Film Room: Using Artificial Intelligence to Facilitate Video Review	Henning (2024)[37]	General urology	Not reported.	1) Created a user-friendly, protected surgical video forum using a surgical intelligence platform, leading	This group describes the implementation of an AI-assisted "Film Room" video review platform at Mayo

for Urology Trainees				to considerable uptake and usage across Mayo Clinic's urology training program.	Clinic's urology department to enhance surgical training through automated video annotation and structured feedback. The system, powered by Theator Inc.'s surgical intelligence platform, automatically captures, indexes, and annotates laparoscopic, robotic, and endoscopic procedure videos, enabling trainees to search, comment on, and review specific surgical steps.
A warning system for urolithiasis via retrograde intrarenal surgery using machine learning: an experimental study	Jeong (2022)[38]	Urinary endoscopy (Ureteroscopy Retrograde Intrarenal Surgery)	~20,000 data subsets for training and validation.	1) The model was able to distinguish laser-stone/-tissue/ and in idle (accuracy>95%; AUC >0.98).	This study proposes a novel machine learning-based warning system for retrograde intrarenal surgery that detects laser-tissue vs. laser-stone interactions based on vibration signals. Using a simulated surgical model (raw chicken filled with water and human stones), the authors collected 18 hours of signal data via an accelerometer mounted on the

					endoscope. The model was able to rapidly recognize (in 0.5 s) the current laser exposure state with high accuracy.
Accuracy of warm ischemia time measurement using a surgical intelligence software in partial nephrectomies: A validation study	Khandekar (2024)[39]	RAPN	61 RAPN cases.	1) For all procedures, AI-computed WITs were within 1 min of the ground truth within 30 s in 97%, and within 10 s in over 80%. 2) The difference between AI-computed WIT and ground truth was 8.3 s, lower than the 2.45 min difference found in operative reports ($p < 0.001$).	The study compares the accuracy of warm ischemia times (WITs) derived by a surgical artificial intelligence (AI) software to those documented in surgeon operative reports during partial nephrectomy procedures
Automated Operative Reports for Robotic Radical Prostatectomy Using an Artificial Intelligence Platform	Khanna (2023)[40]	RARP	117 RARP cases.	1) AI reproduces major components of the human operative report with 91.5% accuracy (107 out of 117) 2) Out of 9 relevant discrepancies, experts review 3 discrepancies in the human report (2.6%); 6 discrepancies in the AI report (5.1%).	This study demonstrates a fully automated creation of surgical operative reports using AI.

Artificial Intelligence-Enabled Automated Identification of Key Steps in Robotic-Assisted Radical Prostatectomy	Khanna (2023)[41]	RARP	107 full-length RARP videos (70: training dataset; 14: internal validation; 23 evaluating algorithm accuracy).	1) Concordance between AI-enabled automated video analysis and manual human video annotation was 87.6%. 2) Algorithm accuracy was highest for the VUA step (98.6%) and lowest for the final inspection and extraction step (63.0%).	This study describes an AI-enabled computer vision algorithm for automated phase recognition of full-length RARP surgical video.
Automated operative reports for robotic radical prostatectomy using artificial intelligence	Khanna (2024)[42]	RARP	158 RARP cases.	1) Compared to ground-truth, clinically significant discrepancies were observed in 43 (27.2%) operative reports written by surgeons, as compared to 20 (12.7%) operative reports generated by AI (p=0.001).	This study demonstrates an automated creation of surgical operative reports using AI. AI-generated reports achieved higher accuracy than those written by surgeons.
Automated Identification of Key Steps in Robotic-Assisted Radical Prostatectomy Using Artificial Intelligence	Khanna (2024)[43]	RARP	474 RARP videos.	1) The AI algorithm achieved 92.8% overall accuracy compared to human annotation for the full-length video. 2) AI performance was highest for the VUA step	This study developed an AI tool to automate the identification of surgical steps in RARP

				(97.3%) and lowest for the final inspection step (76.8%).	
Initial Experience With a Fully Integrated Artificial Intelligence Platform During Minimally Invasive Surgery	Khanna (2024)[44]	RARP, endoscopic BPH procedures, RAPN, TURBT	766 RARP videos,395 endoscopic BPH procedures, 132 RAPN videos, and 29 TURBT videos.	1) The platform labelled steps, benchmarking efficiency of surgical steps across surgeons, and cataloging trends in surgical step duration.	This study describes a new comprehensive integration of a novel AI computer vision platform into real-world surgical practice in the United States
Human AI collaboration for unsupervised categorization of live surgical feedback	Kocielnik (2024)[45]	General urology	3740 instances (Trainer utterance intended to modify trainee thinking or behavior).	1) Discovered 20 AI topics capturing various aspects of surgical feedback. See the original paper for statistics. 2) Notably, feedback on “Handling Bleeding” is linked to improved learner behavioural change.	This study introduces a Human-AI Collaborative Refinement Process to categorize live surgical feedback during robot-assisted procedures, using unsupervised topic modelling (BERTopic) combined with clinician input.
Artificial Intelligence-Based Surgical Phase Recognition in Robot-Assisted Radical Prostatectomy and	Konnai (2024)[46]	RARP	102 RARP(81 from one surgeon, 21 from five other surgeons)	1) For videos of the same surgeon, out of nine phases, precision exceeded 90% in five phases and recall exceeded 90% in eight	This study aimed to develop an AI to identify phases of RARP, and was able to do so with different surgeons.

Cross-Surgeon External Validity Verification			65/81 cases were used for AI development; the remaining 16 and 21 cases from the first and other surgeons, respectively, validated the AI's accuracy and external validity.	phases, with an overall precision of 94% and recall of 93%. 2) For analyzing videos from different surgeons, at least 80% precision was achieved in five phases and at least 80% recall in five phases, with both overall precision and recall reaching 83%.	
Deep Learning Automated Phase Recognition in Robot-Assisted Partial Nephrectomy	Mezzina (2024)[47]	RAPN	86 RAPN videos performed on Intuitive Xi systems across two tertiary centers (62 training, 12 validation, and 12 testing).	1) Baseline CNN helped to recognize surgical phases with modest accuracy (accuracy: 63.4%± 7.5%). 2) CNN with post-processing filtering showed improved predication accuracy (accuracy: 67.8%± 9.9%). 3) CNN+ LSTM Temporal modeling further improved phase recognition (accuracy: 68.5%± 6.9%).	This study presented results for AI phase recognition of RAPN and demonstrated that adding a temporal component to the CNN improves phase prediction accuracy.

				4) CNN+LSTM with post-processing achieved the best performance and ability to recognize phases while minimizing errors (accuracy: 74.0%± 7.1%, precision: 77.4%± 8.2%, recall: 58.8%± 8.8%, specificity: 97.7%± 0.6%).	
AI Navigation System for Living Donor Surgery: Seven Process Classification of Laparoscopic Nephrectomy by Deep Learning	Nishihara (2020)[48]	Laparoscopic nephrectomy	10 laparoscopic donor nephrectomy videos (6 training, 2 validation, and 2 test).	1) The model accurately classified surgical processes across seven steps (AUC values are P0 (0.951), P1 (0.879), P2 (0.914), P3 (0.957), P4 (0.949), P5 (0.995), P6 (0.951) for the six steps of the procedure. Average AUC across all steps was 0.942.	This study aimed to use AI to recognize phases of laparoscopic donor nephrectomy. Authors noted high accuracy out of a small number of training datasets might be associated with little anatomical and procedural variations of the laparoscopic donor nephrectomies.

AI-powered real-time annotations during urologic surgery: The future of training and quality metrics	Zuluaga (2023)[49]	RAPN and RARP	2 live-broadcasted cases; AI model was trained on 7768 annotated videos for development.	1) The model was able to annotate key surgical steps in RARP include preparation, adhesiolysis, Retzius space dissection, anterior bladder neck transection, posterior bladder neck transection, seminal vesicles and posterior dissection, lateral and apical dissection, and urethral transsection, vesico-urethral anastomosis, and final inspection and extraction. 2) The system is also able to identify and annotation steps of RAPN including preparation, adhesiolysis, dissection and mobilization, lesion resection, resection site closure, and final inspection. 3) Annotations were generated with <1 second latency, demonstrating stable performance throughout both cases.	This study assessed real-time intra-operative segmentation on RARP and RAPN with the commercial AI platform. They showed low latency in real-time analysis, but acknowledged that the sample size was limited.

Use of AI: Evaluation (n=17)					
Artificial Intelligence-Supported Video Analysis as a Means to Assess the Impact of DROP-IN Image Guidance on Robotic Surgeons: Radioguided Sentinel Lymph Node versus PSMA-Targeted Prostate Cancer Surgery	Azargoshasb (2023)[50]	RARP (Radioguided prostate cancer (PCa) surgery)	44 patients with Prostate cancer (Group Sentinel Lymph Node=25; Group Prostate-Specific Membrane Antigen=19).	1) Kinematics assessment reveal that the challenges encounter during PSMA-targeted procedures converted to longer target identification times and increase in probe pick-ups (both five-fold; $p < 0.001$), reducing the decision making score ($P < 0.001$).	This study used AI to assess the impact of DROP-IN gamma probe-guided image navigation on surgical decision-making during sentinel lymph node (SLN) and PSMA-targeted prostate cancer surgeries. A neural network was trained to track the probe's kinematics from videos, extracting metrics like search time, probe pickups, and decision-making scores.
A computer vision technique for automated assessment of surgical performance using surgeons' console-feed videos	Baghdadi (2019)[51]	RARC (pelvic lymph node dissection)	20 videos (14 to develop AI, 6 to test model).	1) The accuracy of predicting the expert-based PLACE (Pelvic Lymphadenectomy Assessment and Completion Evaluation) scores was 83.3% from logistic regression analysis.	This study developed AASP, an automated system to assess PLND adequacy during RARC using console-feed videos. Features like nerve/vessel counts, cleared tissue area, color dynamics, and image entropy (disorganization) were analyzed.

Capturing relationships between suturing sub-skills to improve automatic suturing assessment	Cui (2024)[52]	RARP	156 VR videos from 43 residents, fellows, and attending urologic surgeons in a 5-center multi-institutional study.	1) AI-based Joint skill assessment helps improve skill assessment performance.	The study uses EASE score, an AI assessment tool considering kinetic data from videos, to assess gestures of urological procedures. Instead of looking at subskills in isolation, the authors focused on the relationships of sub-skills in combination. This is proposed to be useful for surgical evaluation and training.
Using Surgical Gestures to Build Explainable Artificial Intelligence for Surgical Skills Assessment	Goldenberg (2023)[53]	RARP (nerve-sparing (NS) step)	80 RARP cases from 21 surgeons.	1) The neural network was able to predict DART (the validated Dissection Assessment for Robotic Technique) domains 2) AUCs for predicting tissue handling, tissue retraction, and efficiency were 0.64, 0.66, and 0.64, respectively	This study utilized surgical gestures-trained AI algorithms to predict validated measures of technical skill to improve transparency.
Utilizing machine learning and automated performance metrics to evaluate robot-assisted	Hung (2018)[54]	RARP	78 RARP cases from 9 surgeons.	1) The best algorithm reached 87.2% accuracy in predicting LOS (73 cases as “pExp-LOS” and 5 cases as “pExt-	This study presented a novel machine learning method of processing automated performance metrics to evaluate surgical performance

radical prostatectomy performance and predict outcomes.				LOS"). 2) Patient outcomes predicted by the algorithm had a significant association with the "ground truth" in surgery time ($p < 0.001$, $r = 0.73$), LOS ($p = 0.05$, $r = 0.52$), and Foley duration ($p < 0.001$, $r = 0.45$).	and predict clinical outcomes after RARP. The authors showed that APMs and ML algorithms may help assess surgical RARP performance and predict clinical outcomes.
A deep-learning model using automated performance metrics and clinical features to predict urinary continence recovery after robot-assisted radical prostatectomy	Hung (2019)[55]	RARP	100 patients who underwent RARP was utilized as the training cohort.	1) DeepSurv achieved the best prediction (C-index: 0.599; MAE: 85.9 days), which outperforms the Cox regression and Random Survival Forests. 2)Surgeons with more efficient APMs also achieved significantly higher continence rates at 3 and 6 months (47.5% vs 36.7%, P = 0.034; 68.3% vs 59.2%, P = 0.047), though rates equalized by 12 months.	This study addressed the potential of using an AI model along with automated performance metrics to determine patient outcomes. Using APMs and clinicopathological data, the DeepSurv DL model in the study was able to predict continence after RARP.

A vision transformer for decoding surgeon activity from surgical videos	Kiyasseh (2023)[56]	RARP and RAPN	78 surgical procedures at the University of Southern California (USC), 60 at St. Antonius Hospital (SAH), and 20 at Houston Methodist Hospital (HMH).	1) SAIS reliably decodes surgical subphases with an area under the receiver operating characteristic curve (AUC) of 0.925, 0.945 and 0.951, for needle driving, needle handling and needle withdrawal, respectively. 2) SAIS reliably decodes the skill level of surgical activity, achieving an AUC of 0.849 and 0.821 for the needle handling and driving activity, respectively. 3) SAIS can provide surgeons with actionable feedback, and assessments are linked to patient outcomes.	This study describes the development and feasibility of SAIS, a surgical assessment tool that is verified across medical centres and continents. Also served as a basis for subsequent studies below.
A multi-institutional study using artificial intelligence to provide reliable and fair feedback to surgeons	Kiyasseh (2023)[57]	RARP (VUA)	78 surgical procedures at the University of Southern California (USC), 60 at St. Antonius	1) SAIS generates explanations that often align with human explanations (AUPRC = 0.488 for needle handling and 0.428 for needle driving) 2) SAIS exhibits explanation	The study quantifies the reliability of SAIS(AI assessment tool) explanations on surgical videos from three hospitals across two continents by comparing them to explanations generated by

			Hospital (SAH), and 20 at Houston Methodist Hospital (HMH). Additionally, 69 videos of medical students performing suturing activities in a training environment were included.	bias against surgeons(AUPRC ≈ 0.54 for prostate volumes ≤ 49 ml, AUPRC ≈ 0.47 for prostate volumes > 49 ml) 3) TWIX improves the reliability of AI-based explanations across hospitals (AUPRC increases at multiple sites)	human experts. Researchers used the precision-recall curve (AUPRC) as a measure of the reliability of AI-based explanations. They ultimately found TWIX to be useful in improving reliability.
Human visual explanations mitigate bias in AI-based assessment of surgeon skills	Kiyasseh (2023)[58]	RARP (VUA)	78 surgical procedures at the University of Southern California (USC), 60 at St. Antonius Hospital (SAH), and 20 at Houston	1) SAIS showed underskilling bias (NPV ≈ 0.71 and 0.75 for the prostate volumes ≤49 ml and >49 ml, respectively at USC) 2) TWIX mitigates underskilling bias (NPV increases at multiple sites)	The study examines the bias with the SAIS(AI assessment tool) for surgery, and the potential of eliminating the bias. NPV(negative predictive value) reflects bias. They ultimately found TWIX to be useful in mitigating bias. They found that this improvement is also seen in the training

			Methodist Hospital (HMH), totaling 158 procedures. Additionally, 69 videos of medical students performing suturing activities in a training environment were included.		environment, where medical students' skills were assessed.
Deep learning-based computer vision to recognize and classify suturing gestures in robot-assisted surgery	Luongo (2021)[59]	RARP (VUA)	2395 video clips for identification and 511 videos for classification training.	1) For the Identification task (accuracy of 79%: AUC:0.88), the model is able to differentiate needle driving vs non-needle driving gestures. 2) For the classification task (average accuracy of 62%; AUC:0.87), the model is able to identify key suture types with different gesture classes.	The study demonstrated computer vision's ability to recognize the action of suturing and distinguish between different classifications of suturing gestures. This is useful for future implementation/refinement of automated VUA skill assessment.

Artificial Intelligence-Based Video Feedback to Improve Novice Performance on Robotic Suturing Skills: A Pilot Study	Ma (2024)[60]	RARP (VUA)	42 participants (medical/undergraduate students with no robotic experience; randomized 22 feedback vs 20 control).	1) The feedback group improved much more in needle handling compared to controls (Δ0.30 vs –0.02, p = 0.018), while gains in needle driving were present but not statistically significant (Δ0.17 vs –0.40, p = 0.074). 2) Underperformers showed the biggest benefit, with a large improvement in needle handling when given AI feedback versus almost no change in controls	This study utilized AI in combination with previously established criteria (EASE) to classify performance as well as provide structured feedback. In the study, AI-based feedback facilitates surgical trainees' acquisition of robotic technical skills, especially in the underperformer group.
Deep learning prediction of error and skill in robotic prostatectomy suturing	Sirajudeen (2024)[61]	RARP	54 RARP videos from 8 surgeons; 2,507 errors annotated for gesture-related issues.	1) The DL models wishing to estimate surgeon skill scores reached about 82.08% of accuracy compared to OSATS (MAE=17.92%) and 79.4% with M-Gears (MAE=20.6%). 2) The best performing models (MS-TCN++ and MS-TCN) in error prediction achieved mean absolute	This study aimed to automate the assessment of surgical skill and error detection in robotic-assisted radical prostatectomy (RARP) suturing using deep learning models. The researchers compared AI skill scores with OSATS and M-Gears (validated global rating scale tools). Notably, researchers used Observational Clinical Human Reliability

				precision of 37.14%, area under the curve 65.10% and Macro-F1 58.97%.	Analysis (OCHRA) methodology, a previously validated tool in minimally invasive surgery.
Autonomous Educational System for Surgical Training Utilizing Deep Learning Combined with Extended Reality	Stone (2025)[62]	Laparoscopic partial nephrectomy	17 participants. Training was conducted on 396043 labeled images across four different clamp states.	1) The ESIST system accurately classified bulldog clamp placement with near-perfect performance (99.91% accuracy, Cohen's Kappa 0.994). 2) Trainees received real-time corrective feedback through an XR headset, and survey results showed 84% rated the system as more valuable than standard VR simulators, 73% rating the system as an ideal teaching tool and 100% reporting high educational value.	This study introduced an AI system (ESIST) to guide trainees to place a clamp on the renal artery using a kidney phantom while wearing an extended reality headset. They showed that the ESIST could assist in surgical education, potentially offer a means to monitor procedural progress and enable fast learning.
Development of an Automated Composite Ureteroscopic Efficiency Score Through Simulated	Valovska (2023)[63]	Urinary endoscopy (Flexible ureteroscopy (fURS))	12 urologists (residents and attendings).	1) The Composite Ureteroscopic Efficiency Score, which effectively distinguishes skill levels by combining task time, scope	The study developed a Composite Ureteroscopic Efficiency Score (CUES) using computer vision and machine learning to assess surgical skill

Ureteroscopic Skills Assessment				distance, smoothness, and collisions, shows strong discrimination between early and late attempts (ROC-AUC = 0.86). 2) Differentiated skill-level, residents scored significantly lower than attendings (0.02 vs 0.44, p = 0.04), with clear improvements across repeated attempts (time: 87.5s → 51.2s → 46.5s).	in simulated fURS tasks. Metrics included task time, scope travel distance, movement smoothness, and purposeful collisions. Results suggest CUES could enable automated, objective skill assessment for training.
Evaluating robotic-assisted surgery training videos with multi-task convolutional neural networks	Wang (2021)[64]	RARP (VUA)	18 training videos (74 annotated segments) from 2 faculty, 5 fellows, and 11 residents.	1) The AI model was able to predict GEARS skill scores from trainee videos with good accuracy, matching human raters within one point in 86% of cases . 2) The model reliably distinguished novices and intermediates, though it struggled with expert classification, showing that automated scoring is promising but not yet fully	The study developed neural networks for predicting the surgical proficiency score (GEARS score) from video clips. They found that Evaluation of GEARS sub-categories with AI is possible for novice and intermediate surgeons, but additional research is needed to understand if expert surgeons can be evaluated with a similar automated system.

				reliable (novices 100%, intermediates 80%, experts 0%).	
Evaluating robotic-assisted partial nephrectomy surgeons with fully convolutional segmentation and multi-task attention networks	Wang (2023)[65]	RAPN	150 videos consisting of 50 surgeries which were rated by a total of 5 annotators/ raters.	1) Network trained with self-attention and cross entropy loss (SA-CE) performed substantially better than other architectures. 2) For skill prediction, the best model showed a match accuracy of 0.60 for GEARS and 0.63 for OSATS.	This study uses machine learning to evaluate surgical skill from videos of RAPN. Researchers used cascade neural networks predicting surgical proficiency scores (OSATS and GEARS). Overall, the model performs well for subcategories such as force sensitivity and knowledge of instruments of GEARS and OSATS scoring, but has issues with false positives and negatives. The authors attribute this to limited sample size and training data.
Novel evaluation of surgical activity recognition models using task-based efficiency metrics	Zia (2019)[66]	RARP	2 live-broadcasted cases. The AI model was trained on 7768 annotated videos	1) Using conventional model evaluation, the RP-Net-V2 model segmented robotic prostatectomy with high accuracy, achieving a Jaccard index of 0.85 compared to	This study introduced a CNN-LSTM model to recognize the steps of RARP. The authors propose a novel evaluation method that goes beyond traditional frame-level

			for development.	0.81 for the unfiltered version ($p < 0.01$). 2) Using metrics-based evaluation, the model predicted task boundaries within 4 minutes of ground truth in over 80% of cases, and efficiency metrics computed from its predictions strongly correlated with expert-annotated values.	accuracy, assessing models based on how accurately they compute task-based efficiency metrics from surgical system data. The present model achieved high performance (Jaccard Index of 0.85) and its predictions showed strong correlation with metrics derived from expert annotations. It demonstrates the feasibility of automated, metrics-based evaluation for improving surgeon performance and training scalability.

Key endpoints and summary of each study is provided[8-66]