International Journal of Computer Vision
https://doi.org/10.1007/s11263-022-01640-6

Surgical Tool Datasets for Machine Learning Research: A Survey

Mark Rodrigues[1] · **Michael Mayo**[1] · **Panos Patros**[2]

Received: 1 March 2021 / Accepted: 18 June 2022

Abstract
This paper is a comprehensive survey of datasets for surgical tool detection and related surgical data science and machine learning techniques and algorithms. The survey offers a high level perspective of current research in this area, analyses the taxonomy of approaches adopted by researchers using surgical tool datasets, and addresses key areas of research, such as the datasets used, evaluation metrics applied and deep learning techniques utilised. Our presentation and taxonomy provides a framework that facilitates greater understanding of current work, and highlights the challenges and opportunities for further innovative and useful research.

Keywords Surgical tool datasets · Machine learning · Deep learning · Dataset and algorithms survey · Hospi-Tools Dataset

1 Introduction

There are fourteen surgical specialities recognised by the American College of Surgeons, ranging from orthopaedic surgery through to vascular surgery (ACS, 2021). Each speciality has its own procedures and its own sets of surgical tools, including instruments, implants and screws designed for specific parts of the body, and for specific procedures. Rapid advances in minimally invasive surgery have led to new classifications of robotic or laparoscopic surgery and open surgery (Bhatt et al., 2018), and also to new types of instruments being introduced at a constant rate (Fig. 1).

Consequently, there are many thousands of different types of surgical tool in circulation within a hospital. Stockert and Langerman (2014) reported that just one institution processed over 100,000 surgical trays and 2.6 million surgical tools annually. There were on average 38 surgical instruments per tray, with around 6 trays used for each surgery (Mhlaba et al., 2015). Handling this volume manually in real time and under difficult, mission-critical conditions is a challenging task requiring highly trained surgical technicians. Automating surgical tool detection and recognition through computer vision and machine learning has numerous practical applications therefore, and these applications can lead to improved efficiencies and/or reduced costs. Applications include robotic and computer-assisted surgery (Sarikaya et al., 2017; Zhao et al., 2019a), instrument position recognition in minimal invasive surgery (Zhao et al., 2017), pose recognition in surgical training (Leppanen et al., 2018; Jo et al., 2019), and instrument tracking in hospital inventory management (Ahmadi et al., 2018).

Ward et al. (2021b) discussed the application of computer vision and deep learning to surgery, specifically for the identification of surgical phases and instruments in multiple surgery procedures. van Amsterdam et al. (2021) reviewed methods for automatic recognition of fine-grained gestures in robotic surgery, and highlighted the promising results obtained by deep learning based models. Garrow et al. (2021) provided an overview of deep learning models utilized for automated surgical phase recognition using data inputs such as videos or surgical instrument use, and found that laparoscopic cholecystectomy was the most common operation evaluated. Yang et al. (2020) presented a review of the literature regarding image-based laparoscopic tool detection and tracking using convolutional neural networks (CNNs), including a discussion of available datasets and CNN-based detection and tracking methods. They also presented a quantitative estimation of several performance measures. Our

Communicated by Jan Kybic.

✉ Mark Rodrigues
mark.rodrigues@waikato.ac.nz

Michael Mayo
michael.mayo@waikato.ac.nz

[1] Department of Computer Science, University of Waikato, Hamilton, New Zealand

[2] Department of Software Engineering, University of Waikato, Hamilton, New Zealand

Published online: 12 July 2022

Fig. 1 Open surgery instruments

survey maintains a focus on surgical tools, reviews image based surgical tool detection, and provides an overview of instrument related surgical data science and machine learning techniques and algorithms. It is comprehensive in nature, covering the range of relevant research conducted in our specified time period—which was from 2015 till 2022. In particular, we maintain a focus on surgical tool datasets and on gaps in the research or on open research questions.

In this survey, we address three research questions:

1. What surgical tool datasets are used in machine learning research?
2. What machine learning methods are used in the research?
3. What are the gaps in surgical tool datasets and associated machine learning research?

Our objective, therefore, is to build a comprehensive knowledge hierarchy of applied research in surgical tool detection, classification and segmentation to guide future work. A concrete outcome is an integrated taxonomy of the methods used across the tasks undertaken in the research. We evaluate the pros and cons of each method or set of methods used in each paper, and address what is missing in the research to date. Gaps not just in the research but also in the publicly available datasets are discussed. We provide a comprehensive survey of the various datasets associated with surgical tool detection (Tables 1, 5, and 6). We address the specific challenges faced in this task and evaluate how they have been addressed. Finally, we make recommendations based on the results of the survey to encourage further work in this area.

2 Survey Methodology

As a logical starting point and following the approach used in similar survey work (Egger et al., 2020; Litjens et al., 2017), we rely on both PubMed and Google Scholar to conduct an initial search for literature. We chose PubMed because of its medical focus and Google Scholar because it indexes a range of peer reviewed international journals and conferences across disciplines. We expected that this strategy would provide a broader range of articles than reliance on academic databases. We used keywords to search the databases—an example search could include the keyword {"Surgical" OR "Surgery"} together with the keywords {"tool" OR "instrument"} AND {"detection" OR "classification"} AND {"deep learning OR machine learning"}. Comprehensive combinations of key words were used to ensure diligence in our search. Our reliance on Google Scholar proved to be a good strategy to develop an acceptable starting set of literature which avoided bias or preference towards any specific publisher. We also conducted other complimentary searches, such as reviewing reference lists, searching through conference proceedings, and obtaining leads from prominent researchers and authors in this area (Wohlin, 2014). Once we completed the literature search, we comprehensively summarised the literature set in a spreadsheet, with sample entries shown in Tables 2 and 3. We then read the papers to ascertain if they all actually included surgical tool detection in some form or the other. For example, some of the studies on surgical workflow also included a surgical tool detection component since it has been reported that combining instrument signals with visual features leads to better segmentation, and faster and more accurate detection (Dergachyova et al., 2016). We discarded papers that did not discuss surgical tools or which used external markers for tool detection or tracking. The resultant collection of 161 papers are surveyed in this review (Fig. 2).

3 Dataset Review

Medical image analysis challenges have resulted in many new and innovative approaches to surgical instrument recognition. These challenges are designed to provide a platform for the development of cutting edge machine learning solutions in medical imaging, and research in these challenges has addressed instrument segmentation, detection and localisation, tracking and pose estimation, velocity and instrument state. Al Hajj et al. (2019) highlight the fact that more than twenty annual challenges were hosted, and the CATARACTS, EndoVis and M2CAI challenges specifically addressed the issue of instrument detection. In the medical image challenges, generally a specific task is defined, a dataset is provided, evaluation procedures are defined, algo-

Table 1 Surgical tool datasets—examples of data and instruments

Challenge name	Data available	Instrument nos
ROBUST-MIS 2019 (Ross et al., 2019)	30 surgical procedures from three surgery types	Large biopsy forceps
EndoVis 2018 (Allan et al., 2020)	14 sequences of abdominal porcine procedures	Seven surgical instruments
CATARACTS (Al Hajj et al., 2019)	50 videos of phacoemulsification cataract surgeries	21 surgical tools
Cholec80 (Twinanda et al., 2017)	80 videos of cholecystectomy surgeries	Seven tools or instruments
EndoVis 2017 (Allan et al., 2019)	10 sequences of abdominal porcine procedures	Seven surgical
Lapgyn4 (Leibetseder et al., 2018)	Gynaecological laparoscopy dataset	Zero to three instruments
ATLAS Dione (Sarikaya et al., 2017)	86 full videos and 910 clips of six surgical tasks	Two Tools
RMIT Dataset (Sznitman et al., 2012)	8 in-vivo sequences	Single-instrument dataset

Table 2 Comprehensive literature summary—example entry (A)

Sr.	Authors	Year	Title	Journal/conference	Overview	Dataset
7	Al Hajj et al.	2018	Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks	Med Image Anal 47	Automatic monitoring of tool usage during a surgery: cataract and cholecystectomy	Cataracts, Cholec80 Datasets

Table 3 Comprehensive literature summary—example entry (B)

Sr.	Technique used	CNN used	Instruments	Data type	Results
7	CNN and RNN enriched by progressively adding weak classifiers trained to improve classification accuracy. CNN outputs fed to RNNs - jointly boosts an ensemble of CNNs and of RNNs	Seven CNNs used as weak classifiers	21 Cataract and 7 Cholec80	Videos via microscope (cataract) or endoscope (cholecystectomy)	ROC = 0.9961 in offline mode; ROC = 0.9957 in online mode

rithms are developed and applied, and solutions are tested on a held-out test set. A critical component is the dataset provided, and every attempt is made by the challenge organisers to ensure that this data is representative of the type of data generally encountered in clinical practice. We describe the important Challenge Datasets in the next section.

3.1 Challenge Datasets

ROBUST-MIS 2019, a part of the EndoVis Challenge series, was based on surgical procedures from three types of surgery. The videos were from 30 minimally invasive surgical procedures: 10 rectal resection procedures, 10 proctocolectomy procedures and 10 sigmoid resection procedures. A labelling

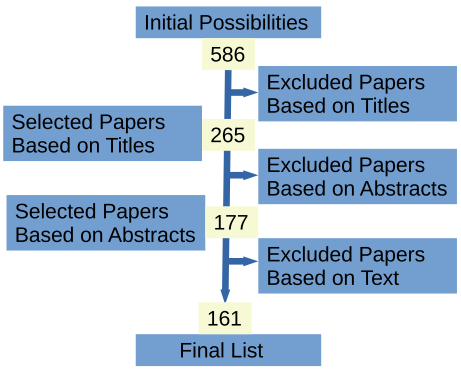

Fig. 2 Paper selection flow

mask and instrument labels were manually created for the 10,040 extracted endoscopic video frames (Ross et al., 2019). This dataset was based on the Heidelberg colorectal data set (Maier-Hein et al., 2021). The Endoscopic Vision 2018 Robotic Scene Segmentation Dataset provided images that were based on actual surgical procedures and included considerable variability in backgrounds, instrument movements, angles, and scales. The entire challenge dataset was made up of 19 sequences of porcine endoscope images and the objective was to perform semantic segmentation of surgical images into a set of medical device classes and a set of anatomical classes (Allan et al., 2020). The EndoVis 2017 Robotic Instrument Dataset was made up of 10 sequences of abdominal porcine procedures, which presented seven different robotic surgical instruments (Table 4). The relatively small size of the dataset was an issue, since it was only made up of 3000 frames in total, out of which 1800 frames were selected as training data. The dataset supported three different segmentation tasks: binary segmentation, parts of instruments (e.g., shaft, wrist, claspers and ultrasound probes) and type segmentation (e.g., needle driver, forceps, scissors, sealer and others). The EndoVis 2015 instrument segmentation and tracking dataset provided data for rigid and articulated robotic instruments in laparoscopic surgery. For rigid instruments, 2D in-vivo images from four laparoscopic colorectal surgeries were provided for segmentation and in-vivo video sequences of four laparoscopic colorectal surgeries were provided for tracking. For articulated instruments, four 45-second 2D images sequences of at least one large Needle Driver instrument in an ex-vivo setup were provided. Relevant annotations and additional test data were also provided.

The Challenge on Automatic Tool Annotation for Cataract Surgery (CATARACTS) Dataset consisted of 50 videos of phacoemulsification cataract surgeries. Cataract surgery is the most common of the surgical procedures, and ophthalmologists use a wider range of tools than surgeons doing robotic or laparoscopic surgeries; consequently this dataset provided a large set of tools. There are more than nine hours of videos with an average duration of almost eleven minutes per surgery. A total of twenty one surgical tools are present in the videos (Table 4); a tool was only considered to be in use when in contact with the eyeball. In any particular frame, up to three tools can be visible at a time. However, this occurs in only 4% of the frames; 45% of the frames show no tools at all, 38% show one tool and 17% show two tools (Al Hajj et al., 2019) (Fig. 3).

The Cholec80 dataset contains 80 videos of cholecystectomy surgeries, and seven tools or instruments are present in the dataset (Table 4). Some tools—such as the grasper and hook—feature in many frames while other tools—such as the scissors and irrigators—are less used and appear with much lower frequency in the videos / frames (Twinanda et al., 2017). The m2cai16-tool dataset is a subset of the Cholec80 Dataset and it consists of fifteen cholecystectomy videos with binary annotations of the seven tools present (Fig. 4).

Table 4 Tools in cataract dataset

Dataset	Instrument
CATARACTS	Biomarker, Charleux cannula, hydrodissection cannula, Rycroft cannula, viscoelastic cannula, cotton, capsulorhexis cystotome, Bonn forceps, capsulorhexis forceps, Troutman forceps, needle holder, irrigation/ aspiration HP, phacoemulsifier HP, Cvitrectomy HP, implant injector, primary incision knife, secondary incision knife, micromanipulator, suture needle, Mendez ring, Vannas scissors, grasper, bipolar, hook, scissors, clipper, irrigator, specimen bag
Cholec80 Dataset	Grasper, hook, bipolar, scissors, clipper, specimen bag and irrigator
EndoVis 2017	Large Needle Driver, Prograsp Forceps, Monopolar Curved Scissors, Cadiere Forceps, Bipolar Forceps, Vessel Sealer and a drop-in ultrasound probe, typically in the jaws of the Prograsp

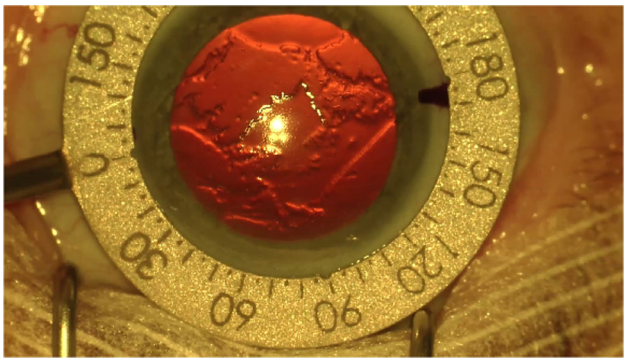

Fig. 3 Cataracts Dataset

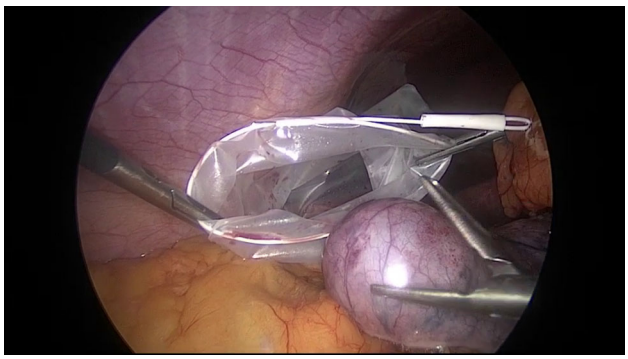

Fig. 4 Cholec80 Dataset

Details of the surgical tool datasets used in the challenges is presented in Table 5. In addition to the metadata provided about each dataset, additional metadata charac-

Table 5 Taxonomy of surgical tool datasets

Metadata characteristic	ROBUST-MIS (EndoVis) 2019	EndoVis 2018 and 2017	MICCAI-2016	CATARACTS	Cholec80
Size or instances	30 videos	2018—14 video sequences; 2017—10 video sequences	15 videos	50 videos	80 videos
Database focus	Rectal resection, proctocolectomy, sigmoid resection surgeries	Abdominal porcine procedures	Cholecystectomy surgeries	Cataract surgeries	Cholecystectomy surgeries
Default task	Segmentation and detection	Segmentation	Detection	Presence detection	Detection
Range—number of objects/classes	2	7	7	21	7
Image acquisition platform/device	–	da Vinci Xi Robotic systems	da Vinci Xi Robotic systems	Toshiba 180I camera and MediCap USB200 recorder	–
Image acquisition location	–	–	University Hospital of Strasbourg	Brest University Hospital	University Hospital of Strasbourg
Image illumination	–	–	Fibre-optic in-cavity	Microscope Illumination	Fibre-optic in-cavity
Distance to object	Close—in-cavity	Close—in-cavity	Close—in-cavity	V. close—surgical microscope	Close—in-cavity
Metrics recommended	DICE	IOU / AUC	AP	AUC	AP
Annotations	Masks	Masks	Binary	Binary	Bounding boxes
Dataset organisation	10,040 frames	2018—15 training and 4 test videos; 2017—1800 training and 1200 test data	23,287 training and 12,541 testing samples	500,000 training and 500,000 test frames	86,304 training and 98,194 test frames
Image resolution	1280 × 1024 pixels	–	–	1920 × 1080 pixels	–

teristics that are common for all the datasets listed in this table are: Image Type—Videos; Image Modality—RGB; Data Types—Images; Attribute Types—Categorical; Dataset Structure—Flat; Collection Methods—Controlled; Annotation Levels—Expert; Data Variety—Specific and Dataset Licence—Register/Public. It is significant that many of the cells in the table are empty, and this highlights the lack of metadata, details and information about the collection and curation of these datasets.

3.2 Other Surgical Tool Datasets

In addition to the challenge datasets described above, many other surgical tool datasets have been developed and present these datasets in Table 6. Again, the blank cells in these tables serves to highlight the shortfall in metadata and details about these datasets. The ATLAS Dione dataset provided video data of ten subjects performing six different surgical tasks. The dataset was described as being challenging as it had camera movement and zoom, free movement of surgeons, a wide range of expertise levels, background objects with high deformation, and annotations that included tools with occlusions, change in pose and articulation or with partially visibility (Sarikaya et al., 2017). The Retinal Microsurgery (RMIT) dataset consisted of 18 in-vivo sequences of retinal procedures; for each sequence, four joints (Tip1, Tip2, Shaft and End Joint) of the retinal instrument were annotated. The RMIT was a single-instrument dataset—specified only as a Retinal Instrument. The dataset was further classified into four instrument-dependent subsets. There were three annotated tool joints and two semantic classes (tool and background).

Lapgyn4 Dataset is a four-part gynaecological laparoscopy dataset comprising collections of images depicting general

Table 6 Surgical tool datasets

Dataset	Focus	Data type	Data quality	Instr.	Licence	Organisation	Annotations
Atlas Dione (Sarikaya et al., 2017)	Urology—urethrovesical anastomosis	86 videos and 910 clips	854 × 480 pixels	3	Open	22,486 frames	Bounding boxes
Bar's Cholecystectomy (Bar et al., 2020)	Cholecystectomy laproscopy	1243 videos	–	–	Private	745, 187 and 311 training, validation and test videos	Phase Annotations
Bouget's NeuroSurgicalTools (Bouget et al., 2015)	Neurological surgery	14 videos	720 × 576 pixels at 25 FPS	7	Private	–	Bounding polygon
CaDIS—Retinal (Grammatikopoulou et al., 2019)	Ophthalmic surgery	4671 frames	960 × 540 pixels	29	Open	–	Annotated tools
Cataracts-101 – Retinal (Schoeffmann et al., 2018)	Ophthalmic surgery	843 frames	1920 × 1080 pixels	11	Open	–	Annotated tools
Choi's Mastoidectomies (Hong et al., 2020)	Otolaryngology	70 videos	1920 × 1080 pixels at 30 FPS	6	Private	–	Masks
Hong's CholecSeg8k (Hong et al., 2020)	Cholecystectomy	8080 frames	854 × 480 pixels	2	Open	–	Semantic segmentation masks
Flapnet (Attanasio et al., 2020)	Lobectomy	2160 images	506 × 466 pixels	1	Open	–	Instrument Presence
Garcia Perez's RoboTool (Garcia-Peraza-Herrera et al., 2021)	Laparoscopic surgeries	20 surgical procedures	–	–	Public	6130 images	514 manually annotated images
Gruijthuijsen's Gynaecology (Gruijthuijsen et al., 2021)	Gynaecology	1180 images	1920 × 1080 pixels	–	Private	1110 training and 70 testing images	Bounding boxes in 379 images
Hasan's ART-Net (Hasan et al., 2021)	Gynaecology	29 videos	–	–	Private	1016 training and 3254 testing images	Segmentation masks
Heidelberg's Colorectal/HeiCo/Hei-Chole (Maier-Hein et al., 2021)	Colon-rectal surgeries	30 videos	960 × 540 pixels	3	Open	10,040 frames	Segmentation masks
Hossain's Knee Arthroplasty (Hossain et al., 2018)	Orthopaedic	16 videos	25 FPS	30	Private	–	Bounding boxes
Hou's SID19 (Hou et al., 2022)	Appendectomy, cholecystectomy and cesarean section	3800 images	3456 × 3456 pixels	19	Open	–	Image labels

Table 6 continued

Dataset	Focus	Data type	Data quality	Instr.	Licence	Organisation	Annotations
Huaulme's MISAW (MICCAI – 2020) (Huaulme et al., 2021)	Anastomosis	27 sequences	920 × 540 pixels	1	Open	17 training and 10 test sequences	–
Jha's Kvasir-Instrument (Jha et al., 2021a)	General surgery	3500 images	128 × 128 pixels	10	Open	–	Bounding boxes, segmentation masks
JIGSAWS (Gao et al., 2014)	General surgery	103 videos	–	–	Open	–	–
Kalavakonda's NeuroID (Kalavakonda et al., 2019)	Neurological surgery	5 videos	720 × 480 pixels at 30 FPS	8	Private	–	Bounding polygon
Kuglers' Cochlear (Kugler et al., 2020a)	Otorhino-laryngology	–	–	2	Private	–	Manually labelled screws
Kurmann's Retinal (Kurmann et al., 2017)	Ophthalmic surgery	4 videos	640 × 480 pixels at 30 FPS	–	Private	1500 frames	Fully annotated
LapGyn4 (Leibetseder et al., 2018)	Gynaecology	55,000 frames	–	4	Open	22,000 labelled frames	Manually annotated
Law's Vesico-Uritheral Anastomosis (Law et al., 2017)	Urology	12 videos	720 pixels	–	Private	146,309 frames	–
Lee's Phantom (Lee et al., 2019a)	Anatomical phantom plus animal studies	1600 frames	1280 × 1024 pixels	–	Private	1000 training and 600 testing frames	Segmentation masks
Lepannen's Micro-Neurological (Leppanen et al., 2018)	Neurological surgery	97,932 frames	720 × 486 pixels at 30 FPS	4	Private	96% training, 2% test, 2% val	–
Lu's SuPer and Hamlyn Heart (Lu et al., 2020)	Cardiothoracic surgery	2 videos	368 × 288 pixels	–	Private	–	–
Matton's BigCat (N. et al., 2022)	Cataract surgery	190 videos	1920 × 1080 pixels	10	Private	114 training, 38 validation and 38 testing videos	Instrument presence ground truth
Meeuwsen's Laparoscopic Hysterectomy (Meeuwsen et al., 2019)	Gynaecology	40 videos	–	12	Private	–	–
Meir-Hein's InstrumentCrowd (Maier-Hein et al., 2014)	Adrenalectomies, pancreatic resections	120 images	–	2	Private	–	2350 instrument segmentations
Murillo's Open Surgery Set (Murillo et al., 2017)	General surgery	7000 Images	480 × 480 pixels	5	Private	2000 training and 5000 testing images	–

Table 6 continued

Dataset	Focus	Data type	Data quality	Instr.	Licence	Organisation	Annotations
Nakawala's Nephrec9 (Nakawala et al., 2019)	Urology	9 videos	720 × 578 pixels at 25 FPS		Private	741,573 frames	Multiple instrument annotated
Qin's Sinus-Surgery (Qin et al., 2020)	Otorhino-laryngology	10 videos	320 × 240 pixels at 30 FPS	1	Open	–	Manually annotated instrument
Ramesh's Neurosurgery (Ramesh et al., 2021a)	Neurosurgery	32 videos	640 × 480 pixels at 1 FPS	4	Open	22 training and 10 testing videos	Bounding boxes
RMIT—Retinal (Sznitman et al., 2012)	Ophthalmic surgery	18 videos	1920 × 1080 pixels	one	Open		Annotated tool joints
SimSurgSkill (MICCAI-2021)	Surgery virtual reality	Simulation videos	1280 × 720 pixels	2	Open		Bounding boxes
UCL DVRK Dataset (Colleoni et al., 2020)		20 videos	538 × 701 pixels	1	Open	8 training, 2 validation and 4 test videos	Tool segmentation masks
Wagner's Hei-Chole (EndoVis 2019) (Wagner et al., 2021)	Cholecystectomy laproscopy	33 videos	960 × 540 pixels; 1920 × 1080 pixels; 720 × 576 pixels;	21	Open	24 training and 9 test videos	6980 instrument occurences
Yamazaki's Laparoscopic Gastrectomy (Yamazaki et al., 2020)	General surgery	62 videos	1920 × 1080 pixels at 60 FPS	14	Private	8572 training and 2144 validation images	Bounding box
Yang's Cardiac (Yang et al., 2019)	Cardiac (Porcine hearts)	93 images	120 × 69 × 92 to 294 × 283 × 202 voxels	2	Private	62 training and 20 test volumes	Segmentation masks
Zadeh's Gynaecological (Zadeh et al., 2020)	Gynaecology	461 images	1920 × 1080 pixels	–	Private	–	–
Zhao's Datasets (Zhao et al., 2019c)	General, retinal, cardiac	6 videos	320 × 240 pixels	10	Private	36,000 frames; 75% training and 25% test	Manually annotated

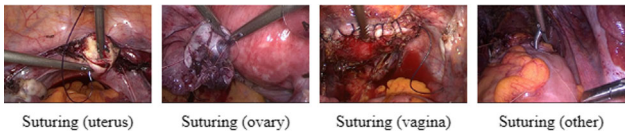

Fig. 5 Lapgyn Dataset

surgical actions, anatomical structures, conducted actions on specific anatomy as well as examples of differing amounts of visible instruments. It is actually four datasets (Surgical Actions, Anatomical Structures, Actions on Anatomy, Instrument Count) of over 500 surgical interventions. The Instrument Count dataset consists of images from gynaecology and cholecystectomy (including samples from Cholec80 dataset) with zero to three instruments (Leibetseder et al., 2018) (Fig. 5).

There were other datasets that we evaluated but did not include since they did not provide sufficient focus or coverage of surgical tools. These included DAISI: Database for AI Surgical Instruction (Rojas et al., 2020), the MISAW dataset used for the MIcro-Surgical Anastomose Workflow recognition on training sessions challenge (Huaulme et al., 2021), the Bypass40 dataset of laparoscopic gastric bypass procedures (Ramesh et al., 2021b), and the EAD2020 dataset (Ali et al., 2021) (Fig. 6).

4 Algorithm Review

Liu et al. (2020a) highlighted the inconsistency in the terminology used in the research, and stated that terms are often differently defined and applied. Some of the terms which were used include detection, presence, localization, recognition, classification, identification, labelling and annotation. The taxonomy of terms used in the literature reviewed is presented in Fig. 7, it is clear that definitions and terminology varies considerably and there is no uniformity in the application or understanding of these terms. When computer vision tasks are considered, multiple problems have been addressed in the literature. Guo et al. (2016) discussed image classification, object detection, image retrieval, semantic segmentation, and human pose estimation as the key computer vision tasks. Chai et al. (2021) similarly listed the main applications as object detection or recognition, visual tracking, semantic segmentation, and image restoration, with image classification providing the basic backbone of each application. Voulodimos et al. (2018) evaluated object detection, face recognition, action and activity recognition, and human pose estimation in their survey of key tasks in computer vision. Al Hajj et al. (2019) state that these tasks can be categorized according the precision of the desired outputs, with the finest or more precise level of surgical tool-based tasks at the tool segmentation level. The next level of precision in tasks is tool localisation, and this often leads to either tool tracking or pose estimation. The coarsest task is tool presence detection or determining which tools are present in each frame of a surgical video. While we considered all these approaches, in actual practice a pipeline using all these types of algorithms would follow a logical flow of tool presence detection, tool localisation, tool tracking, tool segmentation and tool pose estimation. We therefore used this logical flow approach to structure our analysis of the research.

5 Tool Presence Detection Research

In work using the CATARACTS dataset, Roychowdhury et al. (2017) fine-tuned Inception-v4, ResNet-50 and two NASNet-A instances. In their solution, they relied on Markov Random Field (MRF) for modelling long sequences of approximately 20,000 frames. Sahu et al. (2017a) trained ResNet-50 initialised with ImageNet weights on this dataset. Prellberg and Kramer (2018) used the CATARACTS dataset to explore different ways to use ResNet-50, and reported that fine-tuning ResNet achieved consistently better results than using ResNet as a fixed feature extractor in combination with a custom classifier. Al Hajj et al. (2019) reported on the results of surgical instrument presence detection with the CATARACTS Dataset. This included work using VGG-16 (Simonyan & Zisserman, 2014), Inception-v3 (Szegedy et al., 2016b), SqueezeNet (Iandola et al., 2016), DenseNet-161 (Huang et al., 2017), ResNet-34, ResNet-50, DenseNet-169, Inception-v4, ResNet-152, ResNet-101, DenseNet-169, NASNet-A (Zoph et al., 2018) and Inception-ResNet-v2 (Szegedy et al., 2016a). Twinanda et al. (2017) developed and used the Cholec80 dataset to test EndoNet, an architecture based on AlexNet, for tool detection. Sahu et al. (2017b) fine-tuned AlexNet on the m2cai16-tool dataset; using an approach similar to EndoNet. The Cholec80 dataset was used by Alshirbaji et al. (2018) to fine tune AlexNet for surgical tool classification. Mondal et al. (2019) used Cholec80 to train a multi-task learning framework based on ResNet50 trained on the ImageNet Dataset. The features extracted from the fully connected layer of ResNet50 were used to train a multitask Bi-LSTM. The final classification result was generated through combining the score results produced by both the LSTM hidden layers. Alshirbaji et al. (2021a) tested VGG-16, ResNet-50, DenseNet-121 and EfficientNet-B0 for surgical tool presence classification. This was tested on the Cholec80 and Cholec20 datasets. Alshirbaji et al. (2020a) generated synthetic data and used it to augment the Cholec80 dataset. AlexNet was fine-tuned using cross-dataset validation to improve tool presence detection. Vardazaryan et al. (2018) used ResNet18 pre-trained on ImageNet data and further trained the network on a Cholec80 sub-set of five videos annotated with image-level instru-

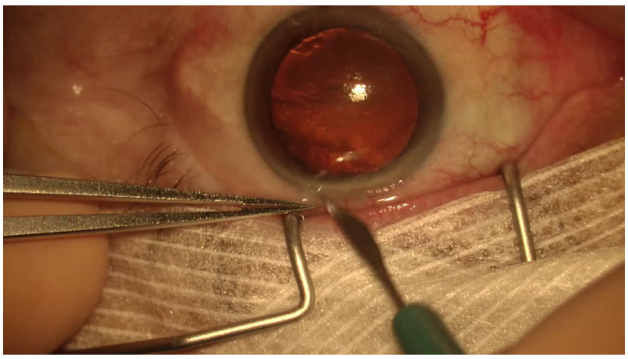

Fig. 6 CaDIS Dataset

ment bounding boxes for binary tool presence classification. Nwoye et al. (2019) adopted a similar approach but modified it with long short-term memories for better performance. Bodenstedt et al. (2018) used surgical tool presence in endoscopic video as a cue for surgery duration predictions, used ResNet152 for tool presence detection and evaluated their architectures on the Cholec80 dataset. Jin et al. (2020) presented a multi-task recurrent convolutional network with correlation loss (MTRCNet-CL) to exploit the relatedness of surgical tool presence and surgical phase to simultaneously boost the performance of the tasks of tool detection and phase recognition. The model was tested on the Cholec80 dataset. Al Hajj et al. (2019) used both the CATARACTS and the Cholec80 datasets for monitoring tool usage during a surgery. Their system jointly boosted an ensemble of CNNs and an ensemble of RNNs. Seven CNN architectures were used as weak classifiers—VGG-16, VGG-19, ResNet-101, ResNet-152, Inception-v4, Inception-ResNet-v2, NASNet-A. For RNN boosting, LSTM and GRU was used. Alshirbaji et al. (2020b) developed three balanced datasets by applying image transformations and substituting image backgrounds on instrument images extracted from the Cholec80 dataset. Wang et al. (2019) developed a deep neural network model, based on DenseNet121 pre-trained from ImageNet, utilizing both spatial and temporal information from surgical videos for surgical tool presence detection. They evaluated their model on two datasets: m2cai-tool and Cholec80.

Using the m2cai16-tool dataset, Raju et al. (2016) fine-tuned GoogleNet and VGG16 and used ten trained models (with 5-fold cross validation for both VGGNet and GoogleNet) in an ensembling process to obtain their final results. Zia et al. (2016) fine-tuned AlexNet, VGG-16 and Inception-v3 and presented a comparison of these different deep network architectures for surgical tool detection. Namazi et al. (2019) developed LapTool-Net, which was a contextual detector for surgical tools based on recurrent convolutional neural networks. The method exploited correlations among usage of tools in the m2cai16-tool dataset, as well as the context of the tools' usage for different tasks. Choi et al. (2017) proposed a real-time detection model for surgical instruments during laparoscopic surgery by using a CNN based on YOLO pre-trained on ImageNet. This was trained on the m2cai16-tool dataset. Hu et al. (2017) developed an attention-guided network (AGNet) and successfully tested it on the m2cai16-tool dataset. The method first extracted regions in images with high probability of containing surgical tools by a deep neural network (the global prediction network) and then analysed these regions via another deep neural network (the local prediction network) which provided a prediction for each tool. Lin et al. (2019) addressed surgical tool presence detection with the m2cai16-tool dataset as a multi-label classification problem. The authors relied on a pre-trained DenseNet201 with a classification layer whose output corresponds to the confidences of the presence of the seven tools in the image. Mishra et al. (2017) proposed a framework to detect tool presence in laparoscopy videos which consisted of a CNN based on ResNet50 for extracting visual features, and a Long Short-Term Memory network to encode temporal information. This was tested on the m2cai16-tool dataset.

Leibetseder et al. (2018) used GoogLeNet (Szegedy et al., 2015) to classify images in the LapGyn4 dataset. Kletz et al. (2019a) used the Lapgyn4 Dataset for the task of binary classification to recognise video frames as either instrument or non-instrument image, and trained GoogLeNet for instrument classification. Murillo et al. (2018) developed a tree-structured convolutional neural network for the classification of 10 open surgery instruments. Eight separate CNNs were trained on ten surgical instruments, and four CNNs on five instruments. Murillo et al. (2017) used 5 open surgery tools for testing the performance of CNNs and Haar Classifiers (Viola & Jones, 2001) for surgical instrumentation classification. A tree based tool classifier was designed using four CNNs for presence detection of the five surgical instruments.

Kurmann et al. (2017) presented a U-Net based surgical instrument detector which estimated instrument joint positions and instrument presence using a cross-entropy loss function. This was evaluated on a retinal and EndoVis 2015 datasets. Qiu et al. (2019) used the m2cai16-tool dataset and built a new dataset called the STT dataset with sequential frame annotations using bounding boxes. The authors then developed RT-MDNet, a real-time multi-domain convolutional neural network with three convolutional layers, a Region of Interest Alignment (RoIAlign) layer and three fully connected layers, and tested it on the STT Dataset. Hou et al. (2022) introduced an attention-based deep neural network—SKA-ResNet—composed of a feature extractor with a selective kernel attention module and a multi-scale regularizer to exploit the relationships between feature maps. Their SKA-ResNet was tested on a new surgical instrument dataset called SID19 for the classification of surgical tools.

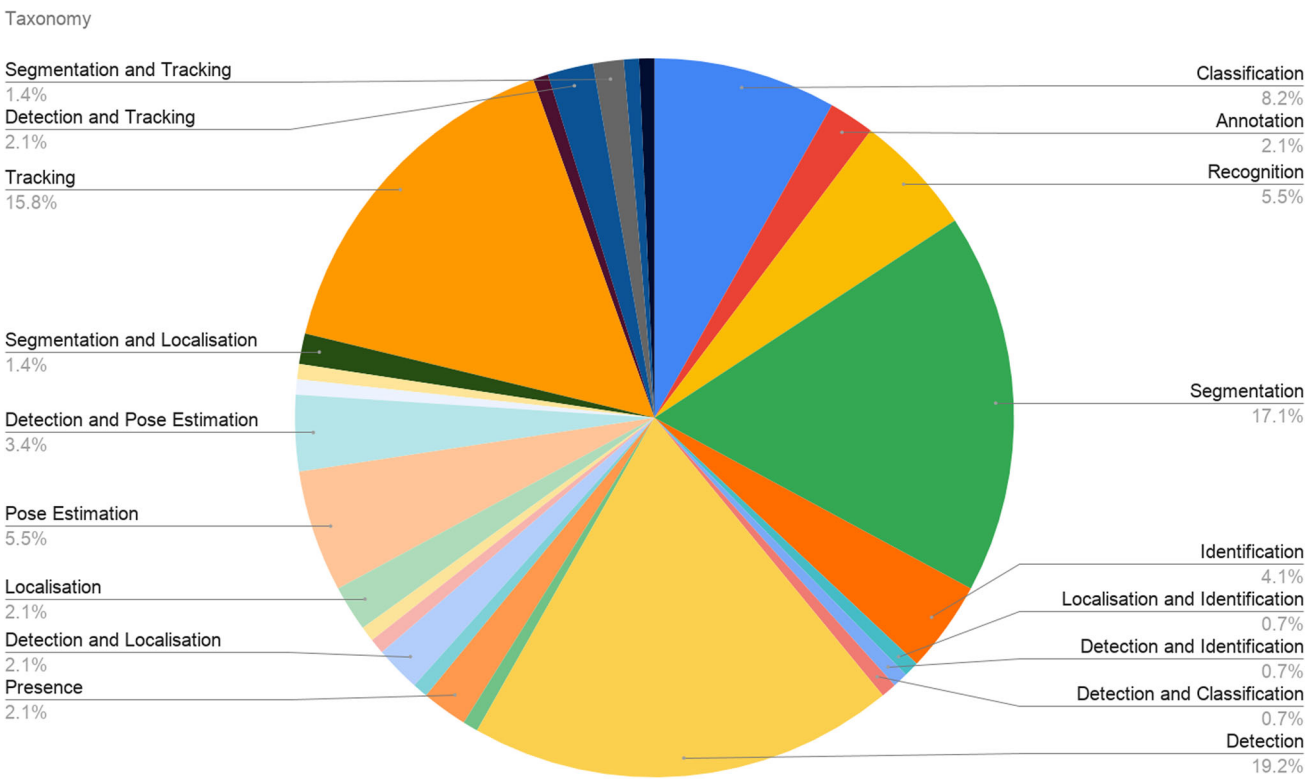

Fig. 7 Taxonomy of approaches

6 Tool Localisation Research

Banerjee et al. (2019) used the CATARACTS dataset for a multi-label multi-class classification task, and developed a framework for localization and detection of tools. A tool counter was implemented using ResNet-18. Using activation maps, three smaller regions of interest were used to train a new CNN which predicted the tool type among the given 22 classes. Three baseline models were trained for the task—AlexNet, VGGNet and ResNet-18/50/152.

Xue et al. (2022) proposed a pseudo supervised surgical tool detection (PSTD) framework, which used pseudo bounding box generation, box regressor, weighted mean boxes fusion and a classifier with bi-directional channel adaption for surgical tool detection. This weakly supervised surgical tool detection (WSTD) approach was successfully tested on the Cholec80 dataset using image-level tool category labels. Alshirbaji et al. (2021b) evaluated the generalisation ability of a VGG-16 model on images from different datasets for surgical tool detection. The datasets used were Cholec80 and a Gyna05 dataset which consisted of 5 videos of gynaecologic procedures, and target tools were the four surgical tools which were present in both datasets.

Nwoye et al. (2021a) developed the CholecTriplet2021: the endoscopic vision challenge for the recognition of surgical action triplets in laparoscopic videos. The focus was on fine-grained surgical activity recognition, modelled as a triplet—instrument, verb, target. This was defined in terms of surgical activities as triplets of the actual instrument that was used, the actions performed, and the target anatomy for each surgery, and was provided as part of the EndoVis2021 grand-challenge. Nwoye et al. (2021b) developed a model which recognized triplets from these surgical videos by leveraging attention at two different levels—a Class Activation Guided Attention Mechanism (CAGAM) and a Multi-Head of Mixed Attention(MHMA). This method used cross and self attentions to capture relationships between the triplets. Nwoye et al. (2020) used class activation modules which used the instrument activation maps to guide the verb and target recognition. They used a dataset based on Cholec80 annoted with 135K action triplets—termed the CholecT40 dataset—and developed a multitask learning (MTL) network with three branches for the instrument, verb and target recognition.

Liu et al. (2020b) used anchor-free convolutional neural network, based on a compact stacked hourglass network, for surgical tool detection. This was tested on the ATLAS Dione and Endovis Challenge datasets, and compared to results using Faster RCNN, Yolov3 (Darknet-53) and CenterNet (Hourglass-104). In surgical tool detection work associated with the ATLAS Dione dataset, Sarikaya et al. (2017) developed a framework with a Region Proposal Network (RPN) and a multimodal two stream convolutional network

for object detection and localization, based on image and temporal motion cues. Fast R-CNN (Girshick, 2015) was used for the object detection task, and the region proposal boxes of RPN with the convolutional features were used as input for the detection network streams on both modalities. Using the EndoVis Challenge dataset and the ATLAS Dione dataset, Zhao et al. (2019a) adopted a frame-by-frame detection method using a cascading convolutional neural network (CNN) which consisted of two different CNNs for real-time multi-tool detection. The method was tested—along with Faster R-CNN (Ren et al., 2017), Yolov3 (Redmon et al., 2016), and RetinaNet (Lin et al., 2017)—on the two datasets. Liu et al. (2020c) proposed an anchor-free convolutional neural network (CNN) architecture using a compact stacked hourglass (Newell et al., 2016) network for surgical tool detection, and tested it on the ATLAS Dione and EndoVis 2015 datasets. The authors also tested five backbones—ResNet-18, ResNet101, Deep Layer Aggregation or DLA-34 (Yu et al., 2018), Hourglass-104 (Law & Deng, 2020), and lightweight Hourglass—and achieved good accuracy and speed for real-time surgical tool detection.

Ciaparrone et al. (2020) tested 12 different combinations of CNN backbones and training hyper-parameters for surgical tool detection on a dataset derived from 13 high-quality endoscopic/laparoscopic videos. Mask R-CNN was used with ResNet-50, ResNet-101 and ResNet-152 as backbone networks. Their best results were obtained using a ResNet101 and training the network for 25 epochs. Shimizu et al. (2021) employed three modules for localization, selection, and classification for detection and classification task of surgical tools from egocentric images for open surgery analysis. Two tools—scissors and needle holders—were detected using Faster R-CNN and were classified using a convolutional neural network and long short-term memory (LSTM) module.

Ramesh et al. (2021a) developed a Yolov5-based system to detect micro-surgical tools from neurosurgical videos. Tool characterization was also reported based on tool on-off time, tool usage time and tool trajectory. Garcia-Peraza-Herrera et al. (2017) introduced two novel lightweight architectures, ToolNetMS and ToolNetH, defined in terms of multi-scale and holistically-nested CNN architectures, for the real-time segmentation of robotic surgical tools. These architectures were evaluated on the EndoVis 2015 dataset. Pakhomov et al. (2019) converted a residual image classification Convolutional Neural Network (ResNet-101) into a Fully Convolutional Network (FCN), performed simple bilinear interpolation of the feature maps for semantic image segmentation, and tested it for binary-segmentation performance on the EndoVis 2015 dataset.

Bouget et al. (2015) used the NeuroSurgicalTools dataset and developed a two step approach for surgical tool detection, where the first stage of the approach performed pixel-wise semantic labelling while the second stage matched global shapes. Leppanen et al. (2018) pioneered work for surgical instrument detection under high microscope magnification using CNNs in micro-neurosurgical videos. Two CNNs were trained—one for instrument detection and instrument tip location detection by classifying small parts of the frame at a time, and the second to detect whether the instrument is present in the frame using the full frame image. Law et al. (2017) trained a stacked hourglass network to detect the key-points of the robotic instruments in vesico-urethral anastomosis surgery videos using crowd-sourced annotations. They also trained a support vector machine (SVM) to classify the skill of a surgeon using the tracking results.

Nakawala et al. (2019) used their Nephrec9 dataset to test a "Deep-Onto" network for surgical workflow and context recognition, including instruments. The network was an ensemble of deep learning models (Inception-V3 pre-trained on ImageNet) with knowledge management tools, ontology and production rules, including usage of instruments. This combined use of deep learning, knowledge representation and reasoning techniques was found to be effective for automatic surgical workflow analysis on robot-assisted urological surgery.

Hossain et al. (2018) relied on CNNs for real-time surgical tools recognition in Total Knee Arthroplasty (TKA), and exploited region based convolutional neural networks to perform real time tool detection. The method was based on Faster R-CNN with VGG-16 as base network, and RGB image convolutional features were used to train a Region Proposal Network (RPN) that generated object proposals, the output was the coordinates of bounding boxes around the deployed surgical tools. Yamazaki et al. (2020) created a dataset from 52 laparoscopic gastrectomy videos, and used this to test Yolov3 for surgical instrument detection. Bar et al. (2020) used an approach based on inflating ResNet-50 into a 3D ConvNet model (I3D) for surgical phase classification. This was termed the short-term model, and the long-term model was a Long Short-Term Memory (LSTM) network. The approach used surgical tool presence as cues for each phase, and was tested on their laparoscopic cholecystectomy dataset.

Yang et al. (2019) relied on a Pyramid-UNet to localize a cardiac intervention instrument (RF-ablation catheter or guidewire) in a 3D ultrasound image for cardiac electrophysiology (EP) and transcatheter aortic valve implantation (TAVI) procedures. This was tested on their dataset of cardiac ultrasound images from porcine hearts. Colleoni et al. (2019) proposed a 3D FCNN architecture for surgical-instrument joint and joint-connection detection, using spatio-temporal features for robotic tool detection and articulation estimation. This was trained and tested on the EndoVis 2015 and the UCL dVRK datasets. Jin et al. (2018) extended the m2cai16-tool dataset by providing labels for 2532 of the frames

with the coordinates of spatial bounding boxes around the tools, and made a new m2cai16-tool-locations dataset available. Their approach for instrument localization was based on Faster R-CNN. In work that utilised the m2cai16-tool-locations and m2cai16-tool-datasets, Jo et al. (2019) applied two algorithms—YOLO9000 (Redmon & Farhadi, 2017) and missing tool detection—to perform detection of surgical instruments in real time.

7 Tool Tracking Research

Tang et al. (2022) leveraged multimodal imaging and deep-learning to dynamically detect surgical instrument positions in ophthalmic surgical maneuvers. In their system, they combined spectrally encoded reflectometry (SER) and cross-sectional OCT imaging for automated instrument-tracking, and tested it on 4730 manually-labelled SER images of a 25-gauge internal limiting membrane (25 G ILM) forceps.

Al Hajj et al. (2019) defined tool tracking work in terms of monitoring tool location over time. Gruijthuijsen et al. (2021) trained a U-Net CNN to segment instruments, training it on their gynaecology dataset. They converted the segmentation prediction into a graph and used this for tool tip prediction in their autonomous instrument tracking framework. Meeuwsen et al. (2019) developed a dataset of 40 laparoscopic hysterectomy (LH) surgeries and built a Random Forest surgical phase recognition model. Lee et al. (2019a) collected three phantom frame-sequence datasets using tracked surgical tools over an anatomical phantom. These datasets were used to test U-Net, TernausNet-11 with a pre-trained VGG-11 network, LinkNet-34 and LinkNet-152 for the semantic labelling, binary segmentation and real-time tracking of surgical tools without any human intervention.

Using a subset of the m2cai-2016 dataset, Zhang and Gao (2020) developed a surgical instrument tracking framework based on object extraction via deep learning, where a segmentation model extracted the end-effector and shaft of the surgical instrument in real time. The model was based on LinkNet with ResNet-18, pre-trained on ImageNet.

Chen et al. (2017b) proposed a visual tracking method for surgical tool tracking based on a CNN with line segment detector (LSD) for the detection part and a spatio-temporal context (STC) learning algorithm for the tracking part. They successfully tested this system on three laparoscopic surgical datasets—a simulation dataset, a real in-vivo dataset and a standard dataset. Zhao et al. (2017) considered a surgical instrument as consisting of two parts: an end-effector and a shaft. Edge-points and line features were used for the shaft detection and a CNN based on AlexNet (Krizhevsky et al., 2012) was used to track and detect the end-effector.

Hiasa et al. (2016) proposed and evaluated a method for segmentation of surgical instruments from RGB-D Endoscopic Images using CNNs. The method used RGB and depth images from stereo endoscope images, and the output was a likelihood image, where white pixels indicated a high probability of instruments and black pixels indicated high probability of background. Segmentation was seen as a critical task for 3D surgical tool tracking and reconstruction.

Zhao et al. (2019c) used two CNNs and six datasets to develop a coarse to fine method for surgical tool tracking. The first CNN, based on AlexNet, classified 10 surgical tool classes, and the second or fine CNN was a regression network for tracking of the tool tip area. This was tested on six different datasets—the first five were in-house surgical videos and the sixth was the Endo-Vis 2015 challenge dataset. Their method was compared with four other methods—Fast R-CNN with filter tracking in convolutional features using VGGNet, data-driven visual tracking, tracking with an active testing filter, and tracking with online multiple instance learning.

Zhao et al. (2019b) developed an automatic real-time method for two-dimensional tool detection and tracking based on a spatial transformer network (STN) and spatio-temporal context (STC), and tested this on eight video datasets from in-house surgical videos. The authors tested their method and four other solutions—correlation filter tracking with convolutional features using VGGNet, data-driven visual tracking, tracking with an active testing filter and tracking with online multiple instance learning—on these datasets. Lu et al. (2020) tested a two deep neural networks framework for surgical tool tracking on the Surgical Perception (SuPer) and Hamlyn Centre Video Datasets. Using these datasets and a two CNN pipeline, a Pyramid Stereo Matching Network (Chang & Chen, 2018) was used to find and match features for stereo reconstruction, and DeepLabCut (Mathis et al., 2018) was used to detect point features for surgical tool tracking.

8 Tool Segmentation Research

For tool segmentation work, Luengo et al. (2021) added pixel-wise semantic annotations for anatomy and also surgical tools for 4670 images from 25 videos of the CATARACTS training set. This CATARACTS Semantic Segmentation dataset was used for the EndoVis 2020 challenge. Chen et al. (2021) developed a method that was based on exploiting cross-consistency in microscopic image segmentation, and used the consistency between the main decoder and auxiliary decoder to leverage unlabeled images. This was used to improve the Deeplabv3 plus network and was tested on the CATARACTS-Semantic-Segmentation 2020 data set. Zisimopoulos et al. (2017) used a FCN-VGG network that was trained to perform supervised semantic segmentation in 14 classes that represented the different tools present in

their simulated cataract dataset. This dataset was used to train CNN models and then transfer learning techniques were used for training on the CATARACTS Dataset. Fox et al. (2020) used the CaDIS and the Cataract-101 dataset with Mask R-CNN to localize and segment surgical tools in ophthalmic cataract surgery. They compared four backbone networks (Inceptionv2, Inception-ResNetv2, ResNet50, and ResNet101—all with pre-trained COCO (Lin et al., 2014) weights—and different data augmentation strategies for multi-class instance segmentation of surgical tools. Grammatikopoulou et al. (2019) developed the CaDIS dataset for semantic segmentation in cataract surgery, based on the CATARACTS dataset. Pissas et al. (2021) highlighted that the main issue in using the CaDIS dataset was the extreme class imbalance in the granular semantic segmentation labels, and they addressed this challenge with two data oversampling strategies. They demonstrated that the choice of the loss function and data sampling strategy were paramount in training their ResNet based encoder-decoder networks.

Ross et al. (2019) discussed segmentation solutions based on the ROBUST-MIS challenge, including the use of Mask R-CNN (He et al., 2017), a Dense Pyramid Attention Network (Li et al., 2018), a Refined Attention Segmentation Network (RASNet), a residual 2D U-Net (Ronneberger et al., 2015), DeepLabV3+ (Chen et al., 2017a), TernausNet (Iglovikov & Shvets, 2018), and Mask R-CNN with FlowNet2 (Ilg et al., 2017). Best results were reported by the U-Net based solutions. Jha et al. (2021b) tested a dual decoder attention network (DDANet) and nine different methods on the ROBUST-MIS dataset. They reported that the DDANet architecture provided the highest metric and best real-time performance over the other methods. Ceron et al. (2021) introduced a YOLACT architecture for real-time instance segmentation of surgical instruments, and tested its accuracy on the ROBUST-MIS dataset. They used criss-cross attention modules (CCAMs) with a ResNet-101 backbone to develop three models—CCAM-Backbone, CCAM-FPN and CCAM-Full—plus a baseline YOLACT++ model. Isensee and Maier-Hein (2020) relied on a 2D U-Net architecture that used residual blocks in the encoder and generated segmentation maps at several resolutions in the convolutional based decoder architecture. This method achieved a mean Dice score of 87.41 (94.35) on the ROBUST-MIS dataset. Sahu et al. (2021) used a teacher-student learning approach that learned from annotated simulation data and unlabeled real data. They redesigned their Endo-Sim2Real framework based on a teacher-student approach, and used a TerNaus11 as the backbone segmentation model. They tested this on a simulated dataset as well as on the Robust-MIS, EndoVis 2015 and Cholec80 datasets.

Allan et al. (2020) reported segmentation results using the EndoVis18 dataset. The solutions included the use of the ResNeXt-101 architecture with Squeeze-Excitation blocks; U-Net architecture with a VGG 19 encoder; a global convolutional network (GCN) with ResNet 152 backbone; DeepLab V3+ using multi-scale feature extraction with Xception and atrous convolutions; WideResnet38 encoder and activated batch norm (ABN) with DeepLab V3 as decoder; two ResNet encoder blocks and a stacked convolutional decoder network with a sum-skip connection; 3 U-Net models with final prediction as an ensemble; a 77 layer fully convolutional dense network architecture; DeepLab V3+ and ResNet-50 pre-trained on ImageNet; a U-Net with a ResNet-101 backbone; and a Pix2Pix model for the segmentation with a U-Net as the generator. Most of the architectures were pre-trained on ImageNet. Gonzalez et al. (2020) extended the EndoVis 2018 dataset for fine-grained instrument segmentation by manually annotating each instrument in the dataset, and used this dataset to successfully test their ISINet model which was based on Mask R-CNN.

Shvets et al. (2018) experimented with the U-Net, TernausNet and LinkNet encoder-decoder architectures on the EndoVis 2017 dataset. TernausNet was shown to outperform the other architectures in all three tasks of binary, part-based and type-based segmentation. Hasan and Linte (2019) used U-Net but modified it to U-NetPlus model by introducing both VGG11 and VGG16 as an encoder with batch-normalized pre-trained weights and nearest-neighbour interpolation as the replacement of the transposed convolution in the decoder layer. This was tested on the EndoVis 2017 dataset. Mohammed et al. (2019) proposed a multi encoder and single decoder convolutional neural network, which they termed StreoScenNet. The architecture consisted of two ResNet50 encoder blocks, pre-trained on ImageNet, and a stacked convolutional decoder network connected with a sum-skip connection. The input to the encoder was a set of left and right frames, and the output of the decoder was a mask for the instrument, part and binary segmentation tasks. This was tested on the EndoVis 2017 dataset. Zhang et al. (2021b) proposed a GAN-based method for unpaired image-to-image translation (I2I), and used it for surgical tool image segmentation and repair. They tested this on three endoscopic surgery datasets and on the EndoVis17 dataset. Kong et al. (2021) optimised Mask R-CNN model with anchor optimization and improved Region Proposal Network for surgical instrument segmentation. They evaluated their architecture on the EndoVis17 and an in-house hysterectomy dataset.

Kurmann et al. (2021) proposed a encoder–decoder network for segmentation and classification of surgical instruments in endoscopic images. Their "segment first, classify last" approach used a shared encoder, two decoders for instance segmentation, and a classifier for instance classification, and it provided good results on the EndoVis 2017 dataset. Ni et al. (2019) introduced a Refined Attention Segmentation Network (RASNet)—based on ResNet-50 pre-trained on ImageNet—to simultaneously segment and

classify surgical instruments. An Attention Fusion Module (AFM) was used to fuse multi-level features by utilizing the global context of high-level features as guidance information, and this was tested on EndoVis 2017. Islam et al. (2019) developed a light-weight cascaded convolutional neural network to segment surgical instruments from the EndoVis 2017 data. The authors developed a Multi-resolution Feature Fusion (MFF) block to fuse feature maps from their auxiliary and main branches, and combined auxiliary loss and adversarial loss to regularize the segmentation model. A spatial pyramid pooling unit was used to aggregate rich contextual information in their intermediate stage. Islam et al. (2021) proposed a Spatio-Temporal Multi-Task Learning (ST-MTL) model with a shared encoder and spatio-temporal decoders for real-time surgical instrument segmentation and tested it on EndoVis 2017. Comparative tests were also conducted on other models using identical pre-processing and augmentation techniques. Lee et al. (2019b) presented a "Two-phase Deep learning Segmentation for Laparoscopic Images" (TDSLI) model and tested it on the EndoVis 2017 dataset and an additional dataset of four retrospectively collected laparoscopic image sequences in different animal surgeries. The LinkNet-34 network was used in a convolutional encoder-decoder architecture, with a pre-trained ResNet-34 network used for the encoder.

Jha et al. (2021a) released the "Kvasir-Instrument" dataset with annotated bounding box and segmentation masks of GI diagnostic and surgical tools, and tested it using the U-Net and DoubleUNet architectures for semantic segmentation. Andersen et al. (2021) reported the success of Mobile-U-Net for the segmentation of surgical tools and suture needles, and tested it on a laboratory dataset and JIGSAWS (Gao et al., 2014) dataset. Choi et al. (2021) used the YOLOv4 and YOLACT-based models for real-time object detection and semantic segmentation of six surgical tools in a mastoidectomy surgery dataset. Zadeh et al. (2020) used a gynaecological dataset to train Mask R-CNN, which was then tested on laparoscopic images from 2 additional surgeries not included in the training set. Qin et al. (2020) used the EndoVis 2017 dataset and the Sinus-Surgery-C Dataset for evaluation of DeepLabv3+ with ResNet-50 and MobileNet, TernausNet with VGG-16, and LWANet with MobileNet with a Multi-Angle Feature Aggregation (MAFA) method. Qin et al. (2019) used a similar setup to the Sinus-Surgery-C Dataset, and a ToolNet-C segmentation model—designed by cascading a feature extractor and a pixel-wise segmentor—was trained to learn features from the unlabelled images and segmentation from the small number of labelled images. Rocha et al. (2019) deployed a two-step algorithm for surgical tool segmentation using kinematic information and tested it on several phantom and in vivo robotic endoscopy datasets. Kalavakonda et al. (2019) evaluated three different deep architectures for binary segmentation—using U-Net, UNet-VGG16 and UNet-MobileNetV2 (Sandler et al., 2018)—on the NeuroID dataset and the EndoVis 2017 dataset.

Jin et al. (2019) leveraged instrument motion information for accurate surgical tool segmentation. The model worked by integrating prior knowledge from motion flow into a temporal attention pyramid network (MF-TAPNet) for surgical instrument segmentation in minimally invasive surgery video. Kletz et al. (2019b) used a ResNet50 architecture as a backbone network with a feature pyramid network (FPN) for instance segmentation task using images of gynaecological surgeries. They also fine-tuned a Mask R-CNN (He et al., 2017) model for seven instrument classes (including "BG" or Background) using a pre-trained model on the COCO dataset. VGG, PSP (Zhao et al., 2016), UPerNet (Xiao et al., 2018) and DeepLab (Chen et al., 2016) were trained and evaluated for anatomical understanding, instrument identification and tracking, and understanding of interactions between surgical instruments and anatomical landmarks.

Sahu et al. (2020) used two datasets—Cholec80 and EndoVis 2015—to test their Endo-Sim2Real method for instrument segmentation. TerNaus11 was used as the DNN model for the instrument segmentation task. Kanakatte et al. (2020) proposed a pixel-wise instance segmentation algorithm for the segmentation and localisation of surgical tool using a spatio-temporal deep network, and tested it on Cholec80. Their model used ResNet pre-trained on ImageNet database and Inflated Inception 3D (I3D) pre-trained on the ImageNet and Kinetics datasets (Kay et al., 2017) to capture spatio-temporal features. They also implemented and tested U-Net and Mask R-CNN on their annotated Cholec80 dataset.

9 Tool Pose Estimation Research

Laina et al. (2017) modelled the tool segmentation and pose estimation problem as a heatmap regression where every pixel represented a confidence proportional to its proximity to the correct landmark location. For encoding, ResNet-50 pre-trained on ImageNet was used and three different CNN variants were defined for the decoding task. The model was tested on the RMIT and EndoVis 2015 datasets. Du et al. (2018) added detailed annotations to existing labels for the RMIT and EndoVis 2015 datasets, and tested a framework with a fully convolutional detection-regression network for articulated multi-instrument 2-D pose estimation. Kayhan et al. (2019) proposed a lightweight deep attention based network architecture and evaluated three SSL algorithms for a deep attention based semi-supervised 2D-pose estimation method for urgical instruments: mean teacher, virtual adversarial training and pseudo-labelling. Analysis was conducted n the RMIT and EndoVis 2015 datasets. A modified U-Net architecture (DAU-Net) that made use of attention echanisms

was used to find each tool joint location via a heatmap output channel.

Kugler et al. (2020a) introduced three datasets: two synthetic Digitally Rendered Radiograph (DRR) Datasets (the first with a screw and the second with two surgical instruments), and a real X-ray Dataset (with manually labelled screws). They used this for a three step approach for surgical pose estimation including the application of a convolutional neural network based on a VGG architecture for information extraction, and then pose reconstruction from pseudo-landmarks. Kugler et al. (2020b) used two of these datasets to test an automatic framework (AutoSNAP) for the discovery of neural network architectures for instrument pose estimation, leading to the development of an improved architecture (SNAPNet).

Hasan et al. (2021) developed a CNN they called ART-Net, for Augmented Reality Tool Network, and combined it with an algebraic geometry approach for generic tool detection, segmentation, and 3D pose estimation. While the CNN ART-Net was used for surgical tool detection and segmentation, geometric primitives were also extracted to compute the 3D pose with algebraic geometry. Gessert et al. (2018) addressed surgical tool pose estimation from optical coherence tomography (OCT) volume data with a deep learning-based tracking framework called Inception3D. The 3D CNN architecture was used to learn accurate regression between volumetric images and object poses, and was then used to estimate object pose from new volumetric images.

10 Open Research Questions

We address our research questions by presenting a comprehensive review of surgical tool datasets. A knowledge hierarchy of machine learning research was then developed using these datasets. However, while robustness or the reliable performance of methods on challenging images has been addressed in the work, there are important questions and research gaps that need to be addressed. These issues are discussed in this section.

10.1 Data Modalities

As we have found in our survey, RGB images or video are the predominant data modalities in the datasets. This is a well understood modality, and it is easy to deploy cameras to capture entire room images, high level views of the procedures, specific images of body parts, or even for internal imaging through endoscopes (Maier-Hein et al., 2020). However, there are many more medical modalities that can be explored for creation of rich and representative datasets. A limited amount of work using other images modalities is reported, and this includes radiograph and X-Ray (Kugler et al., 2020a), optical coherence tomography (OCT) (Gessert et al., 2018), RGB-D depth (Hiasa et al., 2016), and 3D ultrasound images (Yang et al., 2019). Multi-modal datasets could potentially be valuable—for example, in their review of surgical activity recognition research, van Amsterdam et al. (2021) reported that multi-modal data integration demonstrated promising results on small surgical datasets. While image modalities tend to be specific to surgical areas, there are some modalities that could foster innovative work in the surgery domain—for example, the use of IR images to supplement standard RGB images could address issues with illumination and reflection, and could lead to more accurate models being developed. Similarly, depth images could assist in addressing surgical tool counting problems and for segmenting tools from complex and crowded backgrounds.

10.2 Dataset Volume, Variety and Quality

In a white paper on the first annual Conference on Machine Intelligence in Medical Imaging (C-MIMI), Kohli et al. (2017) discussed the impact on machine learning performance due to the unavailability of large and high-quality training data. The lack of data for medical image evaluation with machine learning is a key concern, to the extent that the term "data starved" was used to describe the state of current research in this area. Similarly, van Amsterdam et al. (2021) stated that the availability of large and diverse open-source datasets of annotated data was essential for the development and validation of robust solutions in the surgery domain. A further challenge in medical surgery domains is the great variety of surgeries and the rapid rate of change (i.e. new techniques and tools) which increases the chance that a medical dataset will become obsolete, a problem that is generally not present in traditional object detection domains.

In a workshop on Surgical Data Science (SDS), Maier-Hein et al. (2020) discussed the lack of success stories in surgery, and contrasted it to success with machine learning research in other medical areas, such as radiology, dermatology, gastroenterology and mental health. This lack of success was directly attributed to the lack of quality annotated data, representative of the surgery domain. Participants in the workshop cited the EndoVis, Cholec80 and JIGSAWS datasets as being useful for research but the small size and limited representation provided by the datasets—even in these major initiatives—was reported to be a core issue. It was stated that creating and providing access to larger, more-representative and fully annotated datasets would lead to improved outcomes and success stories in the application of machine learning to surgery.

Bouget et al. (2017) reviewed the surgical tools used in different setups and for different procedures and found that two categories of surgical tools emerged: articulated instruments and rigid instruments. This survey also found two such

categories into which most works fell—we categorise them as either laparoscopic instruments or open surgery tools. Table 7 indicates that the overwhelming majority of work in this area has focused on laparoscopic surgery, and open surgery has received considerably less attention. Even the work that has been accomplished in open surgery focuses on very few instruments; the majority of work detects less than 10 instruments and even the Cataracts dataset provides only 21 instruments (Table 1). There are tens of thousands of instruments in circulation in a hospital at any one time and we would also expect tools to change over time or new tools to be introduced due to new technology or innovations in surgical techniques. Clearly, therefore, larger datasets are required and it would be useful for the research community if more open surgical tool datasets are made available.

Ideally, a surgical tool dataset should have large data volume, expert annotations, reliable ground-truth, and reusability. An issue is the size of available datasets, the benchmark dataset—ImageNet—has 14 million categorized images in a hierarchical arrangement. By contrast, most medical image datasets are limited to hundreds of cases, and datasets with thousands of annotated images are very limited (Maier-Hein et al., 2020). A valuable initiative would be to create and curate a large surgical tool dataset of tens of thousands of tool images across surgical specialities with different modalities of image capture. Further, all the datasets surveyed in our paper have a flat structure. Given that fact that surgery is organised along specialities (Table 7), and each speciality has separate underlying categories, a hierarchical classification of surgical tools in the datasets provided for machine learning research has been shown to be extremely valuable (Rodrigues et al., 2022, 2021a, b).

10.3 Dataset Bias and Generalisation

A major problem highlighted by Barbu et al. (2019) is that most datasets are highly biased. The objects of interest were generally highly correlated with the image backgrounds and objects were presented in stereotypical orientations with limited occlusions and under standardised illumination conditions. These biases were problematic because training on these datasets did not transfer well to real world data where there were variable views, orientations, backgrounds and illumination (Barbu et al., 2019), and there is limited research that tests or addresses this problem. In our survey, we found that benchmark datasets capture very specific image types with similar backgrounds, modalities, controlled collection methods, identical contexts and annotations. A key concern expressed in the literature is about algorithms which are trained on a specific dataset, procedure, intervention or in specific institution being able to generalise to other datasets and procedures (Ross et al., 2019).

To ensure viewpoint invariant object detection, different angles, scales, background clutter, illumination, orientation, pose, occlusion and intra-class variations should be captured in the images. Generalisation can be estimated by conducting research across different datasets using the same model. For example, Sahu et al. (2020) tested the Endo-Sim2Real model for instrument segmentation across two datasets—Cholec80 and EndoVis 2015, Zhao et al. (2019a) tested their method on the EndoVis Challenge dataset and the ATLAS Dione dataset, and Kalavakonda et al. (2019) evaluated three different deep architectures—U-Net, VGG16 and MobileNetV2—on their NeuroID dataset and on the EndoVis 2017 dataset. Du et al. (2018) and Kayhan et al. (2019) developed machine learning solutions and tested them on the RMIT and EndoVis 2015 datasets. More research initiatives across datasets to evaluate issues such as how accuracy or performance changes from one dataset to another, or the dependence of performance on camera or image quality, is essential.

More research is also required across the fourteen surgical specialities as listed in Table 7, since the current research is limited in scope and scale and only addresses a few specialities, but to accomplish this, better surgical tool datasets need to be made available.

10.4 Issues with Annotations

Maier-Hein et al. (2014) highlighted the fact that the performance of deep learning classifiers are heavily dependent on the availability of relevant annotations, and point out that such annotations are difficult and expensive to obtain because they need medical expertise and experience. Since medical resources for this task are limited, available datasets for deep learning are typically small and unable to cover the required range of variance for training deep learning systems for medical applications.

Orting et al. (2020) hypothesised that the high costs associated with annotations is a factor in the limited availability of large-scale, well-annotated datasets. They reviewed 57 papers that used crowd-sourcing for the analysis of medical images and for labelling large quantities of data. They reported that 42% of the papers they surveyed focused on classification, 39% on localisation or segmentation, 12% on both classification and segmentation, and a further 7% on other tasks—each task required specific annotations to be performed, with varying degrees of complexity and difficulty. Hein et al. (2018) state that deep learning based techniques for medical applications require huge amounts of accurate reference segmentation annotations, and completing manual annotations is extremely time consuming. The authors state that crowd-sourcing could result in accurate and cost-effective annotations for radiology images, and showed that even non-experts were able to complete high quality image segmentation in the medical domain.

Table 7 Specialities addressed in the research

Speciality	Open surgery	Laparoscopic	References
Cardiothoracic surgery		✓	Lu et al. (2020)
Colon and rectal surgery		✓	Maier-Hein et al. (2021) and Ross et al. (2019)
General surgery	✓	✓	Jha et al. (2021a), Gao et al. (2014), Murillo et al. (2017), Bar et al. (2020), Hong et al. (2020), Hou et al. (2022), Wagner et al. (2021) and Twinanda et al. (2017)
Gynaecology and obstetrics		✓	Gruijthuijsen et al. (2021), Hasan et al. (2021), Leibetseder et al. (2018), Meeuwsen et al. (2019) and Zadeh et al. (2020)
Gynaecologic oncology			
Neurological surgery		✓	Bouget et al. (2015), Kalavakonda et al. (2019), Leppanen et al. (2018) and Ramesh et al. (2021a)
Ophthalmic surgery	✓		Grammatikopoulou et al. (2019), Schoeffmann et al. (2018), Kurmann et al. (2017), N. et al. (2022), Sznitman et al. (2012) and Al Hajj et al. (2019)
Oral and maxillofacial surgery			
Orthopaedic surgery		✓	Hossain et al. (2018)
Otorhinolaryngology		✓	Kugler et al. (2020a) and Qin et al. (2020)
Paediatric surgery			
Plastic and maxillofacial surgery			
Urology		✓	Sarikaya et al. (2017), Law et al. (2017), and Nakawala et al. (2019)
Vascular surgery			

Nogueira-Rodriguez et al. (2020) reported that the publicly available datasets that could be used for object detection all annotated the object locations as binary masks. These masks were directly used for deep learning solutions but could also be converted to bounding boxes if required for specific training strategies. Annotation costs also vary across types of surgery—for example, annotation of surgical tools in cataract surgery needs to specify if the tool is actually in use or in contact with the eyeball, and this requires expert annotators to define (Al Hajj et al., 2019). This is expensive and tedious, but other surgery types only define the presence of the object in the frame, therefore needing simpler, cheaper annotations. In general terms and as Garcia-Peraza-Herrera et al. (2021) point out, manual annotation of pixel-level segmentation labels is difficult, expensive, tedious and time-consuming, this has led to a shortfall in the availability of quality datasets for deep learning. Since there are no large datasets available for tasks such as deep learning based surgical instrument-background segmentation, advancement in this area has been significantly curtailed.

Ward et al. (2021a) discussed the challenges in annotating spatial, temporal, and clinical elements of surgical videos, and in achieving consistency and reliability of annotations across the data. They also highlighted the requirement for achieving consensus in the development and use of surgical annotations. Meireles et al. (2021) studied current practices in surgical video annotation, and proposed recommendations for the annotation process. This is an on-going effort to create a general framework of recommendations to facilitate uniform annotations and to improve cross-institutional research efforts. Initial recommendations appear to call for increased detail in annotation—for example, to include hierarchical

information of surgical tools, anatomy, and tissue types, as well as for patient-specific factors and intra-operative influencing factors in the annotations.

Kohli et al. (2017) pointed out that there are no generally accepted standards for the creation and cataloguing of medical image datasets. As we demonstrate in Table 6, surgical tool dataset collection, curation and use is typically provided as a one-off solution, directly linked to a specific research project. The metadata provided with these datasets, if at all available, is all too often limited in description, incomplete and inconsistent. Specific domain and speciality expertise as well as knowledge of the context and institution is required to make sense of the data provided. In our Table 5, we provide metadata for the important publicly available machine learning datasets that address surgical tool tasks, more information would be useful and this is perhaps a starting point for future work to make datasets more understandable and useful (Kohli et al., 2017).

10.5 Metrics

There are an extremely wide range of metrics that have been used in the research (Fig. 8). Reinke et al. (2018) reported 14 different metric used by the MICCAI in 75 grand challenges held between 2007 and 2016. The range of metrics, variety of approaches and different reporting criteria made it difficult to directly compare results. For example, Zhang and Gao (2020) reported sensitivity, specificity, dice similarity coefficient (DSC) and model inference time (MIT) for their work on the m2cai2016 dataset, while other researchers reported the Mean Average Precision. Zia et al. (2016) tested AlexNet, VGG and Inception of the m2cai2016 dataset but pointed out that comparisons were not fair since the first two architecture were tested by removing one of the 10 videos, while the third architecture was tested by randomly selecting a percentage of the input data for testing and validation. A standard set of metrics, consistent and fixed splits of datasets into, for example, training, validation and testing, and standard metrics for evaluation would be useful for future research but it is difficult to make a hard recommendation since this is very task and context specific.

10.6 MLOps and Federated Learning

Given the mission critical nature of surgical tool management in a hospital, the deployment of deep learning systems in real time—or MLOps—needs to be addressed (Makinen et al., 2021). We have highlighted the tremendous progress that has been made in the application of deep learning models to surgical tool management in this survey, but the deployment, integration, adoption and testing of such systems in actual hospital conditions remains a significantly under-explored area due to the lack of data, the general messiness or poor

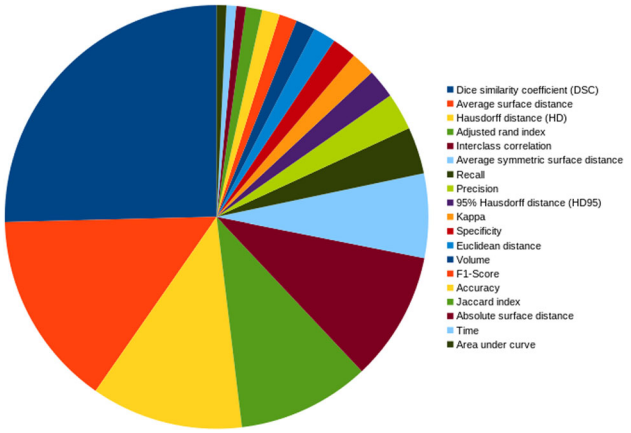

Fig. 8 Range of metrics used

usability of data, and the inaccessibility of data (Makinen et al., 2021). Making sure that consistently high-quality data is available for MLOps, while ensuring coverage of all data cases and creating data annotations that are consistent, is therefore a critical task (Ng, 2021).

Given the fact that the surgical tool datasets used for deep learning are generally small in size, private in nature and distributed across many institutions, federated learning may offer a way to overcome the size and accessibility barrier. With federated learning, local data can be used for local training, and this can then be aggregated with other locally trained models for deep learning (Zhang et al., 2021a). Rieke et al. (2020) highlighted the fact that health related data is difficult to obtain, sensitive in nature, strongly controlled by privacy and other regulations, is expensive to collect, curate and maintain, and therefore generally not available on the scale needed for training deep learning models. Whatever medical data is available tends to be very task- or disease-specific, and of limited utility given license restrictions. Demonstrating the practicality of this approach for biomedical research, Silva et al. (2019) developed a federated learning framework for the analysis of multi-centric, multi-database sub-cortical brain data.

Table 8 summarises the open research questions and opportunities which we identified and detailed in previous sections of this paper.

11 HOSPI-Tools Dataset

The currently available datasets used for surgical tool recognition offer a limited range of instruments to work with, with a maximum of 21 instruments, but—as we have identified in our review—better datasets are required for research. To help in addressing these challenges, we created a new surgical tool dataset named **HOSPITools**—"**H**ierarchically **O**rganised **S**urgical **P**rocedure **I**nstruments and **Tools**" (Rodrigues et

Table 8 Open research questions (ORQs)

No	Research gaps and questions
ORQ 1	Generalisation of algorithms across contexts and dataset
ORQ 2	Open source datasets for surgical tool research—high volume, bias-free, multi-modal with comprehensive coverage of all surgical specialities
ORQ 3	High quality annotations and metadata for datasets
ORQ 4	Standardised taxonomy, metrics, collection, cataloguing and curation of datasets
ORQ 5	Hierarchical machine learning
ORQ 6	MLOps and federated learning

Table 9 HOSPI-Tools Dataset details

Characteristic	Specification
Specialities	Orthopaedic and general surgery
Data type	40,000 images
Data quality	6000 × 4000 pixels
Modality	RGB-DSLR Camera
Location	Hospital Lab (Sterile Services Unit)
Background	Flat colours
Illumination	Sunlight, LED, halogen and fluorescent lighting
Distance	60–150 cms
Instruments	360
Images/class	74 images
Organisation	Hierarchical
Annotations	Various—image labels, bounding boxes and masks

al., 2022, 2021a, b). We created an initial dataset of surgical instrument images: over forty thousand images of surgical tools were captured using under different lighting conditions and with different backgrounds. Meireles et al. (2021) point out that surgical instruments can present significant differences due to their function, and intended possible uses, as well as due to manufacturing variations. They therefore recommended hierarchical annotation at two levels—the general and the specific instrument type—so that research can address device-related complications or surgical issues stemming from any particular device, the outcome from specific instrument choices, and the use of instruments in different surgical procedures. Since instruments could be used for multiple purposes, the authors recommended that additional labels be added to instrument annotations. We instead built the hierarchical structure directly into our dataset and created a four level hierarchy which consisted of speciality (2 classes), pack (12 classes), set (35 classes) and tool (360 classes) levels. We believe that this approach can be valuable for deep learning research and this dataset was therefore designed to offer a large variety of tools, arranged hierarchically to reflect how surgical tools are organised in real-world conditions. We provide details of the HOSPI-Tools Dataset in Table 9, and examples of actual instrument sets and annotations of instruments in Figs. 9 and 10.

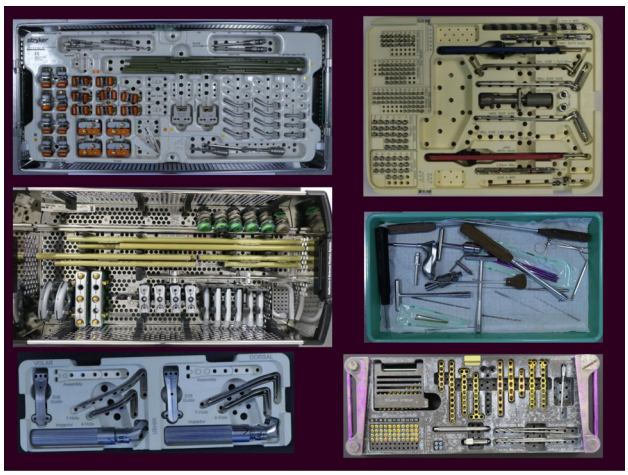

Fig. 9 HOSPI-Tools sets

Images captured included individual object images as well as cluttered, clustered and occluded objects. More images need to be taken by adjusting the DSLR camera position and pose—this would increase the realism and utility of the dataset. Instrument images were captured before and after

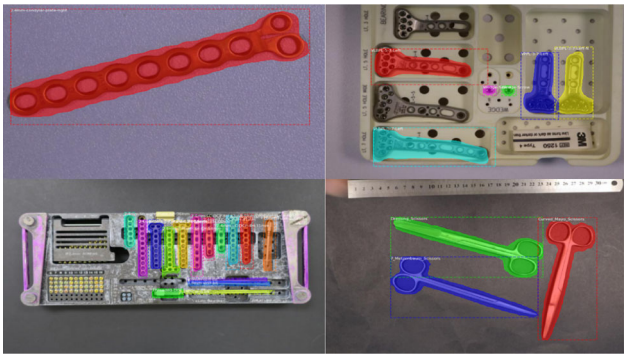

Fig. 10 HOSPI-Tools annotations

use in surgery, it was not possible to take images of the tools in use during actual surgery. Our survey findings have highlighted the need to include more images with occlusions, illumination changes, and the presence of blood, tissue and smoke, to accurately capture complex surgery conditions.

This is one step in the direction of addressing the issues that we have identified in this survey, but much more work needs to be accomplished. We will add other specialities as we develop this dataset, to reflect the complexities inherent in each of the surgical specialities and to address the open research issues and challenges.

12 Conclusions

We presented a comprehensive survey of datasets for surgical tool detection and related surgical data science and machine learning techniques and algorithms. We offered a high level perspective of current research in this area, analysed the taxonomy of approaches adopted by researchers using surgical tool datasets, and addressed key areas of research, such as the datasets used, evaluation metrics applied and deep learning techniques utilised. To ensure that we were rigorous and structured in our approach, we defined an a priori protocol for discovering and selecting the research that we reviewed. Adherence to this protocol prevented any mid-stream shifting of goals and inclusion criteria, and ensured that we presented a comprehensive and robust knowledge hierarchy.

Our survey shows that the application of machine learning to surgical tool detection, localisation, tracking, segmenting and pose estimation is a well explored research subject and many innovative techniques have been applied. However, we also identified and discussed the open research issues and challenges. To help address some of the gaps and shortfalls that we have identified, we make a contribution by creating a new Surgical Tool Dataset and we make this dataset publicly available to encourage more work in this direction.

The dataset is available at: https://doi.org/10.5281/zenodo.5895068.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

ACS. (2021). *What are the surgical specialties?*. Retrieved February 15, 2021 from https://www.facs.org/education/resources/medical-students/faq/specialties.

Ahmadi, E., Masel, D. T., Metcalf, A. Y., & Schuller, K. (2018). Inventory management of surgical supplies and sterile instruments in hospitals: A literature review. *Health Systems, 2018*(8), 134–151. https://doi.org/10.1080/20476965.2018.1496875.

Al Hajj, H., Lamard, M., Conze, P. H., Roychowdhury, S., Hu, X., Maršalkaitė, G., Zisimopoulos, O., Dedmari, M. A., Zhao, F., Prellberg, J., & Sahu, M. (2019). Cataracts: Challenge on automatic tool annotation for cataract surgery. *Medical Image Analysis,52*, 24–41. https://doi.org/10.1016/j.media.2018.11.008

Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., et al. (2021). Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Medical Image Analysis*. https://doi.org/10.1016/j.media.2021.102002.

Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., & Kori, A. (2020). *2018 robotic scene segmentation challenge*. arXiv:2001.11190

Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y. H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., & Herrera, L. (2019). *2017 robotic instrument segmentation challenge*. arXiv:1902.06426

Alshirbaji, T. A., Ding, N., Jalal, N. A., & Moller, K. (2020a). The effect of background pattern on training a deep convolutional neural network for surgical tool detection. In *AUTOMED—Automation in Medical Engineering*.

Alshirbaji, T. A., Ding, N., Jalal, N. A., & Moller, K. (2020b). The effect of background pattern on training a deep convolutional neural network for surgical tool detection. *Proceedings on Automation in Medical Engineering, 1*(1), 24–024.

Alshirbaji, T. A., Jalal, N. A., Docherty, P. D., Neumuth, T., & Moeller, K. (2021a). Assessing generalisation capabilities of CNN models

for surgical tool classification. *Current Directions in Biomedical Engineering, 7,* 476–479.

Alshirbaji, T. A., Jalal, N. A., Docherty, P. D., Neumuth, T., & Moller, K. (2021b). Cross-dataset evaluation of a cnn-based approach for surgical tool detection. In *AUTOMED 2021*.

Alshirbaji, T. A., Jalal, N. A., & Moller, K. (2018). Surgical tool classification in laparoscopic videos using convolutional neural network. *Current Directions in Biomedical Engineering, 4*(1), 407–410.

Andersen, J. K. H., Schwaner, K. L., & Savarimuthu, T. R. (2021). Real-time segmentation of surgical tools and needle using a mobile-u-net. In *20th International Conference on Advanced Robotics (ICAR)*.

Attanasio, A., Scaglioni, B., Leonetti, M., Frangi, A. F., Cross, W., Biyani, C. S. & Valdastri, P. (2020). Autonomous tissue retraction in robotic assisted minimally invasive surgery—A feasibility study. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (Vol. 5, pp. 6528–6535). https://doi.org/10.1109/LRA.2020.3013914

Banerjee, N., Sathish, R., & Sheet, D. (2019). Deep neural architecture for localization and tracking of surgical tools in cataract surgery. *Computer Aided Intervention and Diagnostics in Clinical and Medical Images, Lecture Notes in Computational Vision and Biomechanics, 31,* 31–38. https://doi.org/10.1007/978-3-030-04061-1_4.

Bar, O., Neimark, D., Zohar, M., Hager, G. D., Girshick, R., Fried, G. M., Wolf, T., & Asselmann, D. (2020). Impact of data on generalization of AI for surgical intelligence applications. *Scientific Reports*. https://doi.org/10.1038/s41598-020-79173-6

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., & Katz, B. (2019). Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in neural information processing systems*, Vol. 32 (NeurIPS 2019).

Bhatt, N., Dunne, E., Khan, M. F., Gillis, A., Conlon, K., Paran, S., & Ridgway, P. (2018). Trends in the use of laparoscopic versus open paediatric appendicectomy: A regional 12-year study and a national survey. *World Journal of Surgery, 42,* 3792–3802.

Bodenstedt, S., Ohnemus, A., Katic, D., Wekerle, A.L., Wagner, M., Kenngott, H., & Speidel, S. (2018). *Real-time image-based instrument classification for laparoscopic surgery*. arXiv:1808.00178

Bouget, D., Allan, M., Stoyanov, D., & Jannin, P. (2017). Vision-based and marker-less surgical tool detection and tracking: A review of the literature. *Medical Image Analysis, 35,* 633.

Bouget, D., Benenson, R., Omran, M., Riffaud, L., Schiele, B., & Jannin, P. (2015). Detecting surgical tools by modelling local appearance and global shape. *IEEE Transactions on Medical Imaging, 34*(12), 2603–2617.

Ceron, J. C. A., Chang, L., Ruiz, G. O., & Ali, S. (2021). Assessing yolact++ for real time and robust instance segmentation of medical instruments in endoscopic procedures. In *Annual international conference IEEE engineering in medicine biology society*.

Chai, J., Zeng, H., Li, A., & Ngai, E. W. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*. https://doi.org/10.1016/j.mlwa.2021.100134.

Chang, J. R., & Chen, Y. S. (2018). Pyramid stereo matching network. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 5410–5418).

Chen, H., Ma, X., Xia, T., & Jia, F. (2021). Semi-supervised semantic segmentation of cataract surgical images based on deeplab v3+. In *ICCDA 2021: 2021 the 5th international conference on compute and data analysis*.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2016). *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs*.

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017a). *Rethinking atrous convolution for semantic image segmentation*. arXiv:1706.05587v3

Chen, Z., Zhao, Z., & Cheng, X. (2017b). Surgical instruments tracking based on deep learning with lines detection and spatio-temporal context. In *Proceedings of IEEE*, CAC Jinan, China, p. 2711.

Choi, B., Jo, K., & Choi, S. J. Choi (2017). Surgical-tools detection based on convolutional neural network in laparoscopic robot-assisted surgery. In (ny): IEEE NY (Ed.) *Proceedings of annual international conference of the IEEE engineering in medicine and biology society* (pp. 1756–1759).

Choi, J., Cho, S., Chung, J., & Kim, N. (2021). Video recognition of simple mastoidectomy using convolutional neural nets: Detection and segmentation of surgical tools and anatomic regions. *Computer Methods and Programs in Biomedicine*. https://doi.org/10.1016/j.cmpb.2021.106251.

Ciaparrone, G., Bardozzo, F., Priscoli, M.D., Kallewaard, J. L., Zuluaga, M. R., & Tagliaferri, R. (2020). A comparative analysis of multi-backbone mask r-cnn for surgical tools detection. In *International joint conference on neural networks (IJCNN)*. https://doi.org/10.1109/IJCNN48605.2020.9206854

Colleoni, E., Edwards, P., & Stoyanov, D. (2020). Synthetic and real inputs for tool segmentation in robotic surgery. In *Medical image computing and computer assisted intervention—MICCAI 2020, 23rd international conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III* (pp. 700–710). https://doi.org/10.1007/978-3-030-59716-0_67

Colleoni, E., Moccia, S., Du, X., De Momi, E., & Stoyanov, D. (2019). Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. *IEEE Robotics and Automation Letters, 4*(3), 2714–2721.

Dergachyova, O., Bouget, D., Huaulme, A., Morandi, X., & Jannin, P. (2016). Automatic data-driven real-time segmentation and recognition of surgical workflow. *International Journal of Computer Assisted Radiology and Surgery, 11*(6), 1081–1089.

Du, X., Kurmann, T., Chang, P. L., Allan, M., Ourselin, S., Sznitman, R., Kelly, J., & Stoyanov, D. (2018). Articulated multi-instrument 2-d pose estimation using fully convolutional networks. *IEEE Transactions on Medical Imaging,37*, 5.

Egger, J., Gsaxner, C., Pepe, A., & Li, J. (2020). *Medical deep learning—A systematic meta-review*. arXiv:2010.14881

Fox, M., Taschwer, M., & Schoeffmann, K. (2020). Pixel-based tool segmentation in cataract surgery videos with mask r-cnn. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*.

Gao, Y., Vedula, S., Reiley, C., Ahmidi, N., Varadarajan, B., Lin, H., Tao, L., Zappella, L., Bejar, B., Yuh, D., Chen, C., Vidal, R., Khudanpur, S., & Hager, G. (2014). The JHU-ISI gesture and skill assessment dataset (JIGSAWS): A surgical activity working set for human motion modeling. In Modeling and monitoring of computer assisted interventions (M2CAI)—MICCAI Workshop, 2014.

Garcia-Peraza-Herrera, L., Li, W., Fidon, L., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Poorten, E., 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). IEEE, Vancouver, Canada.

Garcia-Peraza-Herrera, L. C., Fidon, L., D'Ettorre, C., Stoyanov, D., Vercauteren, T., & Ourselin, S. (2021). Image compositing for segmentation of surgical tools without manual annotations. *IEEE Transactions on Medical Imaging, 40,* 1450–1460.

Garrow, C. R., Kowalewski, K. F., Li, L. B., et al. (2021). Machine learning for surgical phase recognition: A systematic review. *Annals of Surgery, 273,* 684–693.

Gessert, N., Schlüter, M., & Schlaefer, A. (2018). A deep learning approach for pose estimation from volumetric oct data. *Medical Image Analysis, 46,* 162–179.

Girshick, R. (2015). Fast R-CNN. In *2015 IEEE international conference on computer vision (ICCV)* (pp. 1440–1448). https://doi.org/10.1109/ICCV.2015.169.

Gonzalez, C., Bravo-Sanchez, L., & Arbelaez, P. (2020). Isinet: An instance-based approach for surgical instrument segmentation. In *Medical image computing and computer assisted intervention MICCAI 2020*. https://doi.org/10.1007/978-3-030-59716-0_57

Grammatikopoulou, M., Flouty, E., Kadkhodamohammadi, A., Quellec, G. E., Chow, A., Nehme, J., Luengo, I., & Stoyanov, D. (2019). *Cadis: Cataract dataset for image segmentation*. arXiv:1906.11586

Gruijthuijsen, C., Garcia-Peraza-Herrera, L. C., Borghesan, G., Reynaerts, D., Deprest, J., Ourselin, S., Vercauteren, T., & Vander Poorten, E. (2021). *Robotic endoscope control via autonomous instrument tracking*. arXiv:2107.02317

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing, 187,* 27–48. https://doi.org/10.1016/j.neucom.2015.09.116.

Hasan, S. K., & Linte, C. A. (2019). U-netplus: A modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images. In *Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*.

Hasan, M. K., Calvet, L., Rabbani, N., & Bartoli, A. (2021). Detection, segmentation, and 3d pose estimation of surgical tools using convolutional neural networks and algebraic geometry. *Medical Image Analysis*. https://doi.org/10.1016/j.media.2021.101994.

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. In *International I (ed) conference on computer vision (ICCV)* (pp. 2961–2969).

Hein, E., Rob, T., Seitel, A., März, K., Stieltjes, B., Eisenmann, M., Lebert, J., Metzger, J., Sommer, G., Sauter, A. W., & Schwartz, F. R. (2018). Large-scale medical image annotation with crowd-powered algorithms. *Journal of Medical Imaging*. https://doi.org/10.1117/1.JMI.5.3.034002

Hiasa, Y., Suzuki, Y., Reiter, A., Otake, Y., Nishi, M., Harada, H., Koyama, K., Kanaji, S., Kakeji, Y., & Sato, Y. (2016). Segmentation of surgical instruments from rgb-d endoscopic images using convolutional neural networks: Preliminary experiments towards quantitative skill assessment. In *Proceedings of medical and biological imaging—JSMBE 2016/3*.

Hong, W. Y., Kao, C. L., Kuo, Y. H., Wang. J. R., Chang, W. L., & Shih, C. S. (2020). *Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80*. arXiv:2012.12453

Hossain, M., Nishio, S., Hiranaka, T., & Kobashi, S. (2018). Real-time surgical tools recognition in total knee arthroplasty using deep neural networks. In *2018 Joint 7th international conference on informatics vision and pattern recognition (icIVPR) and 2018 2nd international conference on imaging electronics and vision (ICIEV)* (pp. 470–474).

Hou, Y., Zhang, W., Liu, Q., Ge, H., Meng, J., Zhang, Q., & Wei, X. (2022). Adaptive kernel selection network with attention constraint for surgical instrument classification. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-021-06368-x.

Hu, X., Yu, L., Chen, H., Qin, J., & Heng, P. (2017). Agnet: Attention-guided network for surgical tool presence detection. In *Deep learning in medical image analysis and multimodal learning for clinical decision support. Lecture notes in computer science*, Cham (pp. 186–194).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2261–2269). https://doi.org/10.1109/CVPR.2017.243

Huaulmé, A., Sarikaya, D., Le Mut, K., Despinoy, F., Long, Y., Dou, Q., Chng, C. B., Lin, W., Kondo, S., Bravo-Sánchez, L., & Arbeláez, P. (2021). Micro-surgical anastomose workflow recognition challenge report. *Computer Methods and Programs in Biomedicine*. https://doi.org/10.1016/j.cmpb.2021.106452

Iandola, F., Han, S., Moskewicz, M., Ashraf, K., Dally, W., & Keutzer, K. (2016). *Squeezenet: Alexnet-level accuracy with 50x fewer parameters and* <0.5mb *model size*. arxiv:1602.07360

Iglovikov, V., & Shvets, A. (2018). *Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation*. arXiv:1801.05746

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2017). *Flownet 2.0: Evolution of optical flow estimation with deep networks*. arXiv:1612.01925

Isensee, F., & Maier-Hein, K. H. (2020). *OR-UNet: An optimized robust residual u-net for instrument segmentation in endoscopic images*.

Islam, M., Li, Y., & Ren, H. (2019). Learning where to look while tracking instruments in robot-assisted surgery. https://doi.org/10.1007/978-3-030-32254-0_46

Islam, M., Vibashan, V., Lim, C., & Ren, H. (2021). ST-MTL: Spatio-temporal multitask learning model to predict scanpath while tracking instruments in robotic surgery. *Medical Image Analysis*. https://doi.org/10.1016/j.media.2020.101837.

Jha, D., Ali, S., Emanuelsen, K., Hicks, S., Thambawita, V., Garcia Ceja, E., Riegler, M., de Lange, T., Schmidt, P., Johansen, H., Johansen, D., & Halvorsen, P. (2021a). *Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. Multi media modeling MMM2021 lecture notes in computer science*, Vol. 12573. Springer, Cham.

Jha, D., Ali, S., Tomar, N. K., Riegler, M. A., Johansen, D., Johansen, H. D., & Halvorsen, P. (2021b). Exploring deep learning methods for real-time surgical instrument segmentation in laparoscopy. arXiv:2107.02319

Jin, Y., Cheng, K., Dou, Q., & Heng, P. A. (2019). Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In *International conference on medical image computing and computer-assisted intervention, Cham* (pp. 440–448).

Jin, A., Yeung, S., Jopling, J., Krause, J., Azagury, D., Milstein, A., & Fei-Fei, L. (2018). Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *IEEE Winter conference on applications of computer vision*. Lake Tahoe, Washington (DC), pp. 691–699.

Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C. W., & Heng, P. A. (2020). Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Medical Image Analysis, 59,* 1. https://doi.org/10.1016/j.media.2019.101572.

Jo, K., Choi, Y., Choi, J., & Chung, J. W. (2019). Robust real-time detection of laparoscopic instruments in robot surgery using convolutional neural networks with motion vector prediction. *Applied Sciences, 9*(14), 2865.

Kalavakonda, N., Hannaford, B., Qazi, Z., & Sekhar, L. (2019). Autonomous neurosurgical instrument segmentation using end-to-end learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops, Long Beach, California* (pp. 514–516). https://doi.org/10.1109/CVPRW.2019.00076

Kanakatte, A., Ramaswamy, A., Gubbi, J., Ghose, A., & Purushothaman, B. (2020). Surgical tool segmentation and localization using spatio-temporal deep network. In *2020 42nd annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, Montreal, QC, Canada.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman,

M., & Zisserman, A. (2017). *The kinetics human action video dataset*.
Kayhan, M., Kopuklu, O., Sarhan, M., Yigitsoy, M., Eslami, A., & Rigoll, G. (2019). *Deep attention based semi-supervised 2d-pose estimation for surgical instruments*. arXiv:1912.04618.
Kletz, S., Schoeffmann, K., Benois-Pineau, J., & Husslein, H. (2019). Identifying surgical instruments in laparoscopy using deep learning instance segmentation. In *International conference on content-based multimedia indexing (CBMI)* (pp. 1–6). Dublin, Ireland.
Kletz, S., Schoeffmann, K., & Husslein, H. (2019). Learning the representation of instrument images in laparoscopy videos. *Healthcare Technology Letters, 6*(6), 197–203.
Kohli, M. D., Summers, R. M., & Geis, J. R. (2017). Medical image data and datasets in the era of machine learning—White paper from the 2016 C-MIMI Meeting Dataset Session. *Journal of Digital Imaging, 30,* 392–399. https://doi.org/10.1007/s10278-017-9976-3.
Kong, X., Jin, Y., Dou, Q., Wang, Z., Wang, Z., Lu, B., Dong, E., Liu, Y. H., & Sun, D. (2021). Accurate instance segmentation of surgical instruments in robotic surgery: Model refinement and cross-dataset evaluation. *International Journal of Computer Assisted Radiology and Surgery*. https://doi.org/10.1007/s11548-021-02438-6.
Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Neural information processing systems red hook* (pp. 1097–1105). Curran Associates Inc.
Kugler, D., Sehring, J., Stefanov, A., Stenin, I., Kristin, J., Klenzner, T., & Mukhopadhyay, A. (2020a). iposnet: Instrument pose estimation from x-ray in temporal bone surgery. *International Journal of Computer Assisted Radiology and Surgery, 15*(7), 1137-1145 3.
Kugler, D., Uecker, M., Kuijper, A., & Mukhopadhyay, A. (2020b). Autosnap: Automatically learning neuralarchitectures for instrument pose estimation. In *23rd international conference medical image computing and computer assisted intervention—MICCAI 2020*, Lima, Peru.
Kurmann, T., Neila, P. M., Du, X., Fua, P., Stoyanov, D., Wolf, S., & Sznitman, R. (2017). Simultaneous recognition and pose estimation of instruments in minimally invasive surgery. In *International conference on medical image computing and computer-assisted intervention, Cham* (pp. 505–513).
Kurmann, T., Marquez-Neila, P., Allan, M., Wolf, S., & Sznitman, R. (2021). Mask then classify: Multi-instance segmentation for surgical instruments. *International Journal of Computer Assisted Radiology and Surgery*. https://doi.org/10.1007/s11548-021-02404-2.
Laina, I., Rieke, N., Rupprecht, C., Vizcaíno, J. P., Eslami, A., Tombari, F., & Navab, N. (2017). Concurrent segmentation and localization for tracking of surgical instruments. In *International conference on medical image computing and computer-assisted intervention* (pp. 664–672).
Law, H., Ghani, K., & Deng, J. (2017). Surgeon technical skill assessment using computer vision based analysis. In *Proceedings of the 2nd machine learning for healthcare conference* (Vol. 68, pp. 88–99).
Law, H., & Deng, J. (2020). Cornernet: Detecting objects as paired keypoints. *International Journal of Computer Vision, 128,* 642–656.
Lee, E. J., Plishker, W., Liu, X., Kane, T., Bhattacharyya, S. S., & Shekhar, R. (2019b). Segmentation of surgical instruments in laparoscopic videos: Training dataset generation and deep-learning-based framework. In *Medical imaging image-guided procedures, robotic interventions, and modeling* (Vol. 10951, p. 109511T). International Society for Optics and Photonics 2019.
Lee, E. J., Plishker, W., Liu, X., Bhattacharyya, S. S., & Shekhar, R. (2019). Weakly supervised segmentation for real-time surgical tool tracking. *Healthcare Technology Letters, 6*(6), 231–236.
Leibetseder, A., Petscharnig, S., Primus, M. J., 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* (pp. 357–362). ACM, NY, USA.
Leppanen, T., Vrzakova, H., Bednarik, R., Kanervisto, A., Elomaa, A. P., Huotarinen, A., Bartczak, P., Fraunberg, M., & Jääskeläinen, J. E. (2018). Augmenting microsurgical training: Microsurgical instrument detection using convolutional neural networks. In *IEEE 31st international symposium on computer-based medical systems (CBMS)* (pp. 211–216). https://doi.org/10.1109/CBMS.2018.00044
Li, H., Xiong, P., An, J., & Wang, L. (2018). Pyramid attention network for semantic segmentation. In *British machine vision conference (BMVC)*, Newcastle upon Tyne.
Lin, X. G., Chen, Y. W., Qi, B. L., Wang, P., & Zhong, K. H. (2019). Presence detection of surgical tool via densely connected convolutional networks. In: 2019 international conference on artificial intelligence and computing science (ICAICS 2019) DEStech transactions on computer science and engineering (pp. 245–253).
Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision—ECCV 2014. Lecture notes in computer science*, Vol. 8693. Springer, Cham.
Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2018.2858826.
Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis (Supplement C), 60–88,* 42.
Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International Journal of Computer Vision, 128,* 261–318.
Liu, Y., Zhao, Z., Chang, F., & Hu, S. (2020). An anchor-free convolutional neural network for real-time surgical tool detection in robot-assisted surgery. *IEEE Access*. https://doi.org/10.1109/ACCESS.2020.2989807.
Liu, Y., Zhao, Z., Chang, F., & Hu, S. (2020). An anchor-free convolutional neural network for real-time surgical tool detection in robot-assisted surgery. *IEEE Access, 8,* 78193–78201.
Lu, J., Jayakumari, A., Richter, F., Li, Y., & Yip, M.C. (2020). *Super deep: A surgical perception framework for robotic tissue manipulation using deep learning for feature extraction*. arXiv:2003.03472
Luengo, I., Grammatikopoulou, M., Mohammadi, R., Walsh, C., Nwoye, C. I., Alapatt, D., Padoy, N., Ni, Z. L., Fan, C. C., Bian, G. B., & Hou, Z. G. (2021). *2020 cataracts semantic segmentation challenge*. arXiv:2110.10965
Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., & Malpani, A. (2020). *Surgical data science—From concepts to clinical translation*. arXiv:2011.02284
Maier-Hein, L., Mersmann, S., Kondermann, D., Bodenstedt, S., Sanchez, A., Stock, C., Kenngott, H., Eisenmann, M., & Speidel, S. (2014). Can masses of non-experts train highly accurate image classifiers? a crowdsourcing approach to instrument segmentation in laparoscopic images. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, vol 17 (2), pp 438–45, https://doi.org/10.1007/978-3-319-10470-6_55.
Maier-Hein, L., Wagner, M., Ross, T., Reinke, A., Bodenstedt, S., Full, P. M., et al. (2021). Heidelberg colorectal data set for surgical data science in the sensor operating room. *Scientific Data, 8,* 1–11.
Mäkinen, S., Skogström, H., Laaksonen, E., & Mikkonen, T. (2021). Who needs mlops: What data scientists seek to accomplish and

how can mlops help? In *2021 IEEE/ACM 1st workshop on AI engineering—Software engineering for AI (WAIN)*.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). Deeplabcut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience, 21,* 1281–1289.

Matton, N., Qalieh, A., Zhang, Y., Annadanam, A., Thibodeau, A., Li, T., et al. (2022). Analysis of cataract surgery instrument identification performance of convolutional and recurrent neural network ensembles leveraging bigcat. *Translational Vision Science and Technology*. https://doi.org/10.1167/tvst.11.4.1.

Meeuwsen, F. C., van Luyn, F., Blikkendaal, M. D., Jansen, F. W., & van den Dobbelsteen, J. (2019). Surgical phase modelling in minimal invasive surgery. *Surgical Endoscopy, 33*(5), 1426–1432.

Meireles, O. R., Rosman, G., Altieri, M. S., Carin, L., Hager, G., Madani, A., et al. (2021). SAGES consensus recommendations on an annotation framework for surgical video. *Surgical Endoscopy*. https://doi.org/10.1007/s00464-021-08578-9.

Mhlaba, J. M., Stockert, E. W., Coronel, M., & Langerman, A. J. (2015). Surgical instrumentation: The true cost of instrument trays and a potential strategy for optimization. *Journal of Hospital Administration, 4,* 6. https://doi.org/10.5430/jha.v4n6p82.

Mishra, K., Sathish, R., & Sheet, D. (2017). Learning latent temporal connectionism of deep residual visual abstractions for identifying surgical tools in laparoscopy procedures. In K. Mishra (Ed.), IEEE Computer Society; 2017 (pp. 2233–2240, DC).

Mohammed, A., Yildirim, S., Farup, I., Pedersen, M., & Hovde, O. (2019). Streoscennet: Surgical stereo robotic scene segmentation. In *Medical imaging 2019: Image-guided procedures, robotic interventions, and modeling*, San Diego, California, United States, SPIE Medical Imaging. https://doi.org/10.1117/12.2512518

Mondal, S., Sathish, R., & Sheet, D. (2019). *Multitask learning of temporal connectionism in convolutional networks using a joint distribution loss function to simultaneously identify tools and phase in surgical videos*. arXiv:1905.08315

Murillo, P., Arenas, J. O. P., & Moreno, R. J. (2018). Tree-structured cnn for the classification of surgical instruments. In *International symposium on intelligent computing systems* (pp. 211–216).

Murillo, P. C. U., Moreno, R. J., & Arenas, J. O. P. (2017). Comparison between cnn and haar classifiers for surgical instrumentation classification. *Contemporary Engineering Sciences, 10*(28), 1351–1363.

Nakawala, H., Bianchi, R., Pescatori, L. E., De Cobelli, O., Ferrigno, G., & De Momi, E. (2019). "deep-onto" network for surgical workflow and context recognition. *International Journal of Computer Assisted Radiology and Surgery, 4*(4), 685–696.

Namazi, B., Sankaranarayanan, G., & Devarajan, V. (2019). *Laptoolnet: A contextual detector of surgical tools in laparoscopic videos based on recurrent convolutional neural networks*. arXiv:1905.08983

Newell, A., Yang, K., & Deng, J. (2016). *Stacked hourglass networks for human pose estimation*. arXiv:1603.06937

Ng, A. (2021). *Mlops: From model-centric to data-centric ai, 2021*. YouTube Video Interview.

Ni, Z. L., Bian, G. B., Xie, X. L., Hou, Z. G., Zhou, X. H., & Zhou, Y. J. (2019). Rasnet: Segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 5735–5738). IEEE.

Nogueira-Rodriguez, A., Dominguez, R., Lopez-Fernandez, H., Iglesias, A., Cubiella, J., Fdez-Riverola, F., et al. (2020). Deep neural networks approaches for detecting and classifying colorectal polyps. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2020.02.123.

Nwoye, C. I., Alapatt, D., Yu, T., Vardazaryan, A., Xia, F., Zhao, Z., Xia, T., Jia, F., Yang, Y., Wang, H., & Yu, D. (2021a). *Cholectriplet2021: A benchmark challenge for surgical action triplet recognition*. arXiv:2204.04746.

Nwoye, C. I., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., & Padoy, N. (2020). Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In *International conference on medical image computing and computer-assisted intervention, MICCAI 2020*.

Nwoye, C. I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., & Padoy, N. (2021b). Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Journal of Medical Image Analysis*.

Nwoye, C. I., Mutter, D., Marescaux, J., & Padoy, N. (2019). Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos. *International Journal of Computer Assisted Radiology and Surgery, 4*(6), 1059–1067.

Orting, S. N., Doyle, A., van Hilten, A., Hirth, M., Inel, O., Madan, C. R., et al. (2020). A survey of crowdsourcing in medical image analysis. *Human Computation Journal, 7*(1), 1–26. https://doi.org/10.15346/hc.v7i1.1.

Pakhomov, D., Premachandran, V., Allan, M., Azizian, M., & Navab, N. (2019). Deep residual learning for instrument segmentation in robotic surgery. In International workshop on machine learning in medical imaging (pp. 566–573).

Pissas, T., Ravasio, C., Da Cruz, L., & Bergeles, C. (2021). Effective semantic segmentation in cataract surgery: What matters most? In *Medical image computing and computer assisted intervention—MICCAI 2021. Lecture notes in computer science*.

Prellberg, J., & Kramer, O. (2018). Multi-label classification of surgical tools with convolutional neural networks. In *2018 international joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Qin, F., Li, Y., Su, Y.H., Xu, D., & Hannaford, B. (2019). Surgical instrument segmentation for endoscopic vision with data fusion of reduction and kinematic pose. In *2019 international conference on robotics and automation (ICRA)* (pp. 9821–9827). IEEE.

Qin, F., Lin, S., Li, Y., Bly, R., Moe, K., & Hannaford, B. (2020). Towards better surgical instrument segmentation in endoscopic vision: Multi-angle feature aggregation and contour supervision. *IEEE Robotics and Automation Letters, 5,* 6639–6646.

Qiu, L., Li, C., & Ren, H. (2019). Real-time surgical instrument tracking in robot-assisted surgery using multi-domain convolutional neural networks. *Healthcare Technology Letters, 6*(6), 159–164.

Raju, A., Wang, S., & Huang, J. (2016). *M2cai surgical tool detection challenge report*. Technical report. University of Texas at Arlington.

Ramesh, A., Beniwal, M., Uppar, A. M., Vikas, V., & Rao. M. (2021a). Microsurgical tool detection and characterization in intra-operative neurosurgical videos. In *43rd annual international conference of the IEEE engineering in medicine and biology society (EMBC)*.

Ramesh, S., Dall'Alba, D., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., Fiorini, P., & Padoy, N. (2021b). Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. *International Journal of Computer Assisted Radiology and Surgery,16*, 1111–1119. https://doi.org/10.1007/s11548-021-02388-z

Redmon, J., & Farhadi, A. (2017). Yolo9000: Better, faster, stronger. In *IEEE conference on computer vision and pattern recognition* (pp. 6517–6525). IEEE Computer Society, Washington, DC.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *IEEE conference on computer vision and pattern recognition* (pp. 779–788). IEEE Computer Society, Washington, DC.

Reinke, A., Eisenmann, M., Onogur, S., Stankovic, M., Scholz, P., Full, P., Bogunovic, H., Landman, B., & Maier, O. (2018). How to exploit weaknesses in biomedical challenge design and organi-

zation. In *International conference on medical image computing and computer-assisted intervention*, Granada, Spain.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137–1149.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., & Ourselin, S. (2020). The future of digital health with federated learning. *Digital Medicine,3*, 119. https://doi.org/10.1038/s41746-020-00323-1

Rocha, C., Padoy, N., & Rosa, B. (2019). Self-supervised surgical tool segmentation using kinematic information. In *International conference on robotics and automation (ICRA)* (pp. 8720–8726). IEEE.

Rodrigues, M., Mayo, M., & Patros, P. (2021a). Evaluation of deep learning techniques on a novel hierarchical surgical tool dataset. In *2021 Australasian joint conference on artificial intelligence*.

Rodrigues, M., Mayo, M., & Patros, P. (2021b). Interpretable deep learning for surgical tool management. In M. Reyes, P. H. Abreu, J. Cardoso, M. Hajij, G. Zamzmi, P. Rahul, & L. Thakur (Ed.), *4th international workshop on interpretability of machine intelligence in medical image computing (iMIMIC 2021). Lecture Notes in Computer Science*, Vol. 12929. Springer, Cham. https://doi.org/10.1007/978-3-030-87444-5_1

Rodrigues, M., Mayo, M., & Patros, P. (2022). Octopusnet: Machine learning for intelligent management of surgical tools. *Smart Health*. https://doi.org/10.1016/j.smhl.2021.100244.

Rojas, E., Couperus, K., & Wachs, J. (2020). *DAISI: Database for AI surgical instruction*. arXiv:2004.02809

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention (MICCAI). LNCS* (Vol. 9351, pp. 234–241). Springer.

Ross, T., Reinke, A., & Full, P.M. (2019). Robust medical instrument segmentation challenge. arXiv:2003.10299

Roychowdhury, S., Bian, Z., Vahdat, A., & Macready, M. (2017). *Identification of surgical tools using deep neural networks*. Technical report, D-Wave Systems Inc.

Sahu, M., Stromsdorfer, R., Mukhopadhyay, A., & Zachow, S. (2020). Endo-Sim2Real: Consistency learning-based domain adaptation for instrument segmentation. In Medical image computing and computer assisted intervention—MICCAI (Vol. 2020, pp. 784–794).

Sahu, M., Dill, S., Mukhopadyay, A., & Zachow, S. (2017). Surgical tool presence detection for cataract procedures. *ZIB Report, 2017,* 30–11.

Sahu, M., Mukhopadhyay, A., Szengel, A., & Zachow, S. (2017). Addressing multi-label imbalance problem of surgical tool detection using cnn. *International Journal of Computer Assisted Radiology and Surgery, 12,* 6.

Sahu, M., Mukhopadhyay, A., & Zachow, S. (2021). Simulation-to-real domain adaptation with teacher-student learning for endoscopic instrument segmentation. *International Journal of Computer Assisted Radiology and Surgery, 16,* 849–859.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Sarikaya, D., Corso, J. J., & Guru, K. A. (2017). Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. *IEEE Transactions on Medical Imaging, 36*(7), 1542–1549. https://doi.org/10.1109/TMI.2017.2665671.

Schoeffmann, K., Taschwer, M., Sarny, S., Münzer, B., Primus, M. J., & Putzgruber, D. (2018). Cataract-101—video dataset of 101 cataract surgeries. In *MMSys'18: 9th ACM multimedia systems conference*, June 12–15, 2018, Amsterdam, Netherlands.

Shimizu, T., Hachiuma, R., Kajita, H., Takatsume, Y., & Saito, H. (2021). Hand motion-aware surgical tool localization and classification from an egocentric camera. *Journal of Imaging*. https://doi.org/10.3390/jimaging7020015.

Shvets, A. A., Rakhlin, A., Kalinin, A. A., & Iglovikov, V. I. (2018). Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)* (pp. 624–628). IEEE.

Silva, S., Gutman, B., Romero, E., Thompson, P., Altmann, A., & Lorenzi, M. (2019). Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In *2019 IEEE 16th international symposium on biomedical imaging*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. In *Proceedings of IEEE conference on computer vision and pattern recognition*. IEEE Computer Society, Washington, DC.

Stockert, E. W., & Langerman, A. J. (2014). Assessing the magnitude and costs of intraoperative inefficiencies attributable to surgical instrument trays. *Journal of the American College of Surgeons, 219*(4), 646–655. https://doi.org/10.1016/j.jamcollsurg.2014.06.019.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016a). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI'17: Proceedings of the thirty-first AAAI conference on artificial intelligence* (pp. 4278–4284).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–9), Boston, MA, USA.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016b). Rethinking the inception architecture for computer vision. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2818–2826). https://doi.org/10.1109/CVPR.2016.308

Sznitman, R., Ali, K., Richa, R., Taylor, R., Hager, G., & Fua, P. (2012). Data-driven visual tracking in retinal microsurgery. In *MICCAI-2012*.

Tang, E. M., El-Haddad, M. T., Patel, S. N., & Tao, Y. K. (2022). Automated instrument-tracking for 4d video-rate imaging of ophthalmic surgical maneuvers. *Biomedical Optics Express*. https://doi.org/10.1364/BOE.450814.

Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., & Padoy, N. (2017). Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging, 36,* 86–97. https://doi.org/10.1109/TMI.2016.2593957.

van Amsterdam, B., Clarkson, M. J., & Stoyanov, D. (2021). IEEE Transactions on Biomedical Engineering. *Gesture recognition in robotic surgery: A review., 68*(6), 2021–2035. https://doi.org/10.1109/TBME.2021.3054828.

Vardazaryan, A., Mutter, D., Marescaux, J., & Padoy, N. (2018). Weakly-supervised learning for tool localization in laparoscopic videos. In I. Imaging & C. Assisted (Eds.), *Stenting and large-scale annotation of biomedical data and expert label synthesis* (pp. 169–179). Springer.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society conference on computer vision and pattern recognition. CVPR 2001* (p. I-I), Kauai, HI, USA.

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2018/7068349.

Wagner, M., Müller-Stich, B. P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D. M., Müller, B., Davitashvili, T., Capek, M., & Reinke, A. (2021). Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. arXiv:2109.14956

Wang, S., Xu, Z., Yan, C., & Huang, J. (2019). Graph convolutional nets for tool presence detection in surgical videos. In *Information processing in medical imaging IPMI 2019 lecture notes in computer science, Vol. 11492* (vol. 10, no (1007), pp. 1–36). Springer, Cham.

Ward, T. M., Fer, D. M., Ban, Y., Rosman, G., Meireles, O. R., & Hashimoto, D. A. (2021a). Challenges in surgical video annotation. *Computer Assisted Surgery,26*(1), 58–68. https://doi.org/10.1080/24699322.2021.1937320

Ward, T. M., Mascagni, P., Ban, Y., Rosman, G., Padoy, N., Meireles, O., & Hashimoto, D. A. (2021b). Computer vision in surgery. *Surgery,169,* 1253–1256.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *ACM international conference proceeding series,* (Vol. 10, p. 1145).

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., & Sun, J. (2018). Unified perceptual parsing for scene understanding. In: V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer vision and pattern recognition. Lecture notes in computer science*, Vol. 11209. Springer, Cham. https://doi.org/10.1007/978-3-030-01228-1_26

Xue, Y., Liu, S., Li, Y., Wang, P., & Qian, X. (2022). A new weakly supervised strategy for surgical tool detection. *Knowledge-Based Systems, 239,* 107860.

Yamazaki, Y., Kanaji, S., Matsuda, T., Oshikiri, T., Nakamura, T., Suzuki, S., et al. (2020). Automated surgical instrument detection from laparoscopic gastrectomy video images using an open source convolutional neural network platform. *Journal of the American College of Surgeons*. https://doi.org/10.1016/j.jamcollsurg.2020.01.037.

Yang, H., Shan, C., Tan, T., & Kolen, A. F. (2019). Transferring from ex-vivo to in-vivo: Instrument localization in 3d cardiac ultrasound using pyramid-unet with hybrid loss. In *International conference on medical image computing and computer-assisted intervention* (pp. 263–271). Cham.

Yang, C., Zhao, Z., & Hu, S. (2020). Image-based laparoscopic tool detection and tracking using convolutional neural networks: A review of the literature. *Computer Assisted Surgery, 25*(1), 15–28.

Yu, F., Wang, D., Shelhamer, E., & Darrell, T. (2018). Deep layer aggregation. In *IEEE/CVF conference on computer vision and pattern recognition, Salt Lake City, UT, USA,* (Vol. 2018, pp. 2403–2412).

Zadeh, S. M., Francois, T., Calvet, L., Chauvet, P., Canis, M., Bartoli, A., & Bourdel, N. (2020). Surgai: Deep learning for computerized laparoscopic image understanding in gynaecology. *Surgical Endoscopy, 34*(12), 5377–5383.

Zhang, Z., Rosa, B., & Nageotte, F. (2021b). Surgical tool segmentation using generativeadversarial networks with unpaired training data. *IEEE Robotics and Automation Letters,6,* 6266–6273.

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021a). A survey on federated learning. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2021.106775.

Zhang, J., & Gao, X. (2020). Object extraction via deep learning-based marker-free tracking framework of surgical instruments for laparoscope-holder robots. *International Journal of Computer Assisted Radiology and Surgery, 15,* 1335.

Zhao, Z., Cai, T., Chang, F., & Cheng, X. (2019a). Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade. *Healthcare Technology Letters,6*, 6.

Zhao, Z., Chen, Z., Voros, S., & Cheng, X. (2019b). Real-time tracking of surgical instruments based on spatio-temporal context and deep learning. *Computer Assisted Surgery,24*, 20–29.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2016). Pyramid scene parsing network. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 6230–6239). https://doi.org/10.1109/CVPR.2017.660

Zhao, Z., Voros, S., & Chen, Z. (2019c). Cheng X (2019c) Surgical tool tracking based on two CNNs: from coarse to fine. *The Journal of Engineering,14*, 467–472.

Zhao, Z., Voros, S., Weng, Y., Chang, F., & Li, R. (2017). Tracking-by-detection of surgical instruments in minimally invasive surgery via the convolutional neural network deep learning-based method. *Computer Assisted Surgery, 22,* 26–35. https://doi.org/10.1080/24699322.2017.1378777.

Zia, A., Castro, D., & Essa, I. (2016). *Fine-tuning deep architectures for surgical tool detection*. Technical report, Georgia Institute of Technology.

Zisimopoulos, O., Flouty, E., Stacey, M., Muscroft, S., Giataganas, P., Nehme, J., & Stoyanov, D. (2017). Can surgical simulation be used to train detection and classification of neural networks? *Healthcare Technology Letters, 4*(5), 216–222.

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. (2018). Learning transferable architectures for scalable image recognition. In *2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/CVPR.2018.00907