

**International Journal of Computational Vision and Robotics**

ISSN online: 1752-914X - ISSN print: 1752-9131

<https://www.inderscience.com/ijcvr>

---

**Localisation and classification of surgical instruments in laparoscopy videos using deep learning techniques**

Avanti Bhandarkar, Priyanka Verma

**DOI:** [10.1504/IJCVR.2023.10057447](https://doi.org/10.1504/IJCVR.2023.10057447)

**Article History:**

Received:	15 September 2022
Last revised:	26 February 2023
Accepted:	25 April 2023
Published online:	02 December 2024

---

## Localisation and classification of surgical instruments in laparoscopy videos using deep learning techniques

---

Avanti Bhandarkar\*

Department of Electronics and Telecommunication Engineering,  
Mukesh Patel School of Technology Management and Engineering,  
Mumbai, Maharashtra, India  
Email: [avantibhandarkar@gmail.com](mailto:avantibhandarkar@gmail.com)

\*Corresponding author

Priyanka Verma

Department of Electronics and Telecommunication Engineering,  
Mukesh Patel School of Technology Management and Engineering,  
Mumbai, Maharashtra, India  
Email: [priyanka.verma@nmims.edu](mailto:priyanka.verma@nmims.edu)

**Abstract:** Surgical trainees often use laparoscopic surgery videos to understand the appropriate use of instruments and visualise the surgical workflow better, but these videos may be difficult to interpret without proper annotations. In recent times, neural networks have emerged as an accurate and effective solution for instrument detection and classification in surgical video frames, which can subsequently be used to automate the annotation process. The proposed implementation uses faster-RCNNs and bidirectional LSTMs with (and without) time-distributed layers and attempts to solve some of the problems commonly faced while developing deep learning models for surgical image and video data: severe class imbalance, inaccuracies during multi-label classification and a lack of spatiotemporal context from adjacent video frames. The bidirectional LSTM with time-distributed layers achieved an average accuracy of 80.20% and an average F1 score of 0.7176 on the M2CAI16 tool dataset, while also achieving 63.49% average accuracy and an average F1 score of 0.522 on unseen data. Jaccard distance and Hamming distance have also been used as object detection-specific metrics; the same model registered the lowest values for both distances, implying accurate localisation and identification of surgical instruments.

**Keywords:** deep learning; surgical instrument detection; surgical instrument classification; surgical instrument localisation; data augmentation; transfer learning; faster-RCNN; region-based convolutional neural networks; bidirectional LSTMs; long short-term memory networks; Jaccard distance; Hamming distance.

**Reference** to this paper should be made as follows: Bhandarkar, A. and Verma, P. (2025) 'Localisation and classification of surgical instruments in laparoscopy videos using deep learning techniques', *Int. J. Computational Vision and Robotics*, Vol. 15, No. 1, pp.75–103.

**Biographical notes:** Avanti Bhandarkar is an incoming graduate student in the Department of Electrical and Computer Engineering at University of California, San Diego. She completed her Bachelor of Technology from the NMIMS University, Mumbai, in 2022. Her areas of interest lie at the intersection of image processing and machine learning, with a focus on their application in healthcare systems and medical analysis.

Priyanka Verma is an Assistant Professor in the EXTC Department at MPSTME, NMIMS University. She has 13+ years of teaching experience and is currently pursuing her PhD in Medical Imaging in collaboration with the Tata Memorial Hospital, Mumbai. Her research areas are the use of machine learning and deep learning, specifically towards medical applications.

## 1 Introduction

Laparoscopic or minimally invasive surgery is performed using tiny cuts or incisions as compared to traditional ‘open’ surgeries that require long incisions (Velanovich, 2000). The smaller incisions result in less pain after surgery, shorter hospital stays, faster recovery, and almost imperceptible scars (Varela et al., 2010; Stiff et al., 1994). These benefits have made laparoscopic surgery a preferred option for managing several abdominal and pelvic conditions requiring surgical treatment (Jönsson and Zethraeus, 2000; Cuschieri, 2005). Surgeons use a laparoscope: a long, thin, tubular telescope attached to a high-resolution camera and a high-intensity light source. Images obtained from within the abdominal cavity are displayed on a monitor and the surgeon operates by looking at this live image. Footage of the surgeries is often recorded and can be used subsequently for training purposes, allowing trainee surgeons to visualise and understand the complex surgical procedures.

**Figure 1** Parts of a standard laparoscope (see online version for colours)

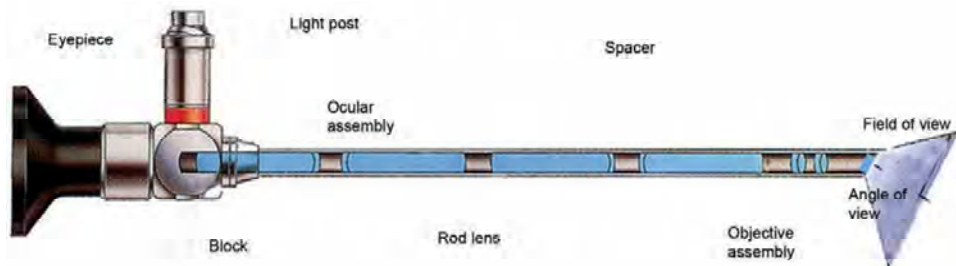


Figure 1 (Yeola et al., 2017) shows a Hopkins rod lens telescope (Dobson and Hopkins, 1989; Ellis, 2007), its field of view (FoV), and its angle of view. The laparoscope’s FoV is a result of its angle of view; a forward oblique angle of the laparoscope tip gives surgeons a better FoV. The diameter of the laparoscope also affects the area visible to the surgeon. Figure 2 is a frame taken from a laparoscopic surgery video and shows the anatomy and instruments visible to the surgeon from the laparoscopic camera feed. Apart from the laparoscope, a variety of other instruments are used to carry out the various steps of the operation. These include graspers, dissectors, clip applicators, hooks, surgical

scissors, etc. These instruments are specially designed to be thin, long, and easy to handle inside the narrow abdominal cavity. In recent times, robotic arms have been used to assist surgeons, especially in surgeries where additional dexterity or precision is required (Taylor and Jayne, 2007; Peters et al., 2018; Longmore et al., 2020).

**Figure 2** Surgeon's view from the laparoscope camera feed (see online version for colours)



As with any other surgery, there are some key challenges faced by surgeons while performing laparoscopic procedures (Ballantyne, 2002). As the surgeon operates looking at a 2D view of the 3D organs, the depth of vision is compromised (Tanagho et al., 2012). Furthermore, the tactile feedback (from tissues and organs) that a surgeon is used to during open surgery is reduced to a minimum when he uses thin, long laparoscopic instruments. As the laparoscope provides a 'tunnel' vision of the operative field, organs at the periphery tend to stay hidden and create a 'blind spot' (Trilling et al., 2021). The small incisions allow only narrow-diameter laparoscopes to be used, resulting in a limited FoV. This view may be further reduced if the laparoscope is not angled correctly. Finally, the quality of the video transmitted by the laparoscopic camera may drop due to the light being absorbed by the blood in the operative field. All these aspects must be considered while implementing a system for the analysis of surgical instruments to obtain a robust model which performs well on real-time data. Taking these challenges into consideration, it is important to identify the instruments being used in each surgical phase and determine their location and orientation with respect to the surrounding anatomy (Allan et al., 2012). The primary applications of analysing the presence and location of surgical instruments are as follows:

- 1 Surgery videos are an important aspect of laparoscopic training (Mota et al., 2018; de'Angelis et al., 2019; Celentano et al., 2019), as surgical trainees use them to understand contextual anatomy (Fitzpatrick et al., 2001) and safe surgical actions (Levy and Mobasher, 2017); spotlighting the correct use and orientation of various instruments would add to the trainees' understanding of the same.
- 2 A surgeon's technique can be evaluated in terms of the correct use of each instrument and the proximity of instruments to vital organs (Baghdadi et al., 2019; Li et al., 2019). A good surgical technique avoids damage to important structures, which may be close to sharp instruments or instruments with electric current flowing through them.

- 3 Understanding the surgical workflow and identifying surgical phases by correlating them with a frame-level presence of instruments (Loukas, 2018; Nakawala et al., 2019; Kitaguchi et al., 2020).

Traditional instrument localisation and classification methods such as physical tagging, image processing-based feature extraction, and haptic signal analysis have drawbacks in terms of efficiency and scope of utility. A modern technique for instrument identification that has gained ground is computer assisted intervention or CAI (Peter et al., 2019). It is the use of digital tools and technology to perform tasks that are challenging for humans, or tasks that can be performed better by a computer as compared to a human. Deep learning and computer vision have several applications in CAI; these include their use in robotic surgeries, computer-assisted expert systems, biopsy guidance systems, diagnostic assistance systems, etc. (Esteva et al., 2021). The computer vision framework proposed by this paper focuses on the task of object detection (Bhagya and Shyna, 2019; Jiao et al., 2019), which consists of a combination of object localisation and classification with various surgical instruments as the objects to be detected.

Despite the popularity of deep learning networks for instrument detection and classification, the nature of laparoscopic surgeries and imaging limitations of laparoscopes result in the following common challenges:

- 1 Class imbalance in the dataset – Certain instruments (graspers and hooks) are most frequently observed in images, which corresponds to their multi-purpose and frequent use in surgeries.
- 2 Presence of multiple instruments in a single frame – A surgeon may use several instruments during a given surgical stage, with two or more of the same instruments being used at the same time; this converts the multiclass problem into a multi-label classification problem.
- 3 Instrument instances are highly correlated – The presence or absence of instruments in a single frame is highly correlated with the surgical phase and can result in redundancy in data in neighbouring frames extracted from a surgery video.
- 4 Physical impedances – Although laparoscopic surgery videos are captured by high-resolution cameras, it is nearly impossible to avoid physical impedances such as gas, occlusions, light reflections, body fluids, etc., both during the operation and during post-surgery video analysis.
- 5 Motion of camera – Unlike other surgeries like cataract surgery where the camera is kept as still as possible, cameras in laparoscopies are frequently moved by surgical assistants to give the surgeon the best view of certain anatomy; this may result in motion blur and frequent shifts in perspective and scale of instruments and anatomy.
- 6 Size and quality of dataset – The dataset affects the performance of the deep learning model significantly. An ideal dataset is of adequate size, with reduced redundancy of data and accurate annotations.

Keeping these challenges in mind, the two-fold objective of this paper is to introduce a computer vision framework that solves some of the previously described problems and to evaluate this framework on noisy, unseen data to determine its suitability for potential real-time applications. We have specifically focused on the lack of temporal context

during training, i.e., a situation in which each frame is analysed and processed in isolation with no information from adjacent frames being considered and our contributions are as follows:

- 1 A single shot detector-bidirectional long-short term memory (LSTM) model is proposed which uses time-distributed data in the spatial feature extraction stage. The LSTM will provide greater temporal context to a given video frame and result in reduced erroneous classifications of instruments with similar handles and different tips, especially in cases where instrument tips are hidden by anatomical structures or other instruments.
- 2 The problem of class imbalance in the M2CAI16 dataset was mitigated through a combination of synthetic minority oversampling technique (SMOTE) oversampling of the majority class and difference hash (Dhash) signature-based under-sampling of the minority class.
- 3 The proposed system was trained and tested on the M2CAI16 tool dataset to obtain baseline observations before being tested on previously unseen surgical video frames. These frames are intended to simulate real-time data the system may receive and to test the robustness of the system.

The rest of this paper is organised as follows. In Section 2, we undertake a comprehensive study of previous work in the domain of instrument detection, tracing the transition from traditional and modern methods. In Section 3, we discuss the details of the proposed system including the dataset used, pre-processing performed, and deep learning models developed. The evaluation metrics and results are presented in Section 4 and Section 5 concludes the paper, with a summary of our work and proposed future improvements.

## 2 Prior work

The earliest effort made towards detecting and tracking instruments in laparoscopy surgery videos was in the form of attaching physical markers such as distinct coloured or patterned tags and LEDs (Speidel et al., 2014; Tonet et al., 2007; Casals et al., 1996; Krupa et al., 2003) onto the instrument. While these methods were simple and cheap, they also made the instrument bulkier and less flexible, which made them difficult to manoeuvre within the narrow abdominal cavity. Alternatively, electromagnetic field sensing phenomenon may be used to sense surgical instruments; every surgical instrument has a unique radio frequency identification (RFID) which acts as the signature of the device, and the RFID signal is identified by the RFID antenna associated with the instrument (Miyawaki et al., 2009; Kranzfelder et al., 2013). The performance of this technique was satisfactory, but it suffered from the same physical drawbacks as the previous method along with the unnecessary and avoidable cost incurred while tagging several sets of instruments with RFID tags. It also required the existing instrument to be modified to accommodate RFID tags and antennas. To overcome these shortcomings, image processing-based extraction of salient features was combined with regression and classification models to localise instruments (Bouget et al., 2017). This method, while

less intrusive and safer than the physical marking of instruments, had a disadvantage in terms of the amount of pre-processing required. Features such as colours, shapes, pixel intensities, edges, etc. had to be carefully selected for each image which added to the detection time and made the method unsuitable for real-time applications. The poor video quality in the perioperative stage complicated the feature extraction as well. However, breakthroughs in the processing power of machines, high-speed computations, and computer vision techniques have made this alternative more promising. Ultrasound signals collected over the course of surgery may also be used to measure the position and orientation of instruments (Tatar et al., 2003; Greenish et al., 2002), but this technique suffers from interference due to noisy signals which reduces localisation accuracy. Moreover, the signal method was only able to localise and orient instruments, not classify them.

Recently, deep learning approaches such as image classification, image segmentation, and object detection have become very popular for the analysis of medical images (Cheplygina et al., 2018; Litjens et al., 2017). This is primarily due to the increase in computational power and graphic processing capacity of modern machines, and the advantage that deep learning models have over traditional methods, i.e., they can automatically detect the most optimal high-level features from all the training images given to them. The trend of using deep learning for surgical instrument detection and classification tasks started with the development of a convolutional neural network-based model called EndoNet (Twinanda et al., 2016), which was based on the AlexNet (Krizhevsky et al., 2017) architecture and was one of the first deep learning models for detecting surgical instruments in laparoscopy videos. Other work towards using CNNs and transfer learning for instrument detection included the use of AlexNet as the base like EndoNet while addressing key gaps in it (Sahu et al., 2016), and comparing the performance of AlexNet, VGG16 and Inception V3 architectures for the task of surgical workflow analysis (Zia et al., 2016). CNNs were also used in combination with specifically extracted features for best results; this was seen in the use of CNNs, and Hough lines (Cai and Zhao, 2020) and the use of fully convolutional networks combined with key point features for optical flow tracking and enhanced instrument localisation (García-Peraza-Herrera et al., 2016). EndoRCN (Jin et al., 2016) and LapTool-Net (Namazi et al., 2022) were used to extract spatiotemporal features between frames using RCNs and region-based CNNs (RCNNs), respectively; these features were then used to perform an analysis of surgical workflow within the video and localise instruments based on the same. One of the authors of EndoNet also experimented with HMMs and RNNs to utilise temporal features extracted from surgical videos (Twinanda, 2017). Mishra et al. (2017) were the first to feed features obtained from a CNN (ResNet50) into an LSTM for tool detection in a video sequence as compared to prior work on still frames. A unique method for spatiotemporal analysis, using a combination of video-level features (from an inflated inception 3D model) and frame-level features (from a ResNet) was proposed by Kanakatte et al. (2020) in the form of the ST-LSTM. The dual-level feature method was also used by Jin et al. (2018) along with an RCNN; this was also one of the first real-time trials of a proposed system for surgical tool analysis.

**Table 1** Comparison of selected, relevant prior work

<i>Paper</i>	<i>Objective</i>	<i>Model(s)</i>	<i>Dataset</i>	<i>Metrics</i>
Twimanda et al. (2016)	Surgical phase recognition and instrument detection	EndoNet: AlexNet + hierarchical hidden Markov model	Cholec80 (introduced) and EndoVis	% accuracy, mean average precision (mAP) score, precision, recall, no. of phases identified correctly
Zia et al. (2016)	Surgical workflow detection by utilising instrument presence	Comparison of VGG16, AlexNet and InceptionV3	M2CAI16 (derived from Cholec80)	% accuracy, average mAP score, top-3 mAP score
Jin et al. (2016)	Surgical workflow detection by utilising instrument presence	EndoRCN: RCNNs and LSTM	M2CAI16	Jaccard score
Mishra et al. (2017)	Instrument detection and segmentation with class imbalance mitigation	CNN + LSTM (with varying transfer learning models)	ILSVRC and M2CAI16	% accuracy, prediction error
Wang et al. (2017)	Instrument detection	Ensemble model (GoogleNet and VGG19)	M2CAI16	mAP score
Jin et al. (2018)	Instrument detection and real-world spatial localisation	Faster RCNN with a VGG16 base	M2CAI16 and M2CAI16-tool location (introduced)	Spatial level per-class average precision, Frame level presence detection average precision, mAP score
Alshirbaji et al. (2018)	Instrument detection with class imbalance mitigation	Partially retrained AlexNet	Cholec80	Imbalance ratio per-class, imbalance ratio
Cai and Zhao (2020)	Instrument detection and part-wise segmentation	Hough transform/Canny edge detection + 2 successive CNNs for instrument shaft and tip detection	EndoVisSub and standard dataset introduced by Du et al. (2016)	% accuracy, orientation error
Kanakatte et al. (2020)	Instrument detection and segmentation	ST-LSTM (Inception3D for video level features and ResNet for image level features)	Cholec80	mAP score
Yamazaki et al. (2020)	Instrument detection and model validation by surgical experts	YOLOv3	Custom dataset	Intersection over union, precision, recall, F1 score, mAP score
Sahu et al. (2021)	Domain adaptation for instrument detection	Endo-Sim2Real: joint learning using real, unlabelled data and simulated data	Simulated dataset, Cholec80 and EndoVis and RobustMIS	Dice score
Sun et al. (2021)	Real-time instrument segmentation	MobileNetV3 + ghost modules (split convolutional layers)	EndoVis	Jaccard index, Dice coefficient, Hausdorff distance
Jha et al. (2021)	Real-time instrument segmentation	DDANet: attention-based network with double decoder architecture	RobustMIS	Dice score, intersection over union
Namazi et al. (2022)	Multi-class instrument detection	LapTool-Net: modified recurrent CNN	M2CAI16 and Cholec80	% accuracy, per-class F1 score



Developing deep learning models for surgical data analysis has some inherent challenges, including severe dataset imbalance and multi-label classification problems. The multi-label classification issue was addressed by Prellberg and Kramer (2018) who used the ResNet50 architecture, Wang et al. (2017) who used an ensemble of GoogLeNet and VGG16, and Jaafari et al. (2022) who used various combinations of VGG19, Inception v4, and NASNet-A as a part of their ensemble models. Majority-class under-sampling and loss-sensitive learning were utilised by Alshirbaji et al. (2018) to develop a model known as ToolMod which addressed class imbalance within the datasets used. Yoon et al. (2020) employed a novel approach by combining semi-supervised learning with bidirectional instrument tracking to mitigate the class imbalance problem. This method independently generated class labels during training and did not limit the authors to the classes defined in datasets, thus facilitating the identification of more instruments based on minute differences in their tooltips. In the past few years, the use of retrained open-source models such as YoloV3 (Yamazaki et al., 2020) and YoloV4 (Wang et al., 2021) has resulted in faster training times and the subsequent integration of instrument identification algorithms into applications such as laparoscopic training modules and assistive surgical robots. Fathabadi et al. (2021) trained faster-RCNNs on data collected from laparoscopic surgical box-trainer devices (Grantner et al., 2019) and used the resulting model to evaluate the performance of surgical trainees at Western Michigan University. Along similar lines, Mohaidat et al. (2022) adapted the laparoscopic box-trainer for the specific task of suturing anatomy inside the body. They used a scaled version of the YoloV4 algorithm in tandem with a specialised centroid tracking system to track the motion of the suturing needle and compared the model's predictions with corresponding trainee data to assess surgical competency in the defined task. Kletz et al. (2019) were able to leverage used a modified form of RCNNs – known as mask RCNN – for the task of instrument segmentation, which built on the task of instrument localisation by generating a pixel-level mask for each instrument. This formed the basis of work done by Lee et al. (2020) and Lam et al. (2022), who applied mask RCNNs on data acquired from surgical robots to mask R-CNN to generate instrument trajectory maps. These maps were then used to quantify surgical skills through real-time motion metrics: moving distance, smoothness of instrument movement, and concentration of instrument movement. The laparoscope manipulating robot – introduced by Myo et al. (2022) – demonstrated an interesting application of surgical data analysis (specifically instrument localisation) by training a robot entirely using continuous learning algorithms and a feedback mechanism. The robot was tested on soft-tissue cadavers and was able to navigate the laparoscope automatically by using the detected instruments as anchor points. Instrument detection also played a crucial role in various robot surgical nurse systems, such as those proposed by Nadhifatul Aini et al. (2019), Badilla-Solórzano et al. (2022) and Nakano and Nagamune (2022). Compared to previous work – which used largely sanitised data – Roß et al. (2021) and Maier-Hein et al. (2021) introduced data that included visual noise in the form of surgical smoke, flecks of blood on the laparoscope, and motion blur artefacts. These datasets simulated the complex and unpredictable environments analysed by intelligent medical systems and allowed for the development of more robust models. Other salient features of recent surgical computer vision frameworks include:

- 1 The ability to generalise well when faced with unseen or rare organ configurations, i.e., domain adaptation (Azqueta-Gavaldon et al., 2020; Sahu et al., 2021; Wang et al., 2022).
- 2 The ability to transfer knowledge suitably when applied to instrument detection tasks in other types of surgeries – Lee et al. (2021) and Markarian et al. (2022) suggest use cases in neurosurgery, while Hossain et al. (2020) provide a model which works with instruments specific to orthopaedic procedures.
- 3 The ability to accurately identify subtle variations within a class of instruments – Su et al. (2023) released the first-of-its-kind dataset covering multiple scalpel types and demonstrated the use of a mask R-CNN for multi-scalpel classification and segmentation.
- 4 The delivery of rapid, yet accurate results at low computational costs (Sun et al., 2021; Jha et al., 2021).

Table 1 summarises the objectives, methodology, and results of selected works; papers were chosen to highlight the progression of models over time and acknowledge previous benchmarks.

### 3 Approach

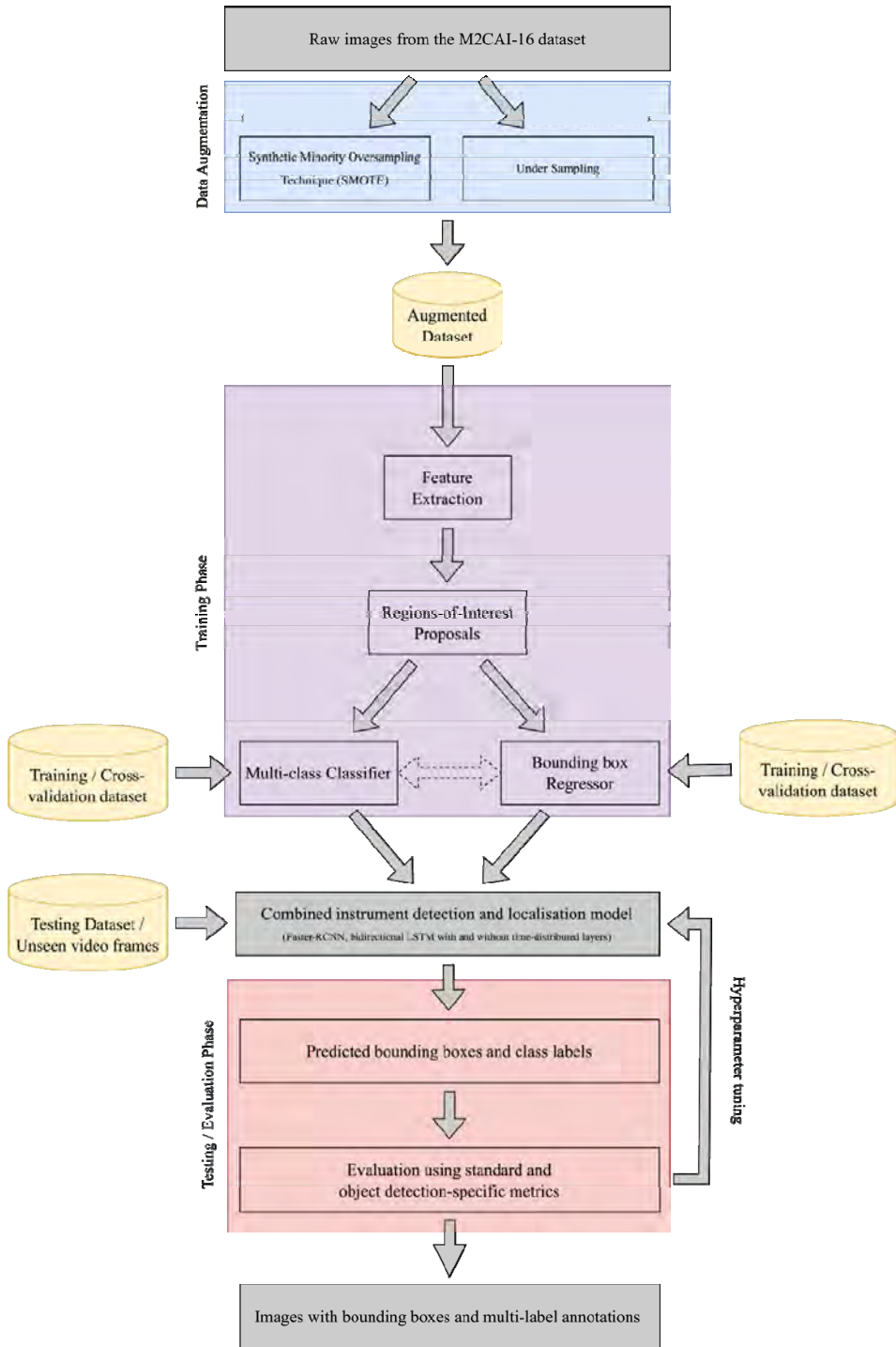
The proposed method is divided into three key phases – data augmentation (oversampling of the underrepresented classes and under-sampling of overrepresented classes), training (feature extraction, region of interest generation, classification, bounding-box regression), and testing (evaluation using standard and custom datasets, hyperparameter tuning). Figure 3 provides a broad overview of the system, while the rest of the section provides the specifics for each phase.

#### 3.1 Dataset








This implementation uses the m2cai16-tool dataset from the MICCAI (M2CAI) 2016 surgical tool detection challenge, which is derived from the Cholec80 dataset introduced by Twinanda et al. (2016) alongside their EndoNet architecture. The dataset consists of 15 videos of laparoscopic cholecystectomies performed by surgeons at the University Hospital of Strasbourg. These videos are divided into ten training videos and five testing videos and contain instruments across seven classes as shown in Table 2. The training dataset is generated by selecting and annotating a frame once every 25 frames from the (2,532 annotated frames in total). The annotations contain the following information:

- 1 image name and relative path
- 2 image length, width and depth
- 3 instrument(s) present in the image and the corresponding bounding box(es).

**Figure 3** Block diagram of the proposed method showing data augmentation, training and testing phases (see online version for colours)

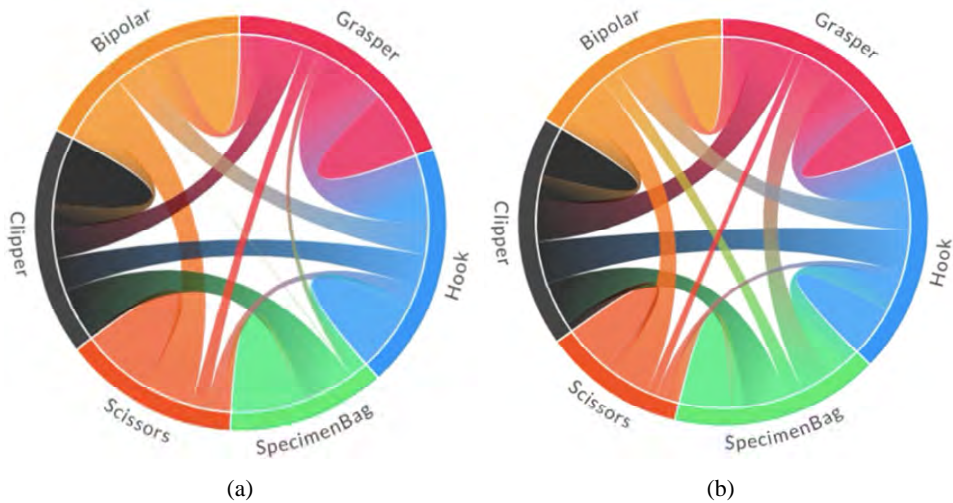


**Table 2** Ground truth labels for instruments and their frequency of occurrence within the dataset (see online version for colours)

<i>Class ID</i>	<i>Class label</i>	<i>Sample image</i>	<i>Instances in the original dataset</i>
0	NA	NA	4,647
1	Hook		21,584
2	Grasper		16,938
3	Specimen bag		1,987
4	Irrigator		1,084
5	Clipper/clip applicator		1,193
6	Bipolar (forceps/cautery)		962
7	Scissors		569

Chord diagrams were generated to visualise the correlation between a pair of instruments in the dataset. It is observed that in the chord diagram for the training dataset [Figure 4(a)], the maximum correlation is between the grasper and hook classes, with a 99% probability of them occurring in the same frame. The minimum correlation was between the specimen bag and bipolar classes, which occurred together with only 0.9% probability. Similarly, the maximum correlation in the chord diagram for the testing dataset [Figure 4(b)] was between the grasper and hook classes (99% probability), while the minimum correlation was between the clip applicators and scissors classes (16.9% probability).

**Figure 4** Chord diagrams for the (a) training data and (b) testing data (see online version for colours)



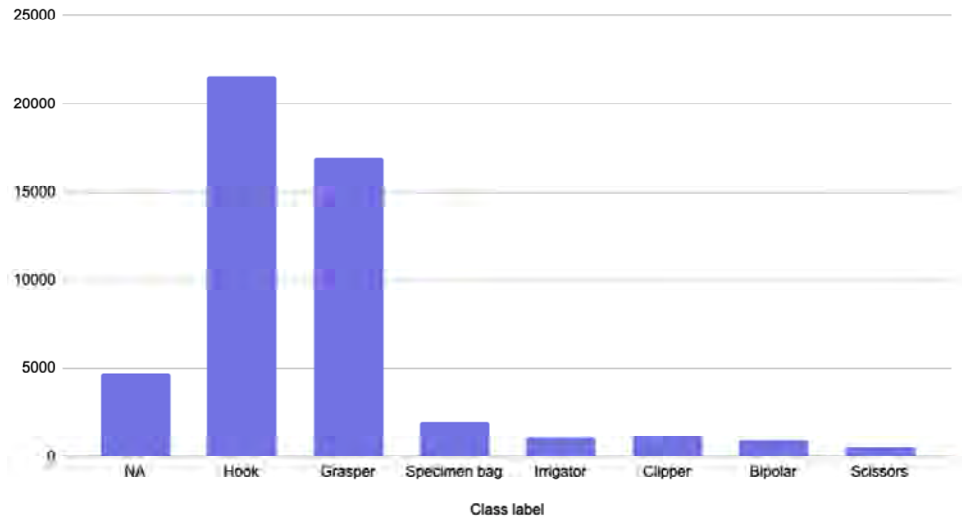
From a surgical point of view, all these correlations make sense as graspers and hooks are used together to lift and move organs, specimen bags are used towards the end of the surgery to retrieve samples and bipolar forceps are used towards the middle of the surgery to cut through fat and tissue. Similarly, clip applicators are used toward the end of the surgery to clip off tubular structures while scissors are used throughout the surgery to cut surface-level tissues. From a model point of view, the strong correlations between certain classes may pose challenges during the training stage. Additional testing data was gathered from raw footage of laparoscopic cholecystectomies performed by local doctors. This data was noisier, i.e., it required more pre-processing and was ‘unseen’ by the model; thus, it simulated a real-world laparoscope feed. A total of five videos were sampled every 30 frames, resulting in a total of 476 frames, which were used to ascertain the generalised instrument detecting and classifying ability of the network across a variety of data.

### 3.2 Data augmentation using SMOTE and under-sampling

As seen in Figure 5, there is a significant difference between the number of instances of hooks and graspers in the dataset as compared to the number of occurrences of other classes. This class imbalance within the data skews the model towards making more

accurate and more frequent predictions of the majority classes, i.e., graspers and hooks while reducing accuracies for minority classes; this resulted in inaccurate models with most of the instruments wrongly being predicted as the graspers.

**Figure 5** Histogram representing class imbalance of instrument classes in the original dataset (see online version for colours)



SMOTE algorithm (Chawla et al., 2002; Liu et al., 2017) was used to synthetically generate additional samples of images which contain minority classes, to better balance the dataset. Unlike simple oversampling algorithms which duplicate minority class samples (i.e., no new information is provided to the model), SMOTE uses the k-nearest neighbours of an image with instruments from the minority classes present in it, to generate the new samples. The minority classes were under-sampled as follows:

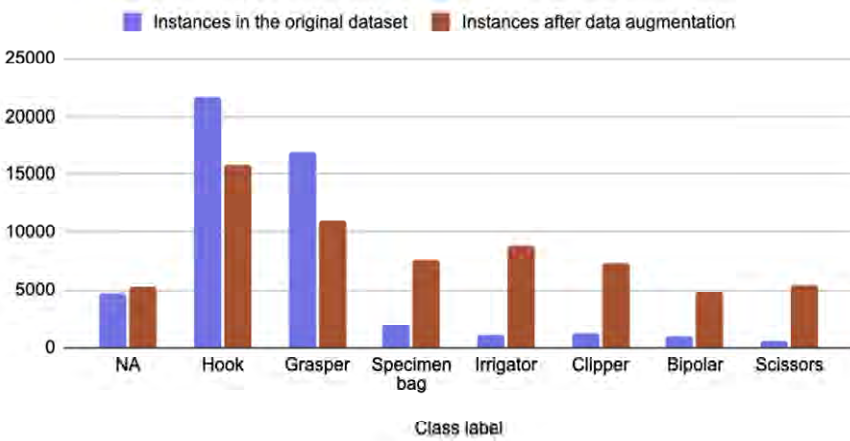
- 1 Images in the same video are separated into sets based on their names.
- 2 Difference hash or Dhash (Weng and Preneel, 2010) signatures are generated for every image in each set.
- 3 A subset of ten images is randomly chosen for every set and hashes of all the other images are matched to hashes of the subset images using brute force.
- 4 Pairs with a degree of similarity greater than 0.8 had one of the images removed from them.
- 5 This process is repeated till all the degrees of similarities are less than 0.8.

The original M2CAI dataset contained 580 frames, which were divided into 480 training frames and 120 testing frames. A total of 1,793 frames were obtained after data augmentation, which were divided into 1,433 training frames and 360 testing frames. Table 3 and Figure 6 show a partial resolution of the class imbalance problem after data augmentation.

**Table 3** Frequency of occurrence of instruments after data augmentation

Class ID	Class label	Instances in the original dataset	Instances after data augmentation
0	NA	4,647	5,230
1	Hook	21,584	15,739
2	Grasper	16,938	11,049
3	Specimen bag	1,987	7,567
4	Irrigator	1,084	8,839
5	Clipper/clip applicator	1,193	7,273
6	Bipolar (forceps/cautery)	962	4,767
7	Scissors	569	5,374

**Figure 6** Comparative histogram of instances of instrument classes in the original dataset (blue) and of instances of instrument classes in the augmented dataset (red) (see online version for colours)



### 3.3 Deep learning models

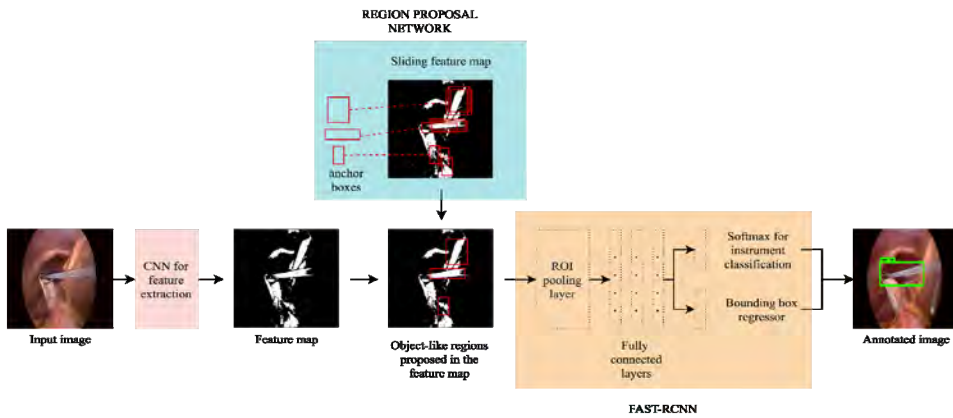
#### 3.3.1 Faster-RCNN

RCNNs take an input image and give a set of bounding boxes at the output, which contain the detected object and the label given to it. Faster-RCNNs (Ren et al., 2015) as seen in Figure 7 are an improved version of the RCNN model, which incorporates a region proposal network (RPN) in place of the selective search algorithm typically used in RCNNs. The steps for generating bounding box and label outputs in a faster-RCNN are:

- 1 CNN extracts features including bounding boxes, class labels, and probabilities for labels/bounding boxes from input images.
- 2 RPN generates region proposals using anchor boxes of various dimensions.
- 3 ROI pooling layer extracts feature vectors for each region proposal.

- 4 Objects in proposed regions are classified into foreground and background instances.
- 5 Classifier and regressor return the probability scores for predicted classes and bounding boxes, respectively.

**Figure 7** Block diagram for faster-RCNN (see online version for colours)



The CNN used in Step 1 is usually a model used for transfer learning which is pre-trained on an enormous dataset of generic images; for this implementation, Inception V3 and B1 Efficient Net architectures were used.

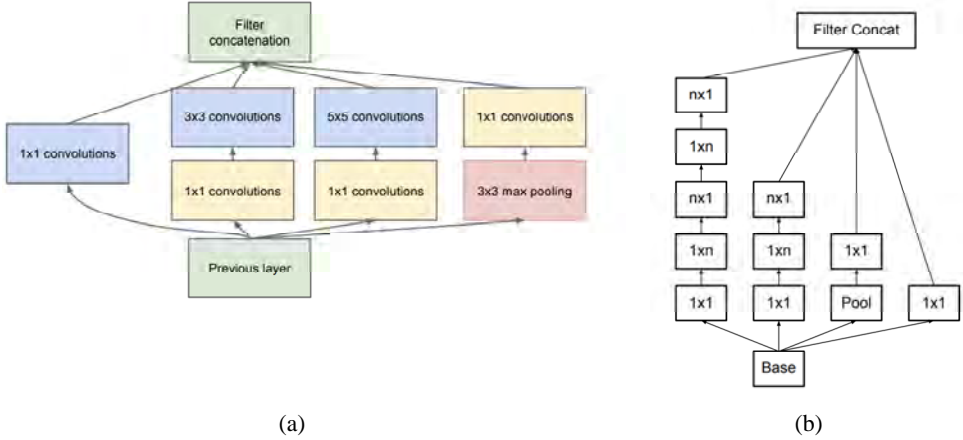
### 3.3.2 The inception architecture

The inception architecture is a series of heavily engineered deep neural networks which aim to improve performance metrics in terms of accuracy and speed through feature engineering, as compared to CNNs which try to improve performance by increasing depth. Inception V1 (Szegedy et al., 2015) was designed to be more ‘wide’ than ‘deep’; it achieved this by performing parallel convolutions using filters of various sizes ( $1 \times 1$ ,  $3 \times 3$  and  $5 \times 5$ ) before pooling the outputs and repeating the process in the next ‘inception’ layer. Figure 8(a) shows an inception block with additional  $1 \times 1$  convolutional filters which help in dimensionality reduction. Inception V1 has nine such blocks between the input and SoftMax output layers. The depth of the network introduced a vanishing gradient problem, which was resolved by adding auxiliary classifiers to the central inception blocks. Inception V2 and V3 (Szegedy et al., 2016) were introduced to solve the problem of information loss which occurred due to drastic changes in input data dimensions because of different filter sizes. They factorised convolution blocks to improve computational efficiency at two levels –  $5 \times 5$  convolutions were factorised into a pair of  $3 \times 3$  convolutions, and all  $N \times N$  convolutions were factorised into  $1 \times n$  and  $n \times 1$  convolutions. Figure 8(b) shows all the changes made in Inception V2, which contributed to its superior performance. Inception V3 changed the location of the auxiliary classifiers from the middle inception blocks to the end inception blocks and added batch normalisation to them. This helped in regularising the model output as accuracies were approaching saturation values towards the output layer. The architecture also introduced another regularisation component called label smoothing, which



prevented over-fitting by preventing the model from becoming too confident about a class prediction.

**Figure 8** (a) Inception V1 block with dimensionality reduction and (b) Inception V2 block with convolution block factorisation (see online version for colours)

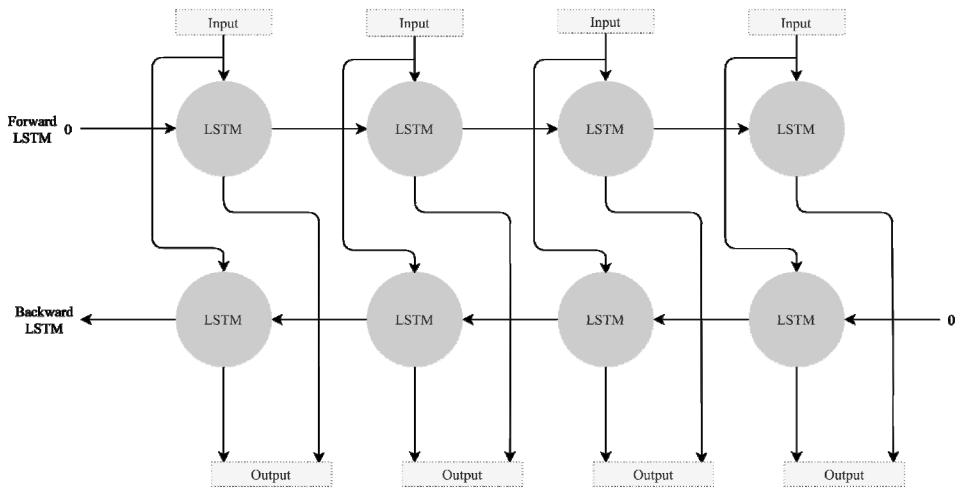


### 3.3.3 The Efficient Net architecture

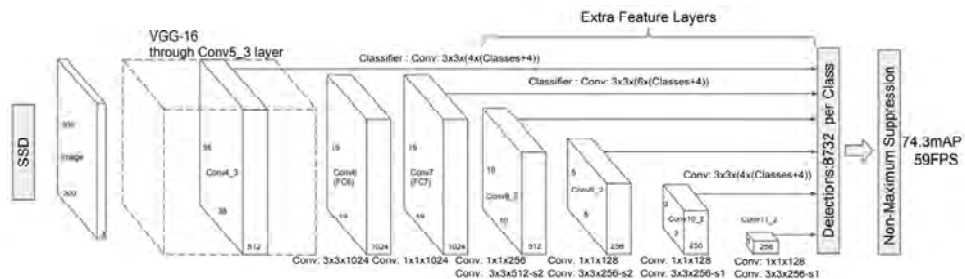
The Efficient Net architecture (Tan and Le, 2019) was introduced as a highly accurate and efficient, yet scalable alternative to traditional small-scale CNNs. It used a technique called compound coefficient to scale up models in a simple but effective manner; instead of randomly scaling up width, depth, or resolution, compound scaling uniformly scaled each dimension with a certain fixed set of scaling coefficients. Using these scaling methods and AutoML, the authors developed seven models of various dimensions (B0 to B7) which surpassed the state-of-the-art accuracy and efficiency of most of the commonly used convolutional neural networks. The architecture also cut down the time wasted by researchers in trial-and-error scaling of their models by developing a uniform set of scaling models.

### 3.3.4 Bidirectional long-short-term memory networks with single-shot detectors

LSTMs networks were introduced by Hochreiter and Schmidhuber (1997) and are a type of RNN that replaces the feedforward connections in neural networks with a back propagation loop. Most RNNs suffer from the problem of short-term memory; this means that if the input is a long sequence, the RNN will not be able to carry information reliably from one time-step to another. LSTMs counter this drawback by using gates, which ascertain whether information from long sequential data is important enough to retain or discard. This allows the model to develop a 'memory' and pass along only important information to future time steps. A bidirectional LSTM is an extension of a traditional LSTM which consists of two LSTMS; one with data flow in the forward direction and one with data flow in the backward direction. Figure 9 illustrates the flow of data in a bidirectional LSTM. The simultaneous forward and backward propagation allows individual LSTM layers to obtain context from both the next and previous image and results in better instrument classification.

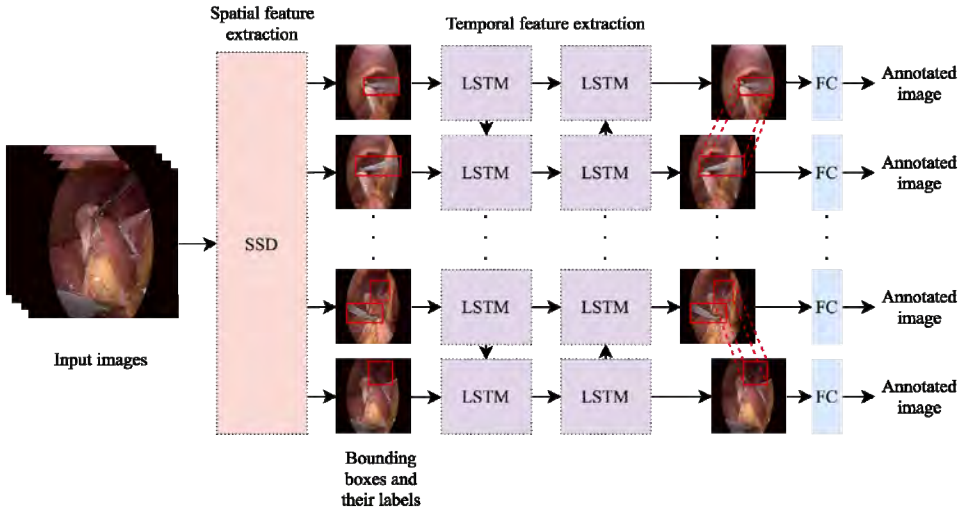
**Figure 9** Data flow in a bidirectional LSTM

This implementation uses bidirectional LSTMs in combination with SSDs, which require a single ‘shot’ or step to detect multiple objects in an image. This makes them faster than two-shot methods such as RCNN which make use of RPNs for region proposal in the first shot and detect objects within each proposed region in the second shot. Figure 10 shows the SSD architecture as proposed by Liu et al. (2016), with a VGG16 (Simonyan and Zisserman, 2014) pre-trained model used as a base and an additional convolutional layer being used for precise feature extraction and classification.

**Figure 10** Single-shot detector architecture

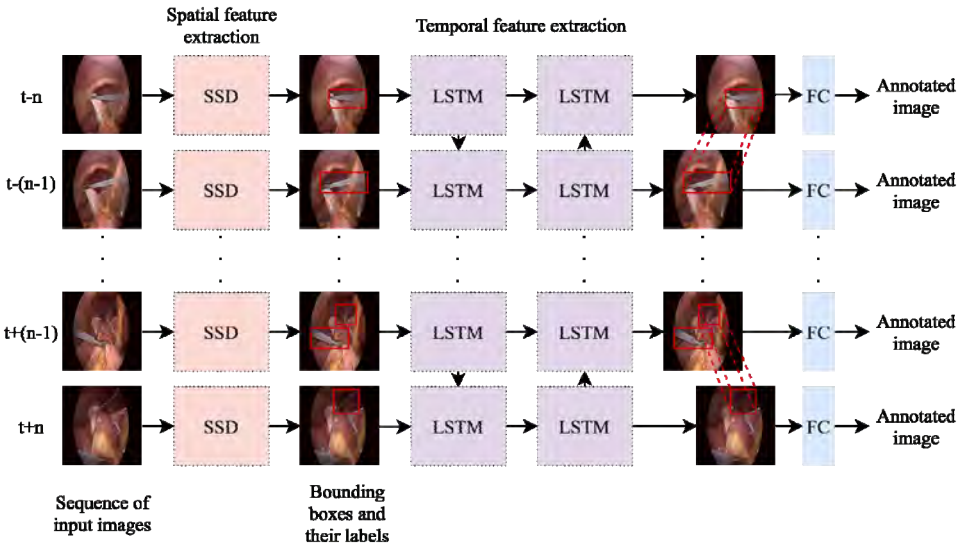
### 3.3.5 Bidirectional LSTM without time-distributed layers

As seen in Figure 11, input images are fed to a VGG16-based SSD for spatial feature extraction. The SSD returns frame-level bounding boxes with a confidence score of more than 0.75 and instrument labels in the form of a feature descriptor vector; these are combined into feature maps which act as the inputs for bidirectional LSTM layers. The bidirectional LSTMs extract temporal features in the form of LSTM states which are constantly updated, and LSTM output as compared to decide the label to be assigned to a bounding box.

**Figure 11** Block diagram of SSD-bidirectional LSTM without time-distributed layers (see online version for colours)

### 3.3.6 Bidirectional LSTM with time-distributed layers

In temporal feature extraction architecture in this model is the same as that of the previous model without time-distributed layers; however, the spatial feature recognition is conducted on a sequence of input images which is separated by a period equal to the sampling rate of the surgery video. As seen in Figure 12, this sequence is fed to individual SSDs, and bounding box predictions are correlated in the bidirectional LSTM layers.

**Figure 12** Block diagram of SSD-bidirectional LSTM with time-distributed layers (see online version for colours)

Time-distributed layers help in correlating the presence and location of instruments between neighbouring frames; this is especially helpful in cases where the tip of an instrument is occluded in a prior frame but visible in the next. Instrument tips are usually the main identifying features of an instrument and comparing adjacent frames helps in reducing the possibility of N/A classifications or incorrect classifications for instruments with similar barrels and different tips.

### 3.4 Evaluation metrics

Model performance was evaluated based on standard metrics (accuracy, precision, recall and F1 score), as well as specifically metrics specific to object detection (Jaccard distance and Hamming distance). This section defines our object detection-specific metrics and describes them in the context of the M2CAI dataset.

#### 3.4.1 Jaccard distance

As defined by equation (1), the Jaccard similarity index (Shi et al., 2014) is a measure of similarity between the members of two sets in terms of intersection over union (IoU). As a ratio of areas, it can take values between 0 and 1.

$$\text{Intersection over union} = \frac{\text{Area of overlap}}{\text{Area of union}} \quad (1)$$

The Jaccard distance is complementary to IoU and can be calculated as shown in equation (2).

$$\text{Jaccard distance} = 1 - \text{Intersection over union} \quad (2)$$

In this implementation, Jaccard distance is calculated as the difference between the actual bounding box coordinates (defined in the metadata files for the M2CAI16 dataset) and predicted bounding box coordinates (obtained from the RCNN). Thus, the lower the Jaccard distance, the closer the predicted coordinates are to the actual coordinates from the annotation files; this makes a lower Jaccard distance a desirable result.

#### 3.4.2 Hamming distance

Hamming distance (Norouzi et al., 2012) measures the difference between a pair of vector elements containing binary data. As described in Section 4.1, the presence or absence of instruments in a frame is annotated as a binary label vector of length = 7. In this implementation, Hamming distance can be calculated by comparing the ground truth and prediction label vectors. E.g., if a frame has one hook and one clip applicator, its ground truth vector is [0, 1, 0, 0, 0, 1, 0] and the predicted vector is [0, 1, 1, 0, 0, 1, 0], we may conclude that a grasper is wrongly predicted to be present. In this case, we can say the Hamming distance is 1, as only one element has taken the incorrect value in the prediction vector. For surgical instrument identification, the values of Hamming distance lie between 0 and 3, as the images in the M2CAI16 dataset have at most three instrument annotations per image. Like Jaccard distance, a low Hamming distance is a desirable metric as it indicates accurate detection of the presence or absence of instruments in a frame.

## 4 Results and discussion

### 4.1 Results

All the implemented models were trained and tested using video frames from the M2CAI16 database, and the best models from the initial testing were further tested on frames from unseen laparoscopic surgery videos. Figure 13 shows selected output images from the augmented M2CAI16 dataset, with overlaid bounding boxes and instrument annotations.

**Figure 13** Output images showing surgical instruments such as graspers, scissors and specimen bag (see online version for colours)

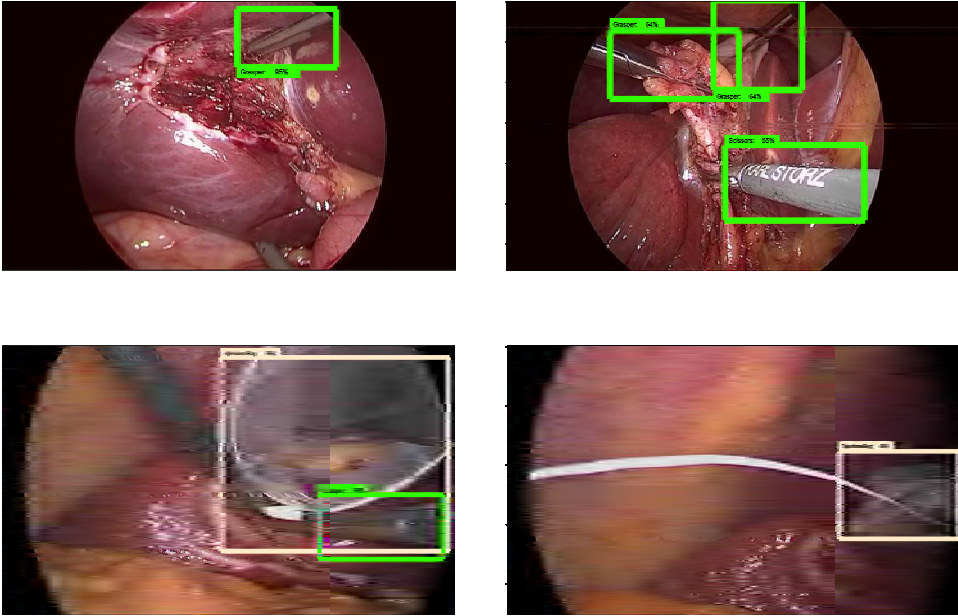


Table 4 summarises results based on the Jaccard and Hamming distance metrics that are chosen specifically for the object detection problem.

Table 5 summarises the results obtained from selected models, based on standard model metrics such as on accuracy, precision, recall and F1 score. It also includes the results for models evaluated on unseen video data.

### 4.2 Discussion

Laparoscopic surgeons make use of instruments such as graspers and hooks more frequently over the course of a surgery due to their versatility, making them the most frequently occurring classes in the MICCAI16 dataset, i.e., the majority classes. On the other hand, instruments such as clip applicators, scissors, and bipolar forceps have very specific functions in the surgery and they do not occur as frequently in the dataset, making them the minority classes. Several instruments are often used simultaneously, causing a single class to occur multiple times in the same frame and further increasing its

counts in the dataset. This imbalance is reflected in the results for all the models trained on data that did not undergo data augmentation; instruments in the majority class had a higher classification accuracy while instruments in the minority class suffered from poor classification accuracy and the issue was most evident when the models were tested on unseen frames from a source outside the M2CAI16 dataset. An exception to this was the specimen bag class; despite having relatively few samples, it had very good results across all models which could be attributed to specimen bags looking the most different from all other instruments in the dataset (which could be mistaken for each other by the model due to them having similar handles and the identifying features being their tips).

The majority classes were robustly identified, as seen by the higher recall and precision shown by models while classifying them, whereas the minority classes were either identified as graspers or not identified at all, as seen by their significantly lower recall and precision. Results for un-augmented data show significant differences between the ground truth and predicted bounding box values resulting in high Jaccard distances, pointing towards inaccurate localisation of the instrument. They also had high Hamming distances due to incorrect labelling of the contents of the bounding boxes. Balancing the dataset resulted in a marked improvement in performance across all models, as seen by the increase in both accuracy and F1 scores along with a drop in the Jaccard and Hamming distances. The best results achieved on the M2CAI16 dataset were an average accuracy of 80.20% and an average F1 score of 0.7176 and were derived from the bidirectional LSTM model with time-distributed layers; for the unseen dataset, the same model had a peak performance of 63.49% average accuracy and an average F1 score of 0.5221. The same model also had minimum Jaccard and Hamming distances and overall, both the bidirectional LSTMs performed better than both the faster-RCNNs due to the additional validation of instrument labels by temporal features, with the bidirectional LSTM model with time-distributed layers being the best-performing model across all metrics for the M2CAI16 dataset. All the models performed poorly on frames from the unseen surgical videos, an expected result that was caused by the significant visual differences between the datasets despite initial image pre-processing efforts. The bidirectional LSTM model with time-distributed layers performed considerably better than the others despite this limitation and achieved an average accuracy and F1 score of 63.49% and 0.5221, respectively. The poor F1 score may be explained by the low recall, which in turn was due to the model being unable to classify most of the unseen instruments. Table 6 compares the performance of selected prior works that have been evaluated on the M2CAI16 dataset and our best-performing model's (bidirectional LSTM with time-distributed layers) results with the data-augmented M2CAI16. It is observed that our model performs better than all the transfer learning models trained by Zia et al. (2016) and equally well as the modified recurrent CNN proposed by Namazi et al. (2022). Although our architecture parallels that proposed by Mishra et al. (2017) with key differences in our use of a bidirectional LSTM and time-distributed layers, their model's superior accuracy can be attributed to the author's superior imbalance mitigation. Jin et al. (2016) were the only other group to use Jaccard distance to evaluate object detection; our score was slightly higher than the score reported by them.

**Table 4**      Comparison of Jaccard and Hamming distance results

Data	Model	Jaccard distance			Hamming distance		
		Average	Max.	Min.	Average	Max.	Min.
M2CA116 (without data augmentation)	Faster-RCNN (InceptionV3)	0.3810	0.6745	0.2902	2	2	1
	Faster-RCNN (efficient net)	0.2893	0.6837	0.2415	2	3	1
	Bidirectional LSTM without time-distributed layers	0.3109	0.4680	0.1722	2	2	1
	Bidirectional LSTM with time-distributed layers	0.2470	0.4435	0.1208	1	2	0
M2CA116 (with data augmentation)	Faster-RCNN (InceptionV3)	0.1942	0.2785	0.0941	2	2	1
	Faster-RCNN (efficient net)	0.2891	0.4448	0.1265	1	2	0
	Bidirectional LSTM without time-distributed layers	0.3034	0.4612	0.0604	1	2	0
	Bidirectional LSTM with time-distributed layers	0.2130	0.3148	0.0859	0	2	0

**Table 5** Comparison of results based on standard metrics

<i>Data</i>	<i>Model</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1 score</i>
M2CAI16 tool dataset (without data augmentation)	Faster-RCNN (InceptionV3)	71.12%	0.6298	0.6515	0.6404
	Faster-RCNN (Efficient Net)	63.47%	0.6885	0.4933	0.5163
	Bidirectional LSTM without time-distributed layers	65.53%	0.5573	0.6127	0.5304
	Bidirectional LSTM with time-distributed layers	79.35%	0.7185	0.6042	0.6472
M2CAI16 tool dataset (with data augmentation)	Faster-RCNN (InceptionV3)	76.81%	0.6982	0.6850	0.6915
	Bidirectional LSTM with time-distributed layers	80.20%	0.7134	0.7629	0.7176
Frames from unseen surgery videos	Faster-RCNN (InceptionV3)	57.14%	0.5800	0.4722	0.4658
	Faster-RCNN (Efficient Net)	53.15%	0.6133	0.4421	0.4501
	Bidirectional LSTM without time-distributed layers	56.30%	0.4556	0.4816	0.3891
	Bidirectional LSTM with time-distributed layers	63.49%	0.5808	0.5030	0.5221

**Table 6** Comparison of our best performing model with selected prior work

<i>Method</i>	<i>Jaccard distance</i>	<i>Accuracy</i>
Zia et al. (2016)	-	AlexNet: 63.78%/VGG: 69.75%/inception: 76.6%
Jin et al. (2016)	78.2	-
Mishra et al. (2017)	-	88.75%
Namazi et al. (2022)	-	Online: 80.95%/offline: 81.84%
Proposed method	78.7	80.20%

## 5 Conclusions

Surgery videos are an important part of surgical training but may not always be easy to interpret for novice trainees and general surgeons who are shifting to a super-specialisation (in this case, laparoscopy). The dynamic nature of anatomy, combined with a surgeon's limited FoV through a laparoscope may further complicate the learning process. The proposed method aims to solve the low-level problem of classification and annotation of instruments using a novel combination of bidirectional LSTM and time-distributed layers, laying the foundation for various applications of surgical video analysis. The results obtained on the M2CAI dataset are at par with previous work of the same nature and may be further improved through future work as follows: using generative adversarial networks to generate new training images for under-represented classes during data augmentation, evaluating the models using other metrics (average precision, mAP score, etc.), combining datasets for access to varied image data and using metaheuristic optimisation algorithms to improve the accuracy metric of our LSTMs (Rashid et al., 2018; Mahmoodzadeh et al., 2022). Another future direction would be incorporating predictions made into real-time laparoscopic surgical training



devices, that may require un-supervised or semi-supervised learning to provide rapid and accurate predictions in complex and dynamic environments faced by surgeons.

## Acknowledgements

The authors would like to thank Dr. Deepraj Bhandarkar (Department of Minimal Access Surgery, P.D. Hinduja National Hospital & Medical Research Centre, Mumbai) for his valuable insight on the challenges faced during laparoscopic surgeries and the importance of surgical videos as a pedagogical tool for trainee surgeons.

## References

- Allan, M., Ourselin, S., Thompson, S., Hawkes, D.J., Kelly, J. and Stoyanov, D. (2012) 'Toward detection and localization of instruments in minimally invasive surgery', *IEEE Transactions on Biomedical Engineering*, Vol. 60, No. 4, pp.1050–1058.
- Alshirbaji, T.A., Jalal, N.A. and Möller, K. (2018) 'Surgical tool classification in laparoscopic videos using convolutional neural network', *Current Directions in Biomedical Engineering*, Vol. 4, No. 1, pp.407–410.
- Azqueta-Gavaldon, I., Fröhlich, F., Strobl, K. and Triebel, R. (2020) *Segmentation of Surgical Instruments for Minimally-Invasive Robot-Assisted Procedures Using Generative Deep Neural Networks*, arXiv 2020 preprint, arXiv: 2006.03486v1.
- Badilla-Solórzano, J., Spindeldreier, S., Ihler, S., Gellrich, N-C. and Spalthoff, S. (2022) 'Deep-learning-based instrument detection for intra-operative robotic assistance', *International Journal of Computer Assisted Radiology and Surgery*, Vol. 17, No. 9, pp.1685–1695.
- Baghdadi, A., Hussein, A.A., Ahmed, Y., Cavuoto, L.A. and Guru, K.A. (2019) 'A computer vision technique for automated assessment of surgical performance using surgeons' console-feed videos', *International Journal of Computer Assisted Radiology and Surgery*, Vol. 14, No. 4, pp.697–707.
- Ballantyne, G.H. (2002) 'The pitfalls of laparoscopic surgery: challenges for robotics and telerobotic surgery', *Surgical Laparoscopy Endoscopy & Percutaneous Techniques*, Vol. 12, No. 1, pp.1–5.
- Bhagya, C. and Shyna, A. (2019) 'An overview of deep learning-based object detection techniques', in *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, IEEE, April, pp.1–6.
- Bouget, D., Allan, M., Stoyanov, D. and Jannin, P. (2017) 'Vision-based and marker-less surgical tool detection and tracking: a review of the literature', *Medical Image Analysis*, Vol. 35, pp.633–654.
- Cai, T. and Zhao, Z. (2020) 'Convolutional neural network-based surgical instrument detection', *Technology and Health Care*, Vol. 28, No. S1, pp.81–88.
- Casals, A., Amat, J. and Laporte, E. (1996) 'Automatic guidance of an assistant robot in laparoscopic surgery', in *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, April, Vol. 1, pp.895–900.
- Celentano, V., Smart, N., Cahill, R.A., McGrath, J.S., Gupta, S., Griffith, J.P., Acheson, A.G., Cecil, T.D. and Coleman, M.G. (2019) 'Use of laparoscopic videos amongst surgical trainees in the United Kingdom', *The Surgeon*, Vol. 17, No. 6, pp.334–339.
- Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) 'SMOTE: synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, Vol. 16, pp.321–357.

- Cheplygina, V., de Bruijne, M. and Pluim, J.P.W. (2018) *Not-so-supervised: A Survey of Semi-supervised, Multi-instance, and Transfer Learning in Medical Image Analysis*, arXiv preprint arXiv: 1804.06353.
- Cuschieri, A. (2005) 'Laparoscopic surgery: current status, issues and future developments', *The Surgeon*, Vol. 3, No. 3, pp.125–138.
- de'Angelis, N., Gavrilidis, P., Martínez-Pérez, A., Genova, P., Notarnicola, M., Reitano, E., Petrucciani, N., Abdalla, S., Memeo, R., Brunetti, F. and Carra, M.C. (2019) 'Educational value of surgical videos on YouTube: quality assessment of laparoscopic appendectomy videos by senior surgeons vs. novice trainees', *World Journal of Emergency Surgery*, Vol. 14, No. 1, pp.1–11.
- Dobson, S.J. and Hopkins, H.H. (1989) 'A new rod-lens relay system offering improved image quality', *Journal of Physics E: Scientific Instruments*, Vol. 22, No. 7, p.450.
- Du, X., Allan, M., Dore, A., Ourselin, S., Hawkes, D., Kelly, J.D. and Stoyanov, D. (2016) 'Combined 2D and 3D tracking of surgical instruments for minimally invasive and robotic-assisted surgery', *International Journal of Computer Assisted Radiology and Surgery*, Vol. 11, No. 6, pp.1109–1119.
- Ellis, H. (2007) 'The Hopkins rod-lens system', *The Journal of Perioperative Practice*, Vol. 17, No. 6, p.272.
- Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J. and Socher, R. (2021) 'Deep learning-enabled medical computer vision', *NPJ Digital Medicine*, Vol. 4, No. 1, pp.1–9.
- Fathabadi, F.R., Grantner, J.L., Shebrain, S.A. and Abdel-Qader, I. (2021) 'Multi-class detection of laparoscopic instruments for the intelligent box-trainer system using faster R-CNN architecture', *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*.
- Fitzpatrick, C.M., Kolesari, G.L. and Brasel, K.J. (2001) 'Teaching anatomy with surgeons' tools: use of the laparoscope in clinical anatomy', *Clinical Anatomy: The Official Journal of the American Association of Clinical Anatomists and the British Association of Clinical Anatomists*, Vol. 14, No. 5, pp.349–353.
- García-Peraza-Herrera, L.C., Li, W., Gruijthuisen, C., Devreker, A., Attilakos, G., Deprest, J., Poorten, E.V., Stoyanov, D., Vercauteren, T. and Ourselin, S. (2016) 'Real-time segmentation of non-rigid surgical tools based on deep learning and tracking', in *International Workshop on Computer-Assisted and Robotic Endoscopy*, Springer, Cham, October, pp.84–95.
- Grantner, J.L., Kurdi, A.H., AlGailani, M., Abdel-Qader, I., Sawyer, R.G. and Shebrain, S. (2019) 'Multi-thread implementation of tool tip tracking for laparoscopic surgical box-trainer intelligent performance assessment system', *2019 IEEE 23rd International Conference on Intelligent Engineering Systems (INES)*.
- Greenish, S., Hayward, V., Chial, V., Okamura, A. and Steffen, T. (2002) 'Measurement, analysis, and display of haptic signals during surgical cutting', *Presence: Teleoperators & Virtual Environments*, Vol. 11, No. 6, pp.626–651.
- Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural Computation*, Vol. 9, No. 8, pp.1735–1780.
- Hossain, B., Nishio, S., Takafumio, H. and Kobashi, S. (2020) 'A deep learning approach for surgical instruments detection in orthopaedic surgery videos using transfer learning', *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*.
- Jaafari, J., Douzi, S., Douzi, K. and Hssina, B. (2022) 'The impact of ensemble learning on surgical tools classification during laparoscopic cholecystectomy', *Journal of Big Data*, Vol. 9, No. 1, Article No. 49.
- Jha, D., Ali, S., Tomar, N.K., Riegler, M.A., Johansen, D., Johansen, H.D. and Halvorsen, P. (2021) 'Exploring deep learning methods for real-time surgical instrument segmentation in laparoscopy', *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*.

- Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z. and Qu, R. (2019) 'A survey of deep learning-based object detection', *IEEE Access*, Vol. 7, pp.128837–128868.
- Jin, A., Yeung, S., Jopling, J., Krause, J., Azagury, D., Milstein, A. and Fei-Fei, L. (2018) 'Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks', in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, March, pp.691–699.
- Jin, Y., Dou, Q., Chen, H., Yu, L. and Heng, P.A. (2016) 'EndoRCN: recurrent convolutional networks for recognition of surgical workflow in cholecystectomy procedure video', *IEEE Trans on Medical Imaging*, Vol. 35.
- Jönsson, B. and Zethraeus, N. (2000) 'Costs and benefits of laparoscopic surgery – a review of the literature', *European Journal of Surgery*, Vol. 166, No. S12, pp.48–56.
- Kanakatte, A., Ramaswamy, A., Gubbi, J., Ghose, A. and 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 & Biology Society (EMBC)*, IEEE, July, pp.1658–1661.
- Kitaguchi, D., Takeshita, N., Matsuzaki, H., Takano, H., Owada, Y., Enomoto, T., Oda, T., Miura, H., Yamanashi, T., Watanabe, M. and Sato, D. (2020) 'Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach', *Surgical Endoscopy*, Vol. 34, No. 11, pp.4924–4931.
- Kletz, S., Schoeffmann, K., Benois-Pineau, J. and Husslein, H. (2019) 'Identifying surgical instruments in laparoscopy using deep learning instance segmentation', *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*.
- Kranzfelder, M., Schneider, A., Fiolka, A., Schwan, E., Gillen, S., Wilhelm, D., Schirren, R., Reiser, S., Jensen, B. and Feussner, H. (2013) 'Real-time instrument detection in minimally invasive surgery using radiofrequency identification technology', *Journal of Surgical Research*, Vol. 185, No. 2, pp.704–710.
- Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) 'ImageNet classification with deep convolutional neural networks', *Communications of the ACM*, Vol. 60, No. 6, pp.84–90.
- Krupa, A., Gangloff, J., Doignon, C., De Mathelin, M.F., Morel, G., Leroy, J., Soler, L. and Marescaux, J. (2003) 'Autonomous 3-D positioning of surgical instruments in robotized laparoscopic surgery using visual servoing', *IEEE Transactions on Robotics and Automation*, Vol. 19, No. 5, pp.842–853.
- Lam, K., Lo, F.P-W., An, Y., Darzi, A., Kinross, J.M., Purkayastha, S. and Lo, B. (2022) 'Deep learning for instrument detection and assessment of operative skill in surgical videos', *IEEE Transactions on Medical Robotics and Bionics*, Vol. 4, No. 4, pp.1068–1071.
- Lee, D., Yu, H.W., Kwon, H., Kong, H-J., Lee, K.E. and Kim, H.C. (2020) 'Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations', *Journal of Clinical Medicine*, Vol. 9, No. 6, p.1964.
- Lee, J-D., Chien, J-C., Hsu, Y-T. and Wu, C-T. (2021) 'Automatic surgical instrument recognition – a case of comparison study between the faster R-CNN, mask R-CNN, and single-shot multi-box detectors', *Applied Sciences*, Vol. 11, No. 17, p.8097.
- Levy, B. and Mobasher, M. (2017) 'Principles of safe laparoscopic surgery', *Surgery (Oxford)*, Vol. 35, No. 4, pp.216–219.
- Li, Z., Huang, Y., Cai, M. and Sato, Y. (2019) 'Manipulation-skill assessment from videos with spatial attention network', in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.
- Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B. and Sánchez, C.I. (2017) 'A survey on deep learning in medical image analysis', *Medical Image Analysis*, Vol. 42, pp.60–88.

- Liu, R., Hall, L.O., Bowyer, K.W., Goldgof, D.B., Gatenby, R. and Ahmed, K.B. (2017) 'Synthetic minority image over-sampling technique: how to improve AUC for glioblastoma patient survival prediction', in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, October, pp.1357–1362.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y. and Berg, A.C. (2016) 'SSD: single shot multibox detector', in *European Conference on Computer Vision*, Springer, Cham, October, pp.21–37.
- Longmore, S.K., Naik, G. and Gargiulo, G.D. (2020) 'Laparoscopic robotic surgery: current perspective and future directions', *Robotics*, Vol. 9, No. 2, p.42.
- Loukas, C. (2018) 'Video content analysis of surgical procedures', *Surgical Endoscopy*, Vol. 32, No. 2, pp.553–568.
- Mahmoodzadeh, A., Nejati, H.R., Mohammadi, M., Hashim Ibrahim, H., Rashidi, S. and Ahmed Rashid, T. (2022) 'Forecasting tunnel boring machine penetration rate using LSTM deep neural network optimized by grey wolf optimization algorithm', *Expert Systems with Applications*, Vol. 209, Article No. 118303.
- 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*, Vol. 8, Article No. 101.
- Markarian, N., Kugener, G., Pangal, D.J., Unadkat, V., Sinha, A., Zhu, Y., Roshannai, A., Chan, J., Hung, A.J., Wrobel, B.B., Anandkumar, A., Zada, G. and Donoho, D.A. (2022) 'Validation of machine learning-based automated surgical instrument annotation using publicly available intraoperative video', *Operative Neurosurgery*, Vol. 23, No. 3, pp.235–240.
- Mishra, K., Sathish, R. and Sheet, D. (2017) 'Learning latent temporal connectionism of deep residual visual abstractions for identifying surgical tools in laparoscopy procedures', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.58–65.
- Miyawaki, F., Tsunoi, T., Namiki, H., Yaginuma, T., Yoshimitsu, K., Hashimoto, D. and Fukui, Y. (2009) 'Development of automatic acquisition system of surgical-instrument information in endoscopic and laparoscopic surgery', in *2009 4th IEEE Conference on Industrial Electronics and Applications*, IEEE, May, pp.3058–3063.
- Mohaidat, M., Grantner, J.L., Shebrain, S.A. and Abdel-Qader, I. (2022) 'Instrument detection for the intracorporeal suturing task in the laparoscopic box trainer using single-stage object detectors', *2022 IEEE International Conference on Electro Information Technology (EIT)*.
- Mota, P., Carvalho, N., Carvalho-Dias, E., Costa, M.J., Correia-Pinto, J. and Lima, E. (2018) 'Video-based surgical learning: improving trainee education and preparation for surgery', *Journal of Surgical Education*, Vol. 75, No. 3, pp.828–835.
- Myo, N.N., Boonkong, A., Hormdee, D., Sonsilphong, S., Sonsilphong, A. and Khampitak, K. (2022) 'Laparoscope manipulating robot (LMR) navigation using deep learning-based surgical instruments detection', *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.
- Nadhifatul Aini, F.A., Zalnika Purwalaksana, A. and Manalu, I.P. (2019) 'Object detection of surgical instruments for assistant robot surgeon using KNN', *2019 International Conference on Advanced Mechatronics, Intelligent Manufacturing and Industrial Automation (ICAMIMIA)*.
- Nakano, A. and Nagamune, K. (2022) 'A development of robotic scrub nurse system – detection for surgical instruments using faster region-based convolutional neural network', *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 26, No. 1, pp.74–82.
- Nakawala, H., Bianchi, R., Pescatori, L.E., De Cobelli, O., Ferrigno, G. and De Momi, E. (2019) '“Deep-Onto” network for surgical workflow and context recognition', *International Journal of Computer Assisted Radiology and Surgery*, Vol. 14, No. 4, pp.685–696.
- Namaz, B., Sankaranarayanan, G. and Devarajan, V. (2022) 'A contextual detector of surgical tools in laparoscopic videos using deep learning', *Surgical Endoscopy*, Vol. 36, No. 1, pp.679–688.

- Norouzi, M., Fleet, D.J. and Salakhutdinov, R.R. (2012) 'Hamming distance metric learning', *Advances in Neural Information Processing Systems*, Vol. 25.
- Peter, J.D., Fernandes, S.L., Thomaz, C.E. and Viriri, S. (Eds.) (2019) 'Computer aided intervention and diagnostics in clinical and medical images', *Lecture Notes in Computational Vision and Biomechanics*, Vol. 31.
- Peters, B.S., Armijo, P.R., Krause, C., Choudhury, S.A. and Oleynikov, D. (2018) 'Review of emerging surgical robotic technology', *Surgical Endoscopy*, Vol. 32, No. 4, pp.1636–1655.
- Prellberg, J. and Kramer, O. (2018) 'Multi-label classification of surgical tools with convolutional neural networks', in *2018 International Joint Conference on Neural Networks (IJCNN)*, IEEE, July, pp.1–8.
- Rashid, T.A., Fattah, P. and Awla, D.K. (2018) 'Using accuracy measure for improving the training of LSTM with metaheuristic algorithms', *Procedia Computer Science*, Vol. 140, pp.324–333.
- Ren, S., He, K., Girshick, R. and Sun, J. (2015) 'Faster R-CNN: towards real-time object detection with region proposal networks', in *Advances in Neural Information Processing Systems*, Vol. 28.
- Roß, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M. et al. (2021) 'Comparative validation of multi-instance instrument segmentation in endoscopy: results of the robust-MIS 2019 challenge', *Medical Image Analysis*, Vol. 70, Article No. 101920.
- Sahu, M., Mukhopadhyay, A. and Zachow, S. (2021) 'Simulation-to-real domain adaptation with teacher-student learning for endoscopic instrument segmentation', *International Journal of Computer Assisted Radiology and Surgery*, Vol. 16, No. 5, pp.849–859.
- Sahu, M., Mukhopadhyay, A., Szengel, A. and Zachow, S. (2016) *Tool and Phase Recognition Using Contextual CNN Features*, arXiv preprint arXiv: 1610.08854.
- Shi, R., Ngan, K.N. and Li, S. (2014) 'Jaccard index compensation for object segmentation evaluation', in *2014 IEEE International Conference on Image Processing (ICIP)*, IEEE, October, pp.4457–4461.
- Simonyan, K. and Zisserman, A. (2014) *Very Deep Convolutional Networks for Large-scale Image Recognition*, arXiv preprint arXiv: 1409.1556.
- Speidel, S., Kuhn, E., Bodenstedt, S., Röhl, S., Kenngott, H., Müller-Stich, B. and Dillmann, R. (2014) 'Visual tracking of Da Vinci instruments for laparoscopic surgery', in *Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling*, SPIE, March, Vol. 9036, pp.47–52.
- Stiff, G., Rhodes, M., Kelly, A., Telford, K., Armstrong, C.P. and Rees, B.I. (1994) 'Long-term pain: less common after laparoscopic than open cholecystectomy', *British Journal of Surgery*, Vol. 81, No. 9, pp.1368–1370.
- Su, B., Zhang, Q., Gong, Y., Xiu, W., Gao, Y., Xu, L. et al. (2023) 'Deep learning-based classification and segmentation for scalpels', *International Journal of Computer Assisted Radiology and Surgery*, Vol. 18, No. 5, pp.855–864, DOI: 10.1007/s11548-022-02825-7.
- Sun, Y., Pan, B. and Fu, Y. (2021) 'Lightweight deep neural network for real-time instrument semantic segmentation in robot assisted minimally invasive surgery', *IEEE Robotics and Automation Letters*, Vol. 6, No. 2, pp.3870–3877.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015) 'Going deeper with convolutions', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1–9.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) 'Rethinking the inception architecture for computer vision', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2818–2826.
- Tan, M. and Le, Q. (2019) 'Efficientnet: rethinking model scaling for convolutional neural networks', in *International Conference on Machine Learning*, PMLR, May, pp.6105–6114.

- Tanagho, Y.S., Andriole, G.L., Paradis, A.G., Madison, K.M., Sandhu, G.S., Varela, J.E. and Benway, B.M. (2012) '2D versus 3D visualization: impact on laparoscopic proficiency using the fundamentals of laparoscopic surgery skill set', *Journal of Laparoendoscopic & Advanced Surgical Techniques*, Vol. 22, No. 9, pp.865–870.
- Tatar, F., Mollinger, J. and Bossche, A. (2003) 'Ultrasound system for measuring position and orientation of laparoscopic surgery tools', in *2003 IEEE Sensors*, IEEE, October, Vol. 2, pp.987–990.
- Taylor, G.W. and Jayne, D.G. (2007) 'Robotic applications in abdominal surgery: their limitations and future developments', *The International Journal of Medical Robotics and Computer Assisted Surgery*, Vol. 3, No. 1, pp.3–9.
- Tonet, O., Thoranaghatte, R.U., Megali, G. and Dario, P. (2007) 'Tracking endoscopic instruments without a localizer: a shape-analysis-based approach', *Computer Aided Surgery*, Vol. 12, No. 1, pp.35–42.
- Trilling, B., Mancini, A., Fiard, G., Barraud, P.A., Decrouez, M., Vijayan, S., Tummers, M., Faucheron, J.L., Silvent, S., Schwartz, C. and Voros, S. (2021) 'Improving vision for surgeons during laparoscopy: the enhanced laparoscopic vision system (ELViS)', *Surgical Endoscopy*, Vol. 35, No. 5, pp.2403–2415.
- Twinanda, A.P. (2017) *Vision-based Approaches for Surgical Activity Recognition Using Laparoscopic and RBGD Videos*, Doctoral dissertation, Strasbourg.
- Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M. and Padoy, N. (2016) 'EndoNet: a deep architecture for recognition tasks on laparoscopic videos', *IEEE Transactions on Medical Imaging*, Vol. 36, No. 1, pp.86–97.
- Varela, J.E., Wilson, S.E. and Nguyen, N.T. (2010) 'Laparoscopic surgery significantly reduces surgical-site infections compared with open surgery', *Surgical Endoscopy*, Vol. 24, No. 2, pp.270–276.
- Velanovich, V. (2000) 'Laparoscopic vs open surgery', *Surgical Endoscopy*, Vol. 14, No. 1, pp.16–21.
- Wang, A., Islam, M., Xu, M. and Ren, H. (2022) 'Rethinking surgical instrument segmentation: a background image can be all you need', *Lecture Notes in Computer Science*, pp.355–364.
- Wang, S., Raju, A. and Huang, J. (2017) 'Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos', in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, IEEE, April, pp.620–623.
- Wang, Y., Sun, Q., Sun, G., Gu, L. and Liu, Z. (2021) 'Object detection of surgical instruments based on yolov4', *2021 6th IEEE International Conference on Advanced Robotics and Mechatronics (ICARM)*.
- Weng, L. and Preneel, B. (2010) 'From image hashing to video hashing', in *International Conference on Multimedia Modeling*, Springer, Berlin, Heidelberg, January, pp.662–668.
- 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*, Vol. 230, No. 5, pp.725–732.
- Yeola, M., Gode, D. and Bora, A. (2017) 'Evolution of laparoscopy through the ages', *International Journal of Recent Surgical and Medical Sciences*, Vol. 3, No. 1, pp.40–47.
- Yoon, J., Lee, J., Park, S., Hyung, W.J. and Choi, M-K. (2020). 'Semi-supervised learning for instrument detection with a class imbalanced dataset', *Interpretable and Annotation-Efficient Learning for Medical Image Computing: Proceedings of Second International Workshop, MIL3ID 2020*, pp.266–276.
- Zia, A., Castro, D. and Essa, I. (2016) 'Fine-tuning deep architectures for surgical tool detection', in *Workshop and Challenges on Modeling and Monitoring of Computer Assisted Intervention (M2CAI)*, Technical report, Athens, Greece, October.