Hierarchical, Informed and Robust Machine Learning for Surgical Tool Management

A thesis
submitted in fulfillment
of the requirements for the degree
of

Doctor of Philosophy

in the

Department of Computer Science
The University of Waikato

by

MARK WILLIAM RODRIGUES

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2023

Abstract

Hierarchical, Informed and Robust Machine Learning for Surgical Tool Management

by MARK WILLIAM RODRIGUES

This thesis focuses on the development of a computer vision and deep learning based system for the intelligent management of surgical tools. The work accomplished included the development of a new dataset, creation of state of the art techniques to cope with volume, variety and vision problems, and designing or adapting algorithms to address specific surgical tool recognition issues. The system was trained to cope with a wide variety of tools, with very subtle differences in shapes, and was designed to work with high volumes, as well as varying illuminations and backgrounds. Methodology that was adopted in this thesis included the creation of a surgical tool image dataset and development of a surgical tool attribute matrix or knowledge-base. This was significant because there are no large scale publicly available surgical tool datasets, nor are there established annotations or datasets of textual descriptions of surgical tools that can be used for machine learning. The work resulted in the development of a new hierarchical architecture for multi-level predictions at surgical speciality, pack, set and tool level. Additional work evaluated the use of synthetic data to improve robustness of the CNN, and the infusion of knowledge to improve predictive performance.

Acknowledgements

The most important aspect of any PhD is the Supervisory Team, and I was blessed and fortunate to get a fantastic team of Supervisors. This Team unfailingly provided me with incredible support, direction and insights. Dr Tony Smith was incredibly helpful in his comments and insights into what could be done to improve the work; I am eternally grateful for his assistance and for his strong and unwavering support. Dr Michael Mayo is without doubt the best Supervisor/Manager I have ever worked with; he always responded promptly to submissions and queries with insightful comments and definitive directions. His wealth of experience, generosity in providing time and sharing knowledge, and willingness to read up new material that could contribute to improving the research was a revelation and a blessing. Dr Panos Patros provided structure, organisation and attention to detail in all my work. He ensured that my research always maintained high standards, and invariably provided cheerful and happy inputs to my work. He provided invaluable professional direction and career support, and I could not have hoped for a better mentor and guide.

I am grateful for Dale Fletcher for putting me onto the Computer Science path, and for the constant encouragement and "stress-relief" hitting on the tennis court. The academic staff who managed and delivered the PGCertInfoTech program – John Thompson, Geoff Holmes, David Bainbridge, Alvin Yeo – thank you for a magnificent program. I sincerely wish that there were many more such programs in academia – relevant, rich in content, practical and incredibly useful.

Thank you, Sanaya. You have always been supportive and loving, and ensured that I could successfully complete this (second) thesis. Now I have more time to devote to do the things that you love.

Contents

List of Figures

List of Tables

List of Abbreviations

AE	AutoEncoder
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representation from Transformers
BILSTM	Bidirectional Long Short-Term Memory
BRNN	Bidirectional Recurrent Neural Network
CNN	Convolutional Neural Network
ConvNet	Convolutional Neural Network
CRNN	Convolutional Recurrent Neural Network
DL	Deep Learning
DNN	Deep Neural Network
DSLR	Digital Single-Lens Reflex
FC	Fully Connected
FCN	Fully Convolutional Network
FC-CNN	Fully Convolutional Convolutional Neural Network
FC-LSTM	Fully Connected Long Short-Term Memory
FPN	Feature Pyramid Network
GAN	Generative Adversarial Network
GAP	Global Average Pooling
GMP	Global Max Pooling
GloVE	Global Vectors
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
NN	Neural Network
PReLU	Parametric Rectified Linear Unit
ResNet	Residual Neural Network
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
VAE	Variational AutoEncoder
VGG	Visual Geometry Group
YOLO	You Only Look Once

Chapter 1

Introduction

This thesis is focused on developing a computer vision and machine learning system for intelligent management of surgical tools. This introduction explains why this is a problem, reviews current methods to tackle the problem, evaluates possible solutions, identifies issues with current solutions, states the research hypothesis of this thesis, and lists the research questions. This section also itemizes contributions of the research, details of publications produced from the work, and outlines the overall organisation of the thesis.

1.1 The Problem

The management of surgical tools is a significant problem in hospitals worldwide. For example, one hospital in New Zealand had 23 Operating Theaters working with multiple surgeries being conducted every day. There are many different surgical procedures, encompassing many different surgical specialities, and each procedure has its own set of tools — on average, there are 38 surgical tools per set, with around 6 sets or trays used for a surgery (Mhlaba et al., 2015). Given that every individual surgeon has his/her own preferences and requirements, manual management of such a vast and complex task is extremely difficult. There are streams of used tools coming out of every theatre, and these tools have to be collected, cleaned, sorted, sterilised and packed. Currently this is done manually, and one hospital in New Zealand had 61 trained technicians working 24/7 to accomplish this task (Unit Manager, personal communication, Nov. 2019), while a US hospital employed 89 staff and processed over 23,000 surgical tools trays per month (Alfred et al., 2021). Challenges included high inventory levels, set assembly errors, inconsistent availability of surgical tools, and non-functional instruments. Large volumes and varieties of surgical tools also pose a formidable challenge for management (Unit Manager, personal communication, Nov. 2019).

Surgical sets -— with up to two hundred objects — are currently assembled manually against checklists but errors are found in many assembled sets in terms of sizes and types of tools included (Guedon et al., 2016; Stockert and Langerman, 2014; Zhou, Rueckert, and Fichtinger, 2019). A U.S. study found 3900 defects in 41,799 surgical sets evaluated. 17.6% of the sets had missing tools, 10.9% had broken, damaged or malfunctioning tools, 8.5 % sets had the wrong tools, and 7.1 % (281) of the sets were assembled incorrectly (Alfred et al., 2021). A system that provides immediate information with pictorial guides for identifying tools and locations in sets would be invaluable. On completion, image scans will provide confirmation of accurate assembly within seconds instead of requiring manual checking. Instrument technicians cannot physically inspect each surgical instrument since they are working under time pressure, but a system that captures tool images and indicates possible damage — or which permits images to be examined at a later time for damage —

can provide a solution. Tracking tools at various key points can lower incidents of lost and missing tools — these include scan stations before and after the tools enter the surgery, at decontamination rooms, in packing areas, and just before sealing and sterilisation. Images of completed sets will reduce current reliance on error-prone labels on sterile, sealed sets. Data on available tools across multiple locations can result in better inventory management. Savings of millions of dollars in direct, indirect and inventory costs can be made across the twenty District Health Boards (DHBs) and over two hundred hospitals in New Zealand by adopting this computer vision and machine learning based surgical tool management system.

Convolutional neural networks (CNNs) have been used in medical imaging to identify fibroids, polyps and tumours, and for classification of cancer cells. Modelling of disease progression — Alzheimer's, multiple sclerosis, and stroke — has been achieved through deep learning based analysis of brain scans (Cao et al., 2019; Pang et al., 2019; Yip et al., 2021; Ali Qadir et al., 2019; Almubarak, Bazi, and Alajlan, 2020). CNNs have also been used to recognise and track surgical instruments (Ahmadi et al., 2018; Sarikaya, Corso, and Guru, 2017; Leppanen et al., 2018; Zhao et al., 2017). These CNNs work well but are specialised systems designed for a specific task, they are governed by rules, and are only able to operate within a particular framework. For example, a CNN trained to detect skin cancer lesions from images is not useful for any other task without retraining or domain adaption. Further, while they can be trained to recognise surgical instruments, they can only provide limited information about the particular instrument based on associated labels and annotations. A solution that can provide detailed and useful information using medical data is needed, and this thesis sets out the work required to develop such a system for surgical tool management.

1.2 The Solution

Convolutional neural networks (CNNs) are used to classify images into specific classes. To accomplish this, CNNs extract a hierarchy of features from input images and use these features to classify or categorize the image into a class. CNNs use a cascade of computing layers, where each layer uses the output from the previous layers to extract relevant features from the image data (He, Zhang, Ren, et al., 2016; Krizhevsky, Sutskever, and Hinton, 2012; Simonyan and Zisserman, 2014). Deep learning is about deeper neural networks that provide hierarchical representation of the data through various convolutions. It adds depth and complexity into the machine learning models, and transforms data with various functions. Feature learning, or the automatic extraction of features from the data, is an important aspect of deep learning, where higher level features are formed or composed from lower level features. Each layer of the deep CNN learns a different level of abstraction in terms of what the features capture, e.g., low-level edges and patterns are captured in the lowest layers and abstract object shapes are captured in the highest layers. It has been demonstrated in prior work that the first layers of a deep CNN learn similar features — early layer features capture more general features, and later layers capture more specific features. This is important, since general features that are common for different classification tasks can be transferred to a new task to provide a head-start for the target learning process (LeCun, Bengio, and Hinton, 2015; Pan and Yang, 2010; Weiss, Khoshgoftaar, and Wang, 2016; Zhou et al., 2014; Zhuang et al., 2019).

Deep neural networks are ideal for the processing of image data, which are arrays of pixel intensities in three colour channels, and have proven to be very successful in

the segmentation, detection and recognition of objects in images (LeCun, Bengio, and Hinton, 2015). Image segmentation divides a visual input or image into sets of pixels, and then these segments are used either to classify images into classes or to detect specific objects in the image. In the case of this research proposal, specific surgical tool objects have to be detected and CNNs have been successfully used for surgical tool segmentation, recognition and detection tasks. Research in surgical instrument identification has addressed robotic and computer-assisted surgery (Sarikaya, Corso, and Guru, 2017), instrument position recognition in minimal invasive surgery (Zhao et al., 2017), pose recognition in surgical training (Leppanen et al., 2018) and instrument tracking in hospital inventory management (Ahmadi et al., 2018). Object-specific learning and classification, with or without CNNs, relies on identifying features of input images that can be aggregated into representations. Representation learning allows a computer to automatically discover the representations needed for detection or classification, without the need for time and expertise intensive hand-engineering of features. Deep-learning methods consist of multiple levels of representation, with non-linear modules that transform representations at each level into a representation at a higher, more abstract level. The key aspect of deep learning is that these features are not hand crafted but are learned from the data by relying on learning procedures (LeCun, Bengio, and Hinton, 2015).

1.2.1 Surgical Tool Recognition

CNNs have been successfully used for surgical instrument detection, segmentation and recognition. The performance of object detection CNNs in recent years has been stated to be extraordinary, even human level on specific tasks (LeCun, Bengio, and Hinton, 2015). Image-based surgical tool detection and tracking methods have been the subject of many applied research efforts, and have progressed alongside advances in general object detection. Transfer learning techniques, where a CNN model pre-trained on general images can be fine-tuned on a surgical instrument database, have been used to achieve good accuracy and predictive performance for surgical instrument recognition tasks (LeCun, Bengio, and Hinton, 2015). While significant amount of work has been conducted in this area, Bouget et al. (2017) stated that much more data needs to be made available for algorithm development, and the lack of quality data is a significant handicap for research. Other researchers have also raised concerns about the lack of quality data for research and algorithm development, and have called for the release of more surgical tool datasets into the research community so that better models can be generated (Twinanda et al., 2016). These concerns are addressed in this thesis.

CNNs are now the predominant approach for computer vision based object recognition and detection, and have been successfully used for the detection, segmentation and recognition of objects and regions in images over the last two decades, including surgical tools detection. However, there are significant problems with this approach for surgical tool management that need to be addressed, and these problems are discussed below:

Volume — Real-world image recognition needs to detect from many thousands of classes, but most CNNs are trained to recognise only a few classes. In many cases, CNNs have been trained on ImageNet which has 1000 classes, and this is not sufficient for real world tasks where tens of thousands of classes have to be recognised. This is particularly true in surgical tool management — the CNNs discussed in surgical tool applications have dealt with very small instrument sets with a maximum of 21 tools

available for research. The Cholec80, EndoVis 2017 and m2cai16-tool datasets have 7 instruments, the CATARACTS dataset has 21 instruments, the NeuroID dataset has 8 instruments and the LapGyn4 Tool Dataset has 3 instruments (Al Hajj et al., 2018; Al Hajj, Lamard, Conze, et al., 2019; Twinanda et al., 2016). There are many thousands of surgical tools in circulation within a hospital but no work has been conducted to train CNNs to recognise these thousands of tools. Stockert and Langerman (2014) reported that one institution processed over 100,000 trays and 2.6 million instruments annually. NYC's Hospital for Special Surgery processed 900 surgical trays per day, and Wythenshawe Hospital in the UK reported 85,000 surgical trays used in 69,000 surgeries every year. The UK Govt reported that there were at least nine million individual surgical trays in circulation in the NHS at any one time. There were approx. 38 instruments per tray (range, 1–188), with about 5.4 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.

Variety and Fine Grained Classification — There are tens of thousands of different surgical tools, and each tool varies in shape, size and functionality. In many cases, the basic shape of two or more different tools is similar with very subtle differences in key attributes differentiating them. CNNs are generally trained on high-level image features and cannot capture fine-grained details of the classes. Fine-grained recognition is required to classify categories that are visually similar, but the current approach of capturing and relying on holistic image features is insufficient for distinguishing fine-grained classes. In these classes, discriminative information is available in only few regions in the image, in small sections of the object, and correspond to a limited number of attributes (Huynh and Elhamifar, 2020). How to recognise these fine-grained discriminative features is a critical aspect that is addressed in this research. It is possible to train multiple CNNs for recognising particular types of tools, but deployment of multiple models for prediction of tools in real world conditions, on systems which may not have adequate memory or computing power, may not be possible. A solution that relies on a single CNN to make predictions across multiple levels would be ideal, and this thesis evaluates how this can be achieved.

Complexity in Management — The organisation or assembly of surgical tools for specific surgical procedures is governed by multiple factors. While the volume and variety of tools presents a significant problem, another important issue is the surgeon preference card, which enumerates each item that a particular surgeon requires when performing a given surgical procedure. Since a single surgery requires many different instrument and tray types, tool management requires evaluation of the specific surgical trays needed for a given surgeon and procedure, the types and numbers of instruments required in each tray type, additional tools needed, and the actual availability of such tools and trays in inventory (Ahmadi et al., 2018). Manual management of these complexities inevitably leads to errors. Global research studies discovered multiple errors in surgical tray assembly, which put surgical procedures and patients at risk (Guedon et al., 2016; Stockert and Langerman, 2014; Zhou, Rueckert, and Fichtinger, 2019). Shifting from the manual handling of surgical tools to a deep learning based system could address these errors.

The costs involved are high — a US institutional cost center reported spending over eight million dollars annually on processing 2.5 million instruments over the course of a year, and total annual institutional cost associated with instrument processing was estimated at over three dollars per instrument. A US Medical Center used

data from 2 neurosurgical procedures to estimate potential institutional savings up to 2.8 million a year through a reduction in instrument processing through sterile supply, demonstrating the scale of savings that are possible (Mhlaba et al., 2015; Stockert and Langerman, 2014; Meter and Adam, 2016; Zhou, Rueckert, and Fichtinger, 2019). Alfred et al. (2021) highlighted how defects in assembled sets could cause increased risks, delays and costs – urgent replacement tool handling could cause surgical site infections, using alternate tools could result in greater risks, delays or deviations in surgery processes, and surgery delays could be expensive. Further, broken and/or defective instruments could lead to in-patient retained objects, as well as skin or tissue tearing and damage, and greater risk of infection (Alfred et al., 2021).

Illumination, Reflection and Occlusions — Illumination variations cause significant problems, particularly in real world conditions where light sources can vary from direct sunlight, filtered natural light, LED lighting, incandescent or fluorescent light or different combinations at different times. These variations, along with the reflective nature of most surgical tools, cause significant problems for effective and accurate tool identification. Added to this is the fact that tools are often occluded or stacked, and also may have foreign matter such as blood, bone or tissue material. Ensuring that a CNN is robust to changes in illumination and background is an important issue that is addressed in this thesis.

Quality Assured ML-Driven Deep Learning — Given the mission-critical nature of surgical tool management, there needs to be validation of deep learning systems in real-world conditions. This was to addressed in this thesis but was not possible due to issues with the global pandemic which restricted access to hospitals and clinics.

1.3 Research Questions

Research on surgical tools has focused on a small number of tools and in limited contexts and scenarios. There is a lack of quality datasets for surgical tool detection tasks, and calls have been made by researchers for the development of high quality surgical tool datasets to facilitate research in this area. A robust and reliable surgical tool detection and identification system that can address the issues and problems enumerated in previous sections would be invaluable. This thesis therefore has its focus on developing a computer vision and machine learning based surgical tool management system.

The research hypothesis is framed as follows:

> A hierarchical, informed, robust machine learning based system can be developed for effective management of surgical tools.

1.3.1 Research Questions

The specific research questions are enumerated below:

- RQ1 — How can a convolutional neural network be designed for recognition of surgical tools, effectively utilising the hierarchical nature of surgical tool classes?

- RQ2 — How can the design of a CNN be improved for interpretable deep learning for intelligent surgical tool management, by incorporating prior information and knowledge of relationships in the ground truth class label arrangements?
- RQ3 — How can the robustness of a CNN be improved for recognition of surgical tools under challenging conditions, addressing volume, variety, complexity and illumination/reflection/occlusion issues?
- RQ4 — How can nominal attribute information be included in a machine learning model to improve the predictions of a CNN for surgical tool management?

1.4 Thesis Contribution, Scope and Limitations

The contributions that this thesis makes are listed as follows:

- The research proposal focuses on designing a computer vision and deep learning based system that could operate effectively in the surgical domain.
- Domain Knowledge is developed in the form of a Surgical Tool Dataset and a Surgery Knowledge Base which will support further research in this area.
- A prototype for computer vision based intelligent management of surgical tools is developed to demonstrate the effectiveness, efficiency and accuracy of the system.
- A key contribution is a deeper insight into how to design and train a CNN for practical deployment while addressing real world problems. The practical results of this research may be useful for hospitals, DHBs and the government, and can contribute to greater efficiencies and cost savings across these organisations.
- A significant limitation was that extensive testing was not possible in actual use conditions due to unforeseen problems with the COVID pandemic, which limited access to hospitals and test areas.

1.4.1 Publications

The work accomplished as part of this thesis has resulted in the following publications:

- 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 — 2nd Place — Best Applied Paper Award.
- Rodrigues M, Mayo M, Patros P (2021b), Interpretable deep learning for surgical tool management. In: Reyes M. et al. (ed) 4th International Workshop on Interpretability of Machine Intelligence in Medical Image Computing (iMIMIC 2021), Springer, Cham., Lecture Notes in Computer Science, vol 12929, DOI — 10.1007/978-3-030-87444-51
- Rodrigues M, Mayo M, Patros P (2022a), OctopusNet: Machine learning for intelligent management of surgical tools. Smart Health Volume 23, DOI — 10.1016/j.smhl.2021.100244

- Rodrigues M, Mayo M, Patros P (2022b) Surgical Tool Datasets for Machine Learning Research: a Survey, International Journal of Computer Vision, Volume 130, Pages 2222–2248 (2022). DOI — 10.1007/s11263-022-01640-6

1.5 Thesis Organisation

- Chapter 1 — Introduction to the thesis, where the problem is introduced, possible solutions are identified and work conducted are discussed. The research hypothesis is defined and the research questions are listed. This section itemizes contributions of the thesis, details of publications produced from the work, and it outlines the overall organisation of the thesis.

- Chapter 2 — This chapter sets up the context of the research, and is a detailed literature review of surgical tool management with CNNs. The work presents a comprehensive survey of datasets for surgical tool detection and related surgical data science and machine learning techniques and algorithms. It 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. The 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.

- Chapter 3 — This chapter explores and implements state of the art methods and frameworks for surgical tool detection. It reports the results of experiments on two important frameworks — Mask R-CNN and YOLOv3 / YOLOv5. This work implements the frameworks, and explores ways in which results can be improved. It then conducts experiments with hierarchical predictions, and evaluates the use of infra-red images. Work accomplished also evaluated real time performance, and the use of differential annotations. In particular, approaches to tackle the problems posed by illumination changes, reflections, background variations and cluttered trays are discussed and evaluated.

- Chapter 4 — This chapter presents a novel convolutional neural network framework for multi-level classification of surgical tools. The classifications are obtained from multiple levels of the model, and high accuracy is obtained by adjusting the depth of layers selected for predictions. The framework enhances the interpretability of the overall predictions by providing a comprehensive set of classifications for each tool. This allows users to make rational decisions about whether to trust the model based on multiple pieces of information, and the predictions can be evaluated against each other for consistency and error-checking. The multi-level prediction framework achieves promising results on a novel surgery tool dataset and surgery knowledge base, which are important contributions of this work. This framework provides a viable solution for intelligent management of surgical tools in a hospital, potentially leading to significant cost savings and increased efficiencies.

- Chapter 5 — This chapter presents a novel densely-connected convolutional neural network architecture, termed OctopusNet, for hierarchical classification of surgical tools. The network shares information across prediction hierarchies to improve classification accuracy, and provides a degree of interpretability by

predicting a set of features for each tool based on multiple classification targets. Important contributions of this chapter work are the OctopusNet architecture, a novel surgical tool dataset and a surgery knowledge base. The architecture provides a viable solution for intelligent management of surgical tools in a hospital, potentially leading to greater patient safety, significant cost savings and increased efficiency.

- Chapter 6 — A new hierarchically organised dataset for artificial intelligence and machine learning research is presented, focusing on intelligent management of surgical tools. In addition to 360 surgical tool classes, a four level hierarchical structure for the dataset was created, defined by 2 specialities, 12 packs and 35 sets. The work employed different convolutional neural network training strategies to evaluate image classification and retrieval performance on this dataset, including the utilisation of prior information in the form of a taxonomic hierarchy tree structure. In this work, the effects of image size and the number of images per class on model predictive performance was evaluated. Experiments with the mapping of image features and class embeddings in semantic space using measures of semantic similarity between classes show that providing prior information results in a significant improvement in image retrieval performance on the dataset.

- Chapter 7 — This chapter evaluates ways of making the OctopusNet robust, so that it can cope with real world conditions. Experimental work to test the CNN performance with challenging images was conducted, and performance was evaluated. Further work developed and used synthetic data and filter based purposeful augmentation, and good results were obtained by training with the synthetic augmented dataset.

- Chapter 8 — This chapter addresses the utility of multi-modal representations in the intelligent management of surgical tools, and the provision of additional information in the training process to improve predictive performance of the CNN. Experiment were conducted on the use of attributes as external sources of knowledge, and the attribute matrix was used to provide prior information in the training regime of the CNN. The work focuses on text and knowledge graphs to formally represent prior knowledge and on a learning algorithm approach for knowledge integration, implemented as a loss function and regularizer in the training process. The approach uses both images and text in the CNN training process in an effective manner.

- Chapter 9 — This chapter presents the thesis conclusions, demonstrates how the research questions have been addressed, reports problems and shortcomings, and the response to the research hypothesis. It also provides directions for future work.

Chapter 2

Surgical Tool Datasets for Machine Learning Research: a Survey

This chapter was published in the International Journal of Computer Vision, Volume 130, Pages 2222–2248 (2022). DOI — 10.1007/s11263-022-01640-6.

It presents 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. This 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.

2.1 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 (Figure 2.1).

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, Corso, and Guru, 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

FIGURE 2.1: Open Surgery Instruments

multiple surgery procedures. Amsterdam, Clarkson, and Stoyanov (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, Zhao, and Hu (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 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 2.1, 2.5, 2.6, 2.7, 2.8 and 2.9). 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.

2.2 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 2.2 and 2.3. 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.

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, Lamard, Conze, 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, algorithms are developed and applied, and solutions are tested on a held-out test set. A critical component is the dataset provided, and every attempt

TABLE 2.1: Surgical Tool Datasets — Examples of data and instruments

Challenge Name	Data Available	Instrument Nos
ROBUST-MIS 2019 (Ross, Reinke, and Full, 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, Lamard, Conze, et al., 2019)	50 videos of phacoemulsification cataract surgeries	21 surgical tools
Cholec80 (Twinanda, Shehata, Mutter, 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, Corso, and Guru, 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 2.2: 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 2.3: 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

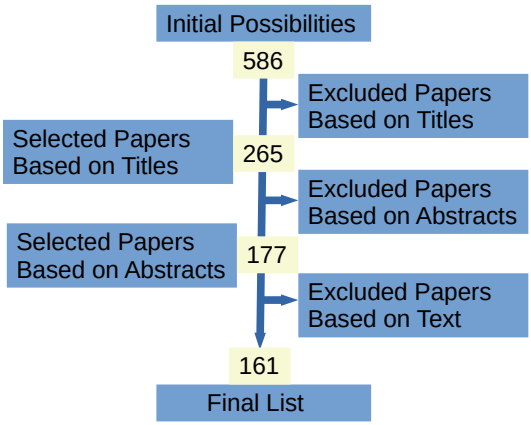


FIGURE 2.2: Paper Selection Flow

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.

2.3.1 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: ten rectal resection procedures, ten proctocolectomy procedures and ten sigmoid resection procedures. A labelling mask and instrument labels were manually created for the 10,040 extracted endoscopic video frames (Ross, Reinke, and Full, 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 2.4). 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

TABLE 2.4: 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

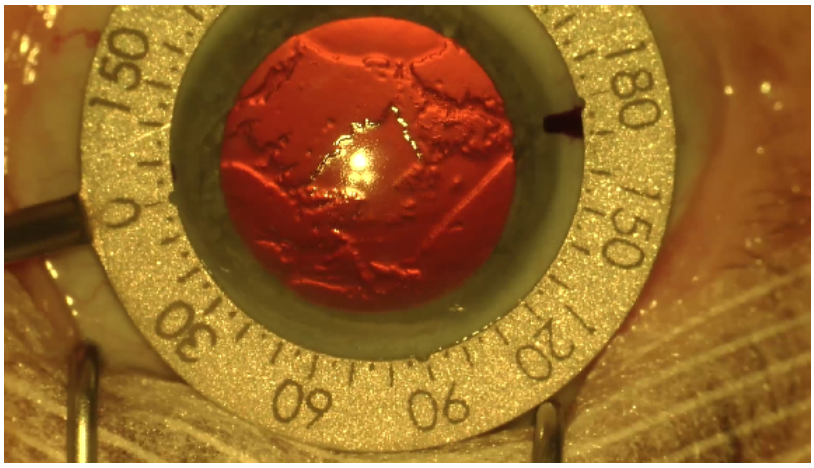


FIGURE 2.3: Cataracts Dataset

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 2.4); 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, Lamard, Conze, et al., 2019).

The Cholec80 dataset contains 80 videos of cholecystectomy surgeries, and seven tools or instruments are present in the dataset (Table 2.4). 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, Shehata, Mutter, et al., 2017). The m2cai16-tool dataset is a subset of the Cholec80 Dataset and it consists of fifteen cholecystectomy videos with binary

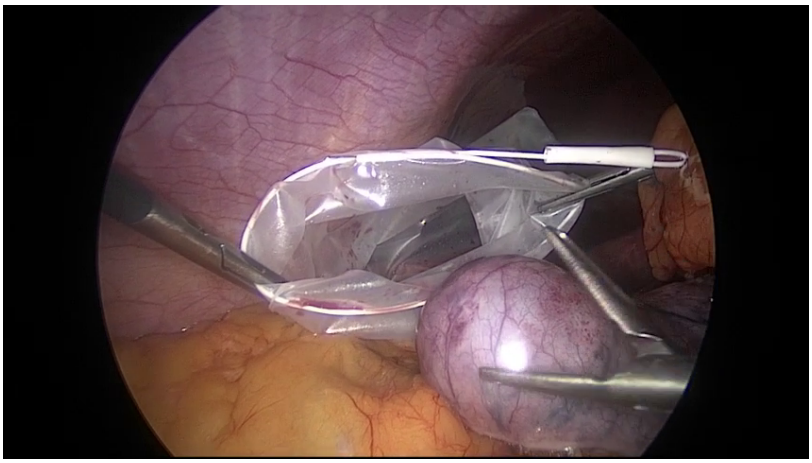


FIGURE 2.4: Cholec80 Dataset

annotations of the seven tools present.

Details of the surgical tool datasets used in the challenges is presented in Table 2.5. In addition to the metadata provided about each dataset, additional metadata characteristics 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.

2.3.2 Other Surgical Tool Datasets

In addition to the challenge datasets described above, many other surgical tool datasets have been developed and we present these datasets in Tables 2.6 and 2.7. 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, Corso, and Guru, 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 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).

TABLE 2.5: 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	—

TABLE 2.6: Surgical Tool Datasets – 1

Dataset	Focus	Data Type	Data Quality	Instr.	Licence	Organisation	Annotations
Atlas Dione (Sarikaya, Corso, and Guru, 2017)	Urology – Urethrovesical Anastomosis	86 videos and 910 clips	854 × 480 pixels	3	Open	22,486 frames	Bounding Boxes
Bar's Cholecystectomy (Bar, Neimark, and 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, Taschwer, and 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	8,080 frames	854 × 480 pixels	2	Open	—	Semantic segmentation masks
Flapnet (Attanasio, Scaglioni, and al., 2020)	Lobectomy	2,160 images	506 × 466 pixels	1	Open	—	Instrument Presence
Garcia Perez's RoboTool (Garcia-Peraza-Herrera et al., 2021a)	Laparoscopic Surgeries	20 surgical procedures	—	—	Public	6130 images	514 manually annotated images
Gruijthuijsen's Gynaecology (Gruijthuijsen, Garcia-Peraza-Herrera, and al., 2021)	Gynaecology	1180 Images	1920 × 1080 pixels	—	Private	1110 Training and 70 Testing Images	Bounding Boxes in 379 images

TABLE 2.7: Surgical Tool Datasets – 2

Dataset	Focus	Data Type	Data Quality	Instr.	Licence	Organisation	Annotations
Hasan's ART-Net (Hasan et al., 2021)	Gynaecology	29 Videos	—	—	Private	1016 training and 3254 testing Images	Segmentation Masks
Heidelberg's Colo-rectal / HeiCo / Hei-Chole (Maier-Hein et al., 2021)	Colon-Rectal Surgeries	30 Videos	960 × 540 pixels	3	Open	10040 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
Huaulme's MISAW (MICCAI – 2020) (Huaulme, Sarikaya, and al., 2021)	Anastomosis	27 sequences	920 × 540 pixels	1	Open	17 training and 10 test sequences	—
Jha's Kvasir-Instrument (Jha et al., 2021b)	General Surgery	3,500 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)	Otorhinolaryngology	—	—	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

TABLE 2.8: Surgical Tool Datasets – 3

Dataset	Focus	Data Type	Data Quality	Instr.	Licence	Organisation	Annotations
Law's Vesico-Uritheral Anastomosis (Law, Ghani, and Deng, 2017)	Urology	12 Videos	720 pixels	—	Private	146,309 Frames	—
Lee's Phantom (Lee et al., 2019b)	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	97932 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, Moreno, and Arenas, 2017)	General Surgery	7000 Images	480 × 480 pixels	5	Private	2000 training and 5000 testing images	—
Nakawala's Nephrec9 (Nakawala et al., 2019)	Urology	9 Videos	720 × 578 pixels at 25 FPS		Private	741,573 Frames	Multiple Instrument Annotated

TABLE 2.9: Surgical Tool Datasets – 4

Dataset	Focus	Data Type	Data Quality	Instr.	Licence	Organisation	Annotations
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, Edwards, and Stoyanov, 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, Muller-Stich, and 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	8,572 training and 2,144 validation images	Bounding Box
Yang's Cardiac (Yang et al., 2019a)	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

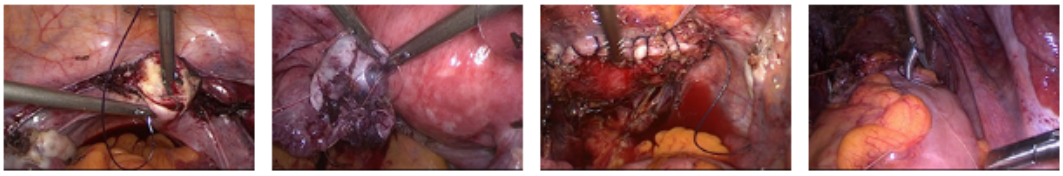


FIGURE 2.5: Lapgyn Dataset

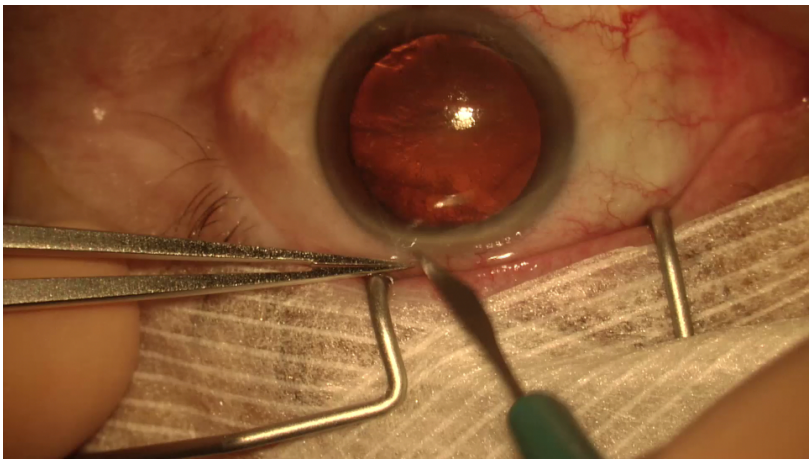


FIGURE 2.6: CaDIS Dataset

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, Couperus, and Wachs, 2020), the MISAW dataset used for the MIcro-Surgical Anastomose Workflow recognition on training sessions challenge (Huaulme, Sarikaya, and al., 2021), the Bypass40 dataset of laparoscopic gastric bypass procedures (Ramesh et al., 2021b), and the EAD2020 dataset (Ali, Dmitrieva, and al., 2021).

2.4 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 Figure 2.7, 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, Lamard, Conze, et al. (2019) state that these tasks can be categorized according the precision of the desired outputs, with the finest or more

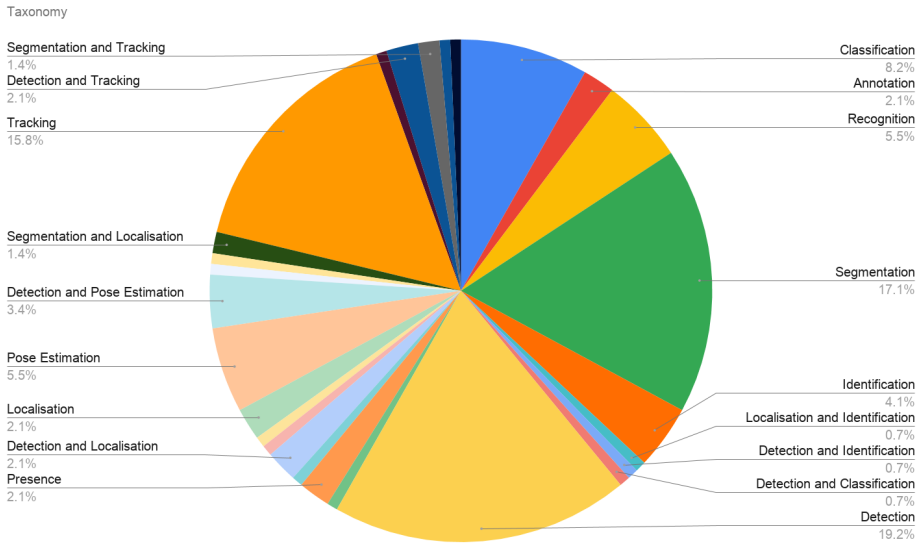


FIGURE 2.7: Taxonomy of Approaches

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.

2.5 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. (2017b) 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, Lamard, Conze, et al. (2019) reported on the results of surgical instrument presence detection with the CATARACTS Dataset. This included work using VGG-16 (Simonyan and 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, Shehata, Mutter, et al. (2017) developed and used the Cholec80 dataset to test EndoNet, an architecture based on AlexNet, for tool detection. Sahu et al. (2017a) fine-tuned AlexNet on the m2cai16-tool dataset; using an approach similar to EndoNet. The Cholec80 dataset was used by Alshirbaji, Jalal, and Moller (2018) to fine tune AlexNet for surgical tool classification. Mondal, Sathish, and Sheet (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. (2021b) 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. (2020b) 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 instrument 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, Li, Dou, 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, Lamard, Conze, 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. (2020a) 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, Wang, and Huang (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, Castro, and Essa (2016) fine-tuned AlexNet, VGG-16 and Inception-v3 and presented a comparison of these different deep network architectures for surgical tool detection. Namazi, Sankaranarayanan, and Devarajan (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, Sathish, and Sheet (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, Schoeffmann, and Husslein (2019) 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, Arenas, and Moreno (2018) developed a tree-structured convolutional neural network for the classification of ten open surgery instruments. Eight separate CNNs were trained on ten surgical instruments, and four CNNs on five instruments. Murillo, Moreno, and Arenas (2017) used five open surgery tools for testing the performance of CNNs and Haar Classifiers (Viola and 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, Li, and Ren (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.

2.6 Tool Localisation Research

Banerjee, Sathish, and Sheet (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. (2021a) 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 EndoVis 2021 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. (2020c) proposed an anchor-free convolutional neural network (CNN) architecture using a compact stacked hourglass (Newell, Yang, and Deng, 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 and Deng, 2020), and lightweight Hourglass – and achieved good accuracy and speed for real-time surgical tool detection. Liu et al. (2020d) 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, Corso, and Guru (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, He, Girshick, et al., 2017), Yolov3 (Redmon et al., 2016), and RetinaNet (Lin et al., 2017) – on the two datasets.

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, Ghani, and Deng (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, Neimark, and 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. (2019a) 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 2,532 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 and Farhadi, 2017) and missing tool detection – to perform detection of surgical instruments in real time.

2.7 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 (25G ILM) forceps.

Al Hajj, Lamard, Conze, et al. (2019) defined tool tracking work in terms of monitoring tool location over time. Gruijthuijsen, Garcia-Peraza-Herrera, and 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. (2019b) 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, Zhao, and Cheng (2017) 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, Sutskever, and Hinton, 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 and 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.

2.8 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. (2021a) 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, Taschwer, and Schoeffmann (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, Reinke, and Full (2019) discussed segmentation solutions based on the ROBUST-MIS challenge, including 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, Fischer, and Brox, 2015), DeepLabV3+ (Chen et al., 2017), TernausNet (Iglovikov and 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. (2021a) 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, Mukhopadhyay, and Zachow (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 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, Bravo-Sanchez, and Arbelaez (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 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, Rosa, and Nageotte (2021) 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, Li, and Ren (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. (2019a) 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. (2021b) 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, Schwaner, and Savarimuthu (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, Padoy, and Rosa (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, J, and Husslein (2019) 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.

2.9 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 surgical 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, Schlüter, and Schlaefer (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.

2.10 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.

2.10.1 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, Schlüter, and Schlaefer, 2018), RGB-D depth (Hiasa et al., 2016), and 3D ultrasound images (Yang et al., 2019a). Multi-modal datasets could potentially be valuable – for example, in their review of surgical activity recognition research, Amsterdam, Clarkson, and Stoyanov (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.

2.10.2 Dataset Volume, Variety and Quality

In a white paper on the first annual Conference on Machine Intelligence in Medical Imaging (C-MIMI), Kohli, Summers, and Geis (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, Amsterdam, Clarkson, and Stoyanov (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 2.10 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 2.1). 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 2.10), 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, Mayo, and Patros, 2022; Rodrigues, Mayo, and Patros, 2021a; Rodrigues, Mayo, and Patros, 2021b).

2.10.3 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, Reinke, and Full, 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

TABLE 2.10: 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, Reinke, and Full (2019)
General surgery	✓	✓	Jha et al. (2021b), Gao et al. (2014), Murillo, Moreno, and Arenas (2017), Bar, Neimark, and al. (2020), Hong et al. (2020), Hou et al. (2022), Wagner, Muller-Stich, and al. (2021), and Twinanda, Shehata, Mutter, et al. (2017)
Gynaecology and obstetrics		✓	Gruijthuijsen, Garcia-Peraza-Herrera, and 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, Taschwer, and al. (2018), Kurmann et al. (2017), N. et al. (2022), Sznitman et al. (2012), and Al Hajj, Lamard, Conze, 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, Corso, and Guru (2017), Law, Ghani, and Deng (2017), and Nakawala et al. (2019)
Vascular surgery			

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 2.10, 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.

2.10.4 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.

Nogueira-Rodriguez et al. (2020) reported that all the publicly available datasets that could be used for object detection 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, Lamard, Conze, 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. (2021a) 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, Summers, and Geis (2017) pointed out that there are no generally accepted standards for the creation and cataloguing of medical image datasets. As we demonstrate in Table 2.6, 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 2.5, 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, Summers, and Geis, 2017).

2.10.5 Metrics

There are an extremely wide range of metrics that have been used in the research. 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, Castro, and Essa (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.

2.10.6 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 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

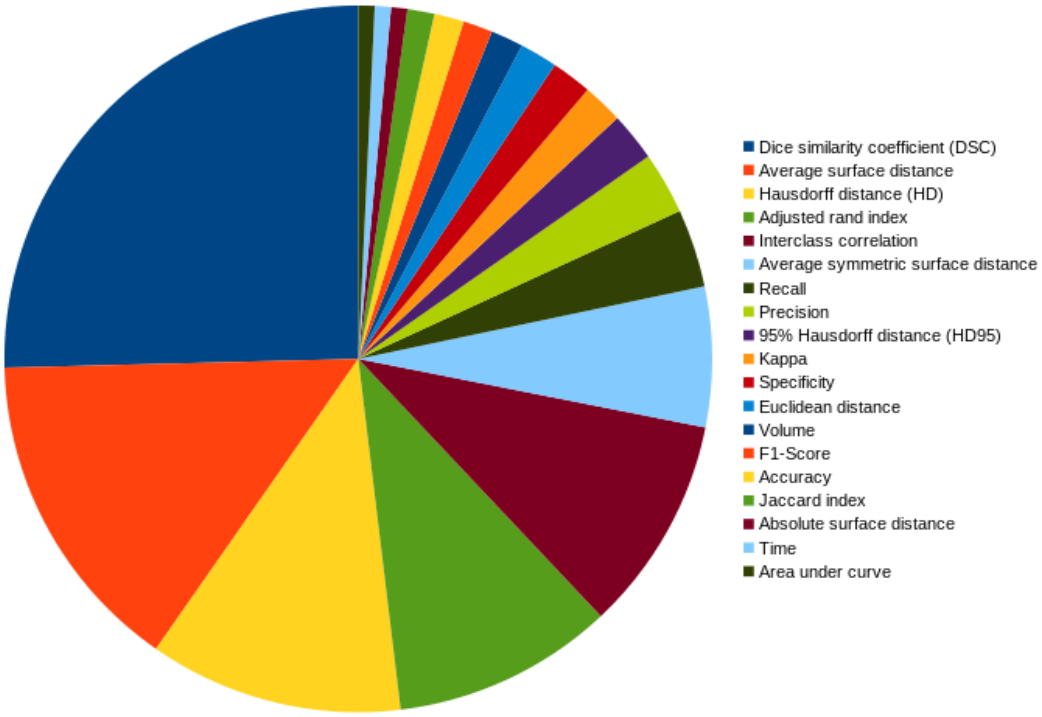


FIGURE 2.8: Range of Metrics Used

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., 2021). 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 2.11 summarises the open research questions and opportunities which we identified and detailed in previous sections of this paper.

2.11 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, Mayo, and Patros, 2022; Rodrigues, Mayo, and Patros, 2021a; Rodrigues, Mayo, and Patros, 2021b). 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

TABLE 2.11: 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 2.12: 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 to 150 cms
Instruments	360
Images/Class	74 images
Organisation	Hierarchical
Annotations	Various - Image labels, Bounding Boxes and Masks

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 2.12, and examples of actual instrument sets and annotations of instruments in Figure 2.9 and Figure 2.10.

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 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,

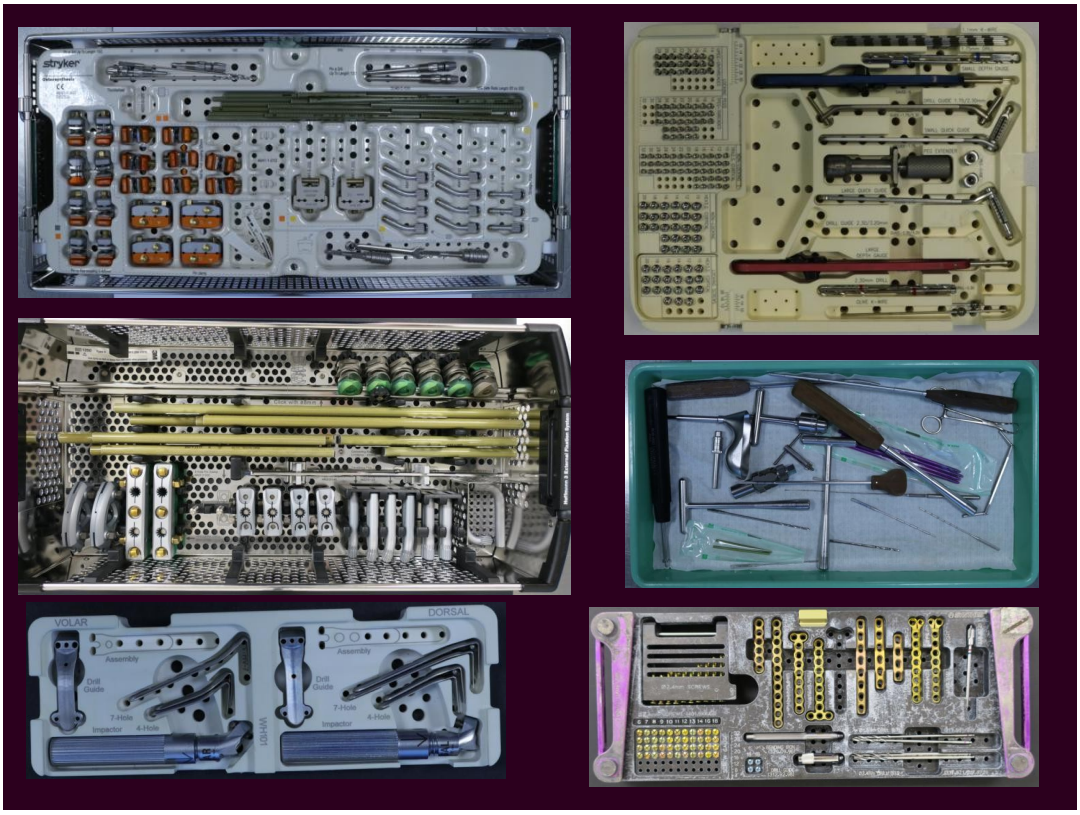


FIGURE 2.9: HOSPI-Tools Sets

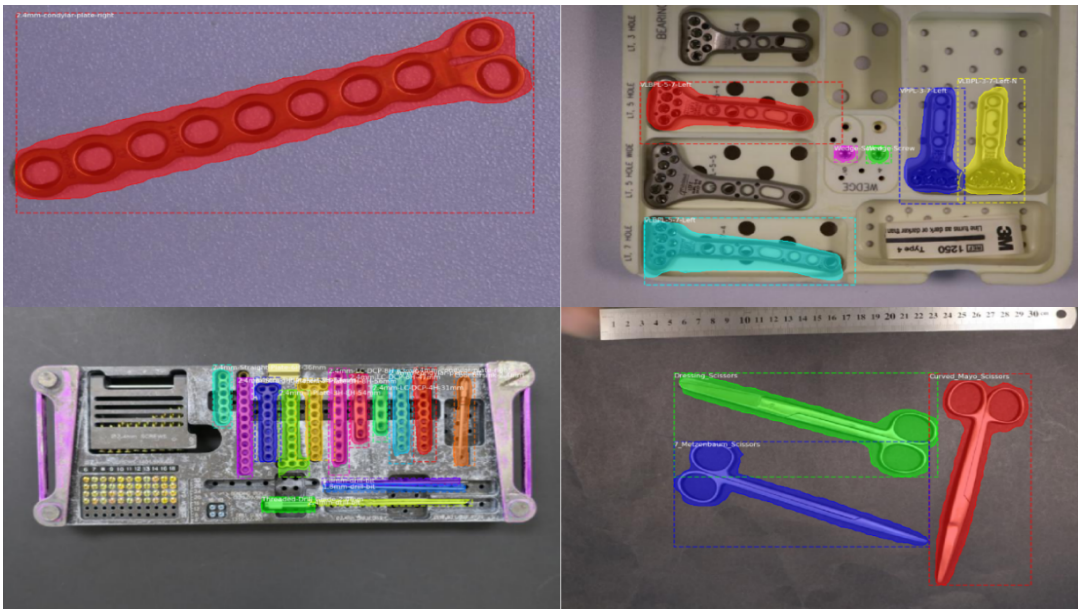


FIGURE 2.10: HOSPI-Tools Annotations

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.

2.12 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`.

Chapter 3

Surgical Tool Detection

This chapter evaluates the performance of state of the art methods and frameworks for surgical tool detection. It accomplishes the following tasks:

1. The work reports the performance of two important frameworks — Mask R-CNN (He et al., 2017) and YOLOv3 (Redmon et al., 2016) — for surgical tool detection;

2. It evaluates how the frameworks perform in real world conditions, and discuss problems posed by illumination changes, reflections, background variations and cluttered trays since these had been identified as important considerations in real world conditions.

3. The research identifies problems and explores ways in which results could be improved. In particular, experiments were conducted with hierarchical predictions, the use of differential annotations, the use of grey scale or binary images, and infra-red images for improved performance.

The results and issues reported in this chapter provide direction and focus for the work that is reported in subsequent chapters.

3.1 Introduction

CNNs are now the predominant approach for computer vision based object recognition and detection, and have been successfully used for the detection, segmentation and recognition of objects and regions in images over the last two decades, including surgical tools detection. This section evaluates current frameworks and approaches for surgical tool detection, and addresses the following two issues:

Volume, Variety and Fine Grained Classification — A major problem in surgical tool management is the sheer volume and variety of tools in use, with new tools continually being added to the asset inventory. Each tool varies in shape, size and functionality, often with very subtle differences in key attributes of tools. This requires fine grained classification, which is a difficult task.

Complex and Cluttered Trays — There are fourteen surgical specialities, ranging from cardiac surgery to neurosurgery (ACS, 2021), and each speciality and procedure uses very specific and purposefully designed surgical tools. On average, a surgical procedure uses 5.4 surgical tool trays or sets, with approx. 38 instruments per tray (range, 1–188) (Mhlaba et al., 2015). Tools are collected after each surgery and presented for cleaning, sorting and sterilisation in a crowded and cluttered manner,

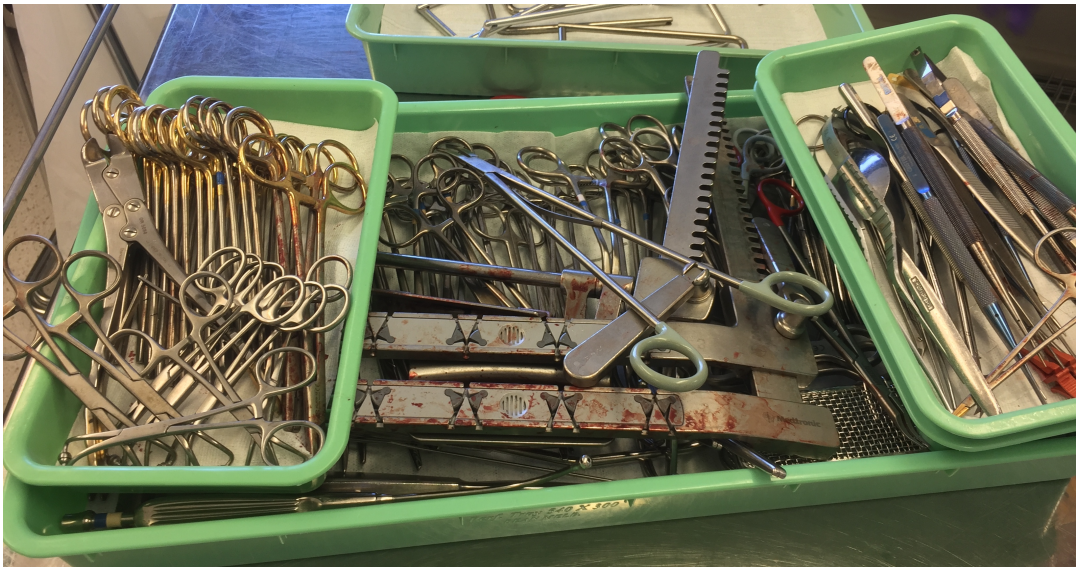


FIGURE 3.1: Cluttered Surgical Tool Tray Example

with minimal separation and isolation after each surgery. This, again, presents a significant problem for surgical tool management.

Object detection is about successfully locating different objects in an image, drawing a rectangle or bounding box around them, and classifying them. Region based CNNs such as R-CNN set new standards and benchmarks in the application of CNNs and deep learning to object detection. This technique used bounding boxes or region proposals to identify the objects in an image; it first proposed a set of boxes within the image, defined as "region of interest" or RoI, and then verified if the RoI actually corresponded to an object. R-CNN used an object proposal algorithm — Selective Search or Edge boxes — which evaluated the image via windows of different sizes, and tried to group adjacent pixels by texture, colour, or intensity to identify objects — 2000 region proposals or bounding boxes were fed to the classifier. R-CNN warped the region to a standard square size and then a CNN based classifier and Support Vector Machine (SVM) classified the object. Finally the system used a linear regression model to define tighter coordinates for the box around the classified object (Girshick, Donahue, Darrell, et al., 2014).

Fast R-CNN improved on R-CNN by training the CNN, classifier, and bounding box regressor in a single model. A final SoftMax layer was used for classification, and a linear regression layer parallel to the SoftMax layer was used for bounding box coordinates. Faster R-CNN added a Fully Convolutional Network on top of the features of the CNN to create a Region Proposal Network, and used anchor boxes to predict the probability of background or foreground. This technique implemented two networks: a region proposal network (RPN) for generating region proposals, and a network which used these proposals to detect objects. Outputs of a region proposal network (RPN) were boxes/proposals that a classifier and regressor checked to verify occurrence of objects (Girshick, 2015; Ren, He, Girshick, et al., 2017).

In SSD, or Single Shot MultiBox Detector, a CNN operated on the input image only once and extracted a feature map. A 3x3 sized convolutional kernel operated on the feature map to predict bounding boxes and classification probabilities. SSD used anchor boxes at various aspect ratios, similar to Faster-RCNN, and learned the offset rather than the actual box. SSD predicted bounding boxes after multiple

convolutional layers, and because each layer used a different scale, it efficiently detected objects of various scales (Liu et al., 2016).

Mask R-CNN extended Faster R-CNN for pixel-level segmentation. In Mask R-CNN, a Fully Convolutional Network (FCN) was added on top of the CNN features of Faster R-CNN to generate a pixel level mask or segmentation output. It added a branch to Faster R-CNN that yielded a binary mask which determined if a pixel was part of an object or not (He et al., 2017). Mask R-CNN has been used in many medical applications, including for detecting cell nuclei in microscopy images (Johnson, 2018), for colon cancer polyp detection and segmentation (Ali Qadir et al., 2019), and in a two stage method for localizing the optic nerve head and segmenting the optic disc/cup in retinal fundus images (Almubarak, Bazi, and Alajlan, 2020). Ciaparrone et al. (2020) tested MASK R-CNN with 12 combinations of CNN backbones and training hyper-parameters for surgical tool detection, and identified best performing configurations in terms of average precision. The Mask R-CNN model used in their work was found to be robust to image artefacts and low-resolution images. Since this is a well established approach, the research conducted in this thesis tested Mask R-CNN for surgical tool detection on a surgical tool dataset, and evaluated the YOLO — or "You Only Look Once" — framework for the tool detection task.

In YOLO, detection was defined as a regression problem which learned the class probabilities and bounding box coordinates. YOLO divided the input image into a grid or regions, and predicted bounding boxes and probabilities in each region. The probabilities or confidence reflected the accuracy of the bounding box and if the bounding box actually contains an object. YOLO also predicted the classification score for each box for every class in training. YOLO used the entire image during training and provided better performance since it used the full context of the object/image, and it was rated to be 1000x faster than R-CNN and 100x faster than Fast R-CNN (Redmon et al., 2016).

Nguyen et al. (2020) evaluated multiple state-of-the-art deep learning models for small object detection and reported that YOLOv3, which performed detection at three different scales, provided impressive real time performance. YOLOv3 has been used for the difficult task of identifying cholelithiasis and classifying gallstones on CT images; a modified architecture achieved 92.7% accuracy in identification of granular gallstones and 80.3% accuracy in muddy gallstones (Pang et al., 2019). Cao et al. (2019) systematically evaluated multiple state-of-the-art object detection and classification frameworks for breast lesions in ultrasound images and reported that YOLOv3 and SSD provided the best performance; YOLOv3 had fewer background errors. In another real time application, Yip et al. (2021) demonstrated that YOLOv3 required minimal training resources and provided fast, accurate neuronal detection on images of live, acute brain slices.

Loncomilla and Solar (2019) proposed a method for object detection based on YOLOv3, and found it to be robust against occlusions, illumination changes, cluttered backgrounds, presence of multiple objects, presence of textured and non-textured objects, and object classes not available in the training set. This is similar to what was needed to be achieved in this thesis, and the work reported in this chapter therefore evaluates the YOLOv3 model for real time and accurate surgical tool identification in complex and cluttered trays.

TABLE 3.1: Results of Mask R-CNN Training Strategies

Exp. No.	Details	Accuracy – Test Set	Accuracy – Real Time
1	Single Mask R-CNN – Single Stage Approach	70.88	
2	Mask R-CNN and ResNet50 Hierarchical classifier	89.00	
3	Mask R-CNN - Hierarchical prediction	90.00	30.00
4	Mask R-CNN trained on binary images with Edges for Pack and Instrument prediction	90.00	70-80

3.2 Methods

A cross section of surgical tools across orthopaedics and general surgery were selected, images were annotated, and MASK R-CNN was trained using the data. However, given the complexity inherent in the dataset, as the number of classes was increased, the accuracy dropped significantly. Experiments were conducted to try to improve Mask R-CNN performance using a limited set of data. A set of 18 instruments were selected from the general surgery tool which reflected complexities, including size variations, fine feature differences (such as toothed or non-toothed forceps), and occlusions. A set of 1400 training images were used for training. This was a small set of images for training a complex model, and strategies to improve training and prevent over-fitting were experimented with in the training methods. In a second approach, YOLOv3 was trained to successfully recognise tools from 18 general surgery classes, even on highly cluttered trays. While performance was tested with YOLOv5, most of the experiments were conducted with the YOLOv3 model.

3.3 Results

Mask R-CNN training on this dataset resulted in a classification accuracy of 31%. Clearly, this is sub-optimal and the incorrect predictions arise from subtle and minor differences in the tool shapes. However, training a different model for each sub-category — for example, for the scissors group — resulted in an accuracy of 72% . The model's performance was improved to 81% by performed 8 rotations of each test image (with each rotation at 45 degrees) and averaging the predictions over the images. Either Mask R-CNN or ResNet50 could be used for this classification. Mask R-CNN was trained for prediction at a higher level, for each category — Scissors, Ring Forceps, Thumb Forceps and Scalpel BP Handle. Four different Mask R-CNN inference models were then trained for sub-category instrument prediction. A hierarchical training approach, in this case a two step approach with Mask R-CNN identifying the first level of the hierarchy and Mask R-CNN or ResNet50 Classifier used for the second stage of identification, averaged for rotated images, then resulted in 96 percent accuracy even with this small training set (Table 3.1 and Figure 3.2).

A further experiment was conducted on using differential annotations – in this case, only the tips of the instruments were annotated, since this was the differentiating factor across the instrument classes that were tested (Figure 3.3). This work resulted in annotation of just the tips of the Ring Forceps category, and subsequent training of

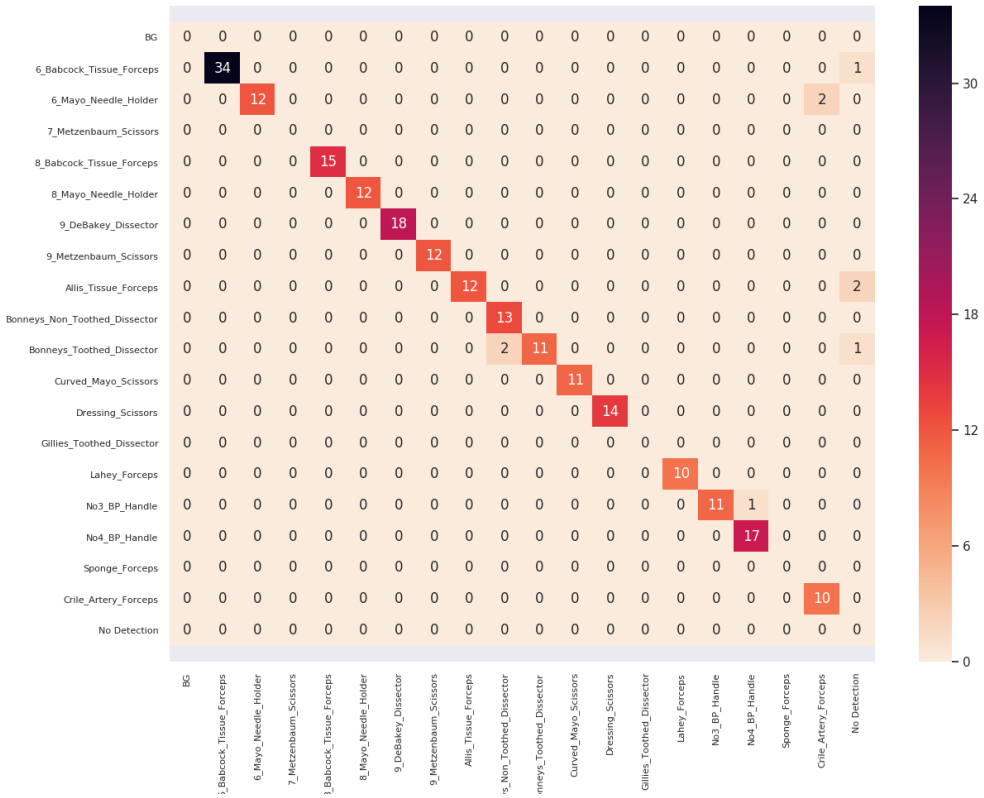


FIGURE 3.2: Mask R-CNN Confusion Matrix

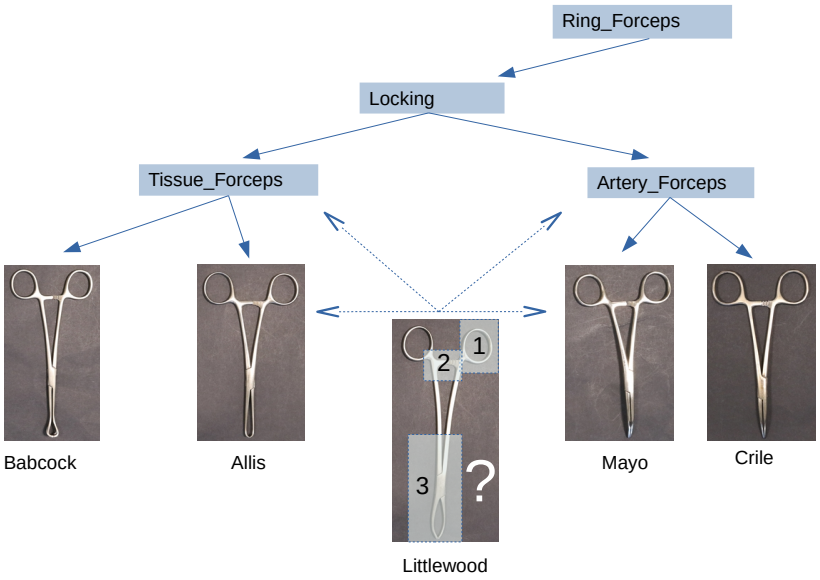


FIGURE 3.3: Examples of Differential Annotations of Surgical Tool Tips

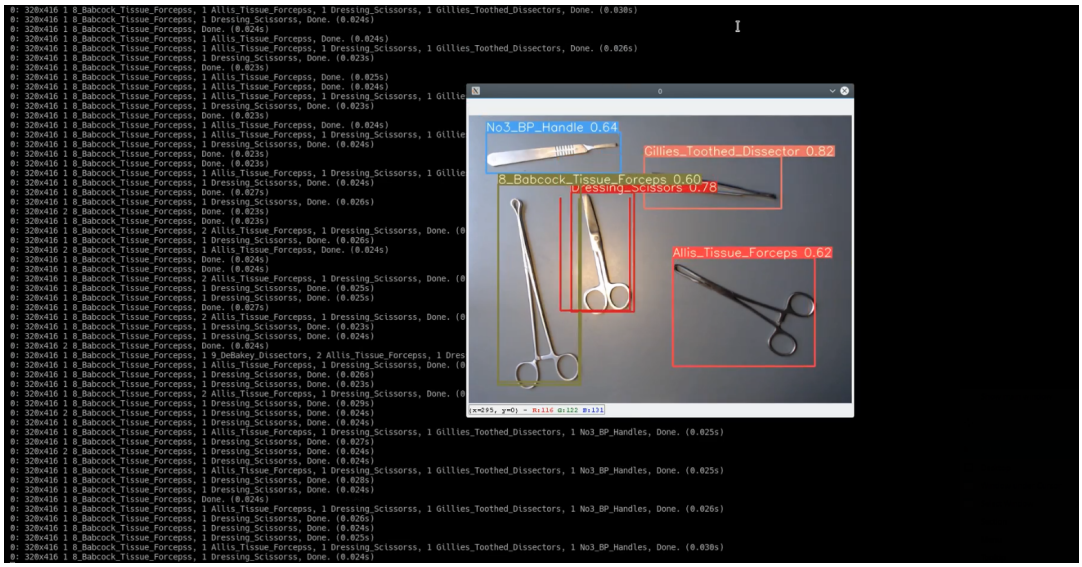


FIGURE 3.4: YOLOv3 Real Time Detection – Results

the model using Mask R-CNN for predictions. This method achieved an accuracy of 65% on independent test data.

Results for YOLOv3 and YOLOv5 were encouraging and the system delivered rapid real time results under challenging conditions. Results were obtained in, on average, 0.0245 seconds – Figure 3.4 and 3.5. Performance was, however, sensitive to light conditions with illumination changes and angles of illumination sources playing a major role in prediction results – Figure 3.6. This is addressed in the next section.

3.4 Illumination and Background Variation Issues

Illumination variations cause significant problems, particularly in real world conditions where light sources can vary from direct sunlight, filtered natural light, LED lighting, incandescent or fluorescent light or different combinations at different times. These variations, along with the reflective nature of most surgical tools, cause significant problems for effective and accurate tool identification within actual hospital conditions. Added to this is the fact that tools are often occluded or stacked, and also may have foreign matter such as blood, bone or tissue material. The framework clearly does not cope well with dedicated, close in lighting and performs best with sunlight, natural and diffused lighting conditions — Figures 3.7 and 3.8.

The YOLOv3 based framework struggled to detect surgical tools in situations where there are coloured, textured or reflective backgrounds, and is effective only with dark-coloured – black and grey, non-textured and non-reflective – backgrounds — Figures 3.9 and 3.10.

Since these frameworks were struggling to cope with changes in illumination and backgrounds, alternative solutions were evaluated and one promising approach was the use of infra-red imaging. Preliminary experiments were conducted to train Mask R-CNN and YOLOv3 on a very small dataset of infrared images, and this used annotations for three general surgery instruments — 6 Mayo Needle Holders, Allis Tissue Forceps and Lahey Forceps — Figures 3.12 and 3.11. Performance was encouraging even acknowledging that this was a very small set of tools and classes (Figure 3.13), but there are practical limitations around deploying an infra-red based system across a hospital.

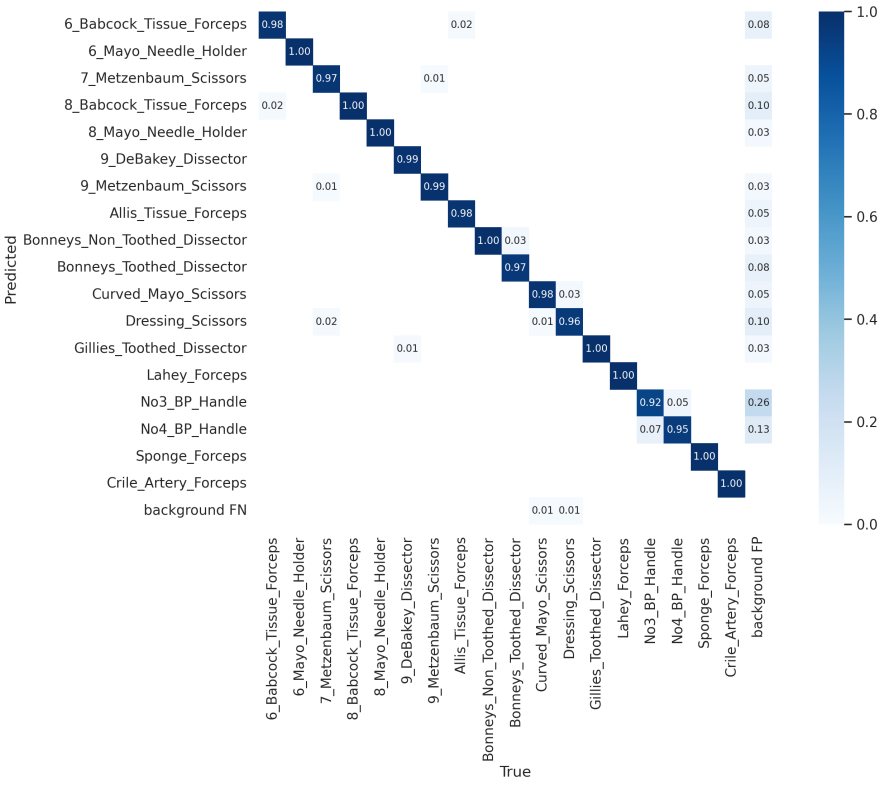


FIGURE 3.5: YOLOv3 Confusion Matrix – Test Set Results

Class	Targets	Precision	Recall	mAP@0.5	F1
all	1350	30.10%	100.00%	0.995	0.448
6_Babcock_Tissue_Forceps	55	8.36%	100.00%	0.995	0.154
6_Mayo_Needle_Holder	68	30.00%	100.00%	0.995	0.461
7_Metzenbaum_Scissors	114	39.00%	100.00%	0.995	0.562
8_Babcock_Tissue_Forceps	57	15.70%	100.00%	0.995	0.272
8_Mayo_Needle_Holder	70	58.80%	100.00%	0.995	0.741
9_DeBakey_Dissector	61	30.70%	100.00%	0.995	0.469
9_Metzenbaum_Scissors	113	23.00%	100.00%	0.995	0.374
Allis_Tissue_Forceps	57	37.00%	100.00%	0.995	0.54
Bonneys_Non_Toothed_Dissector	60	31.60%	100.00%	0.995	0.48
Bonneys_Toothed_Dissector	61	43.60%	100.00%	0.995	0.607
Curved_Mayo_Scissors	94	34.40%	100.00%	0.995	0.512
Dressing_Scissors	108	49.80%	100.00%	0.995	0.665
Gillies_Toothed_Dissector	125	26.40%	100.00%	0.995	0.418
Lahey_Forceps	42	25.50%	100.00%	0.995	0.406
No3_BP_Handle	76	9.09%	100.00%	0.995	0.167
No4_BP_Handle	95	32.60%	100.00%	0.995	0.492
Sponge_Forceps	61	18.00%	100.00%	0.995	0.306
Crile_Artery_Forceps	56	28.60%	100.00%	0.995	0.444

FIGURE 3.6: YOLOv3 Accuracy and Results by Tool Class

FIGURE 3.7: Results Highlighting Problems with Illumination Changes

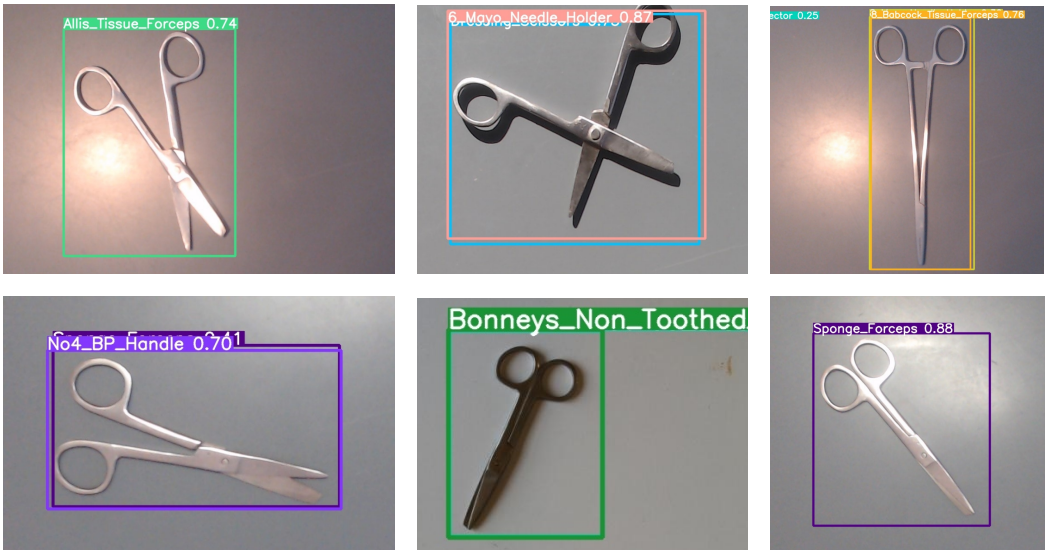


FIGURE 3.8: Incorrect Predictions – Illumination Failures with YOLOv3

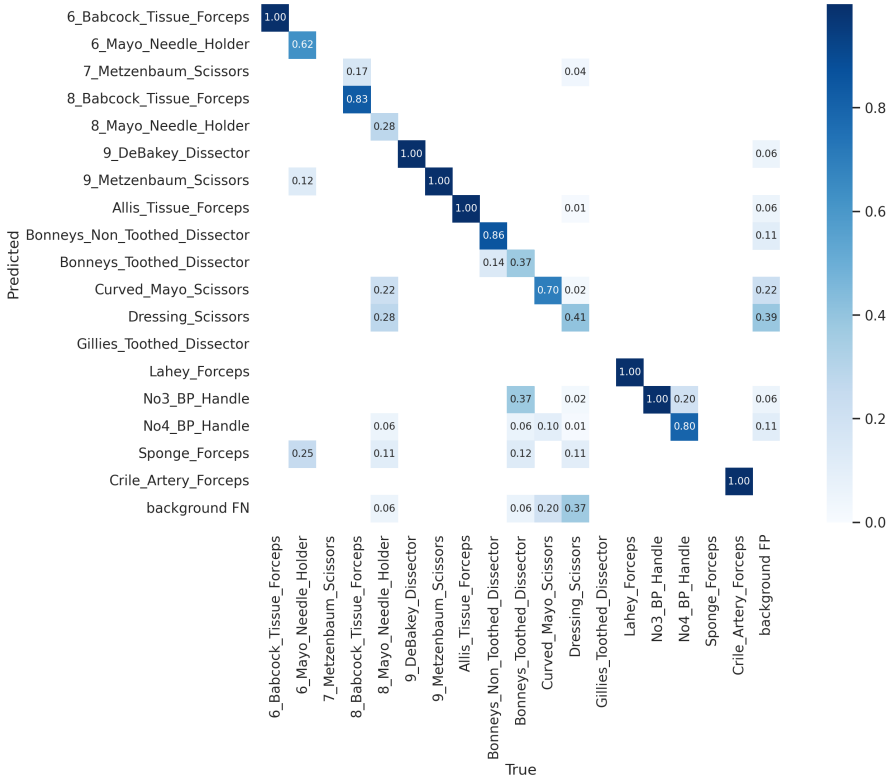


FIGURE 3.9: Confusion Matrix Highlighting Problems with Background Change Results

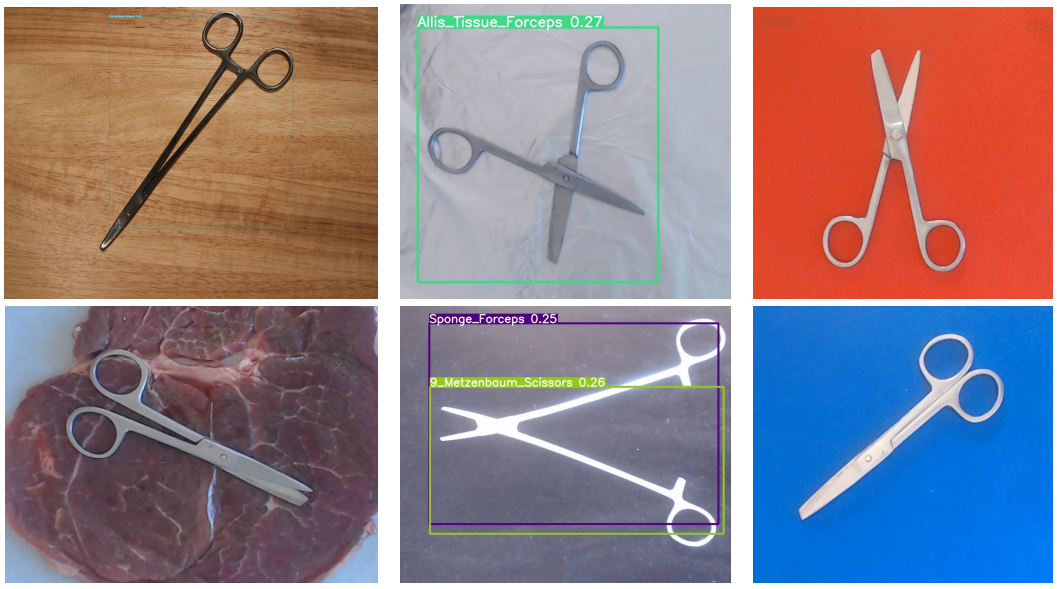


FIGURE 3.10: Prediction Errors with Changes in Background

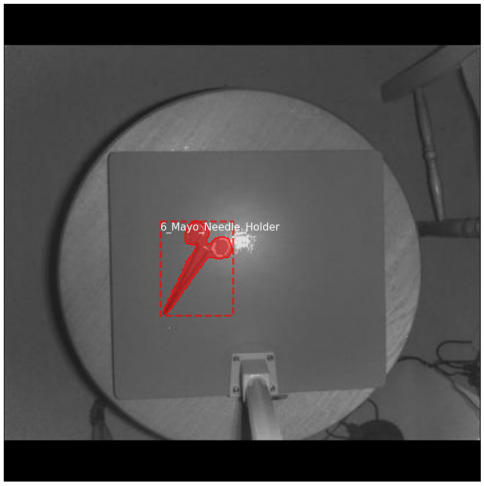


FIGURE 3.11: Infra-Red Results with Difficult Illumination

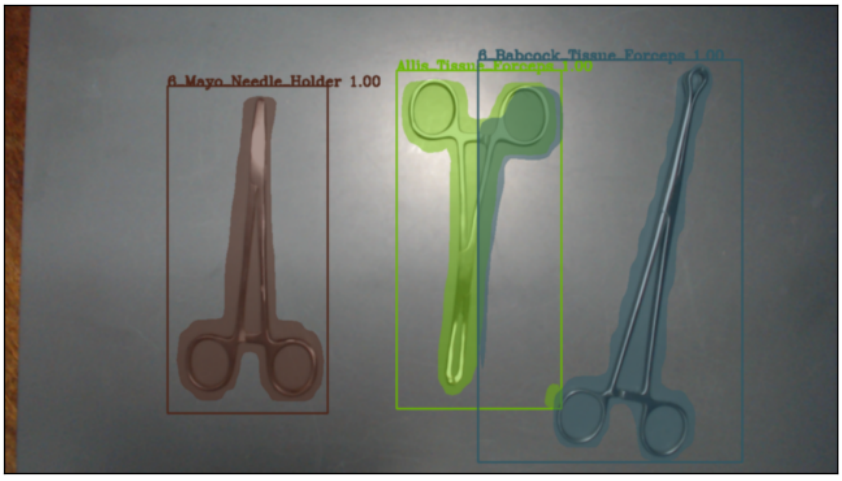


FIGURE 3.12: Infra-Red Image Results with Multiple Tools

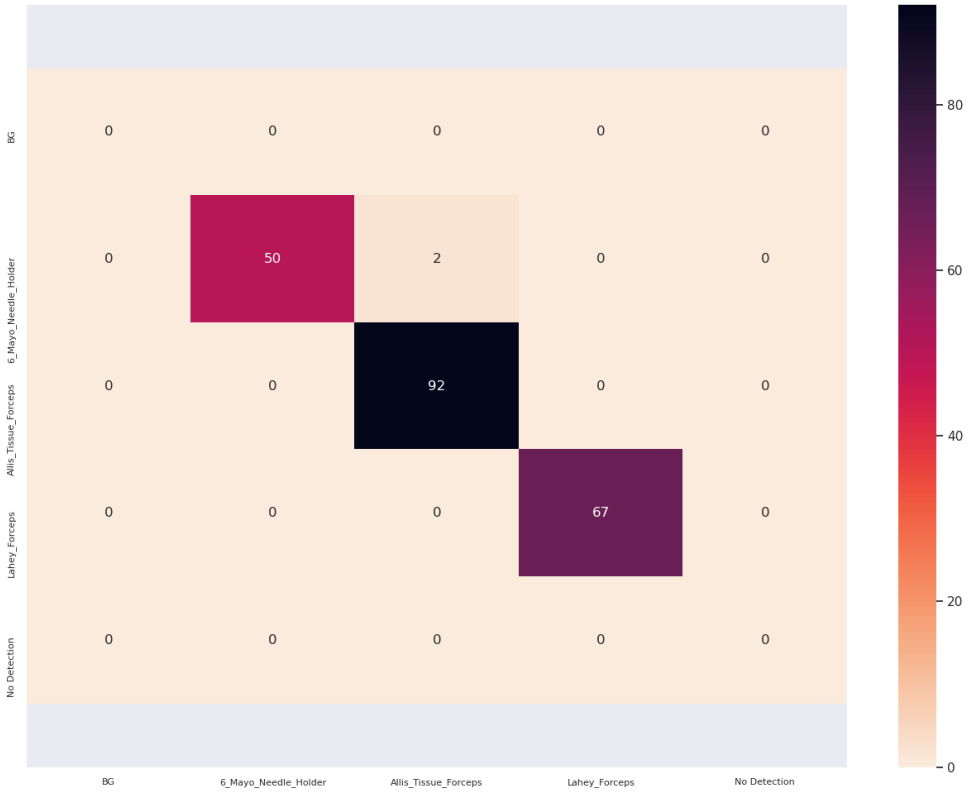


FIGURE 3.13: Infra-Red Confusion Matrix for Small Set of Tools

The results demonstrated the usefulness of using infra red imaging for addressing illumination and reflection problems. This is important in real world condition where illumination sources vary greatly across different regions of a hospital, however much more work has to be conducted in this area and the feasibility of using infrared images in real world conditions needs to be evaluated.

3.5 Discussion

This chapter addressed the following research question:

RQ1 – How can CNNs be trained for recognition of surgical tools while addressing volume, variety, complexity, adaptive self-learning and illumination / reflection / occlusion issues?

The work demonstrated that state of the art frameworks can provide good performance on a range of tools that offer volume and variety, but that there are significant issues with light conditions and backgrounds that need to be addressed. Further, while the system correctly identifies tools and provides basic information to the end-user, a solution that can provide much more information to the end user was needed. The research therefore focused on developing new methods to address this issue, and the work is reported in the next chapters of this thesis.

Chapter 4

Interpretable Deep Learning for Surgical Tool Management

The work in this chapter was presented at the 4th International Workshop on Interpretability of Machine Intelligence in Medical Image Computing (iMIMIC 2021), Springer, Cham., and published in Lecture Notes in Computer Science, vol 12929, DOI — 10.1007/978-3-030-87444-51.

The work that was reported earlier in this thesis (Chapters 2 and 3) highlighted significant challenges in surgical tool management but also provided possible solutions. A particular task was to develop a hierarchical classification strategy that could provide useful and relevant information to an end user, and the work in this section therefore developed a novel convolutional neural network framework for multi-level classification of surgical tools. The framework was designed to enhance the interpretability of the overall predictions by providing a more informative set of classifications for each tool. This allows users to make rational decisions about whether to trust the model based on multiple pieces of information, and the predictions can be evaluated against each other for consistency and error-checking. This is particularly important in hospitals, and could potentially reduce errors and increase efficiencies.

4.1 Introduction

Surgical tool and tray management is recognized as a difficult issue in hospitals worldwide. Stockert and Langerman (2014) observed 49 surgical procedures involving over two-hundred surgery instrument trays, and discovered missing, incorrect or broken instruments in 40 trays, or in 20% of the sets. Guedon et al. (2016) found equipment issues in 16% of surgical procedures; 40% was due to unavailability of a specific surgical tool when needed. Zhu et al. (2019) estimated that 44% of packaging errors in surgical trays at a Chinese hospital were caused by packing the wrong instrument, even by experienced operators. This is significant given the volumes; for example, just one US medical institution processed over one-hundred-thousand surgical trays and 2.5 million instruments annually (Stockert and Langerman, 2014).

There are tens of thousands of different surgical tools, with new tools constantly being introduced. Each tool differs in shape, size and complexity — often in very minor, subtle, and difficult to discern ways, as shown in Fig.4.1. Surgical sets, which can contain 200 surgical tools, are currently assembled manually (Mhlaba et al., 2015) but this is a difficult task even for experienced packing technicians. Given that surgical tool availability is a mission-critical task, vital to the smooth functioning of a surgery, ensuring that the tool is identified accurately is extremely important.

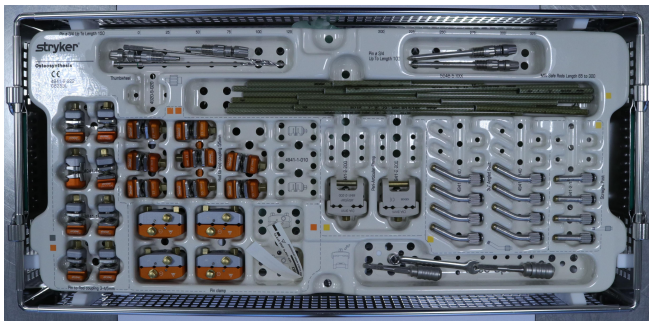


FIGURE 4.1: Surgical tools - Hoffman Compact instruments and implants

Al Hajj, Lamard, Conze, et al. (2019) reviewed convolutional neural network (CNN) architectures and a range of imaging modalities, applications, tasks, algorithms and detection pipelines used for surgical segmentation. They pointed out that hand crafted and hand engineered features had also been used for this task, and Bouget et al. (2017) reviewed predominant features used for object-specific learning with surgical tools, and listed colour, texture, gradient and shape as being important for detection and classification. Yang, Zhao, and Hu (2020) presented a review of the literature regarding image-based laparoscopic tool detection and tracking using CNNs, including a discussion of available datasets and CNN-based detection and tracking methods. While CNNs can therefore provide viable solutions for surgical tool management, understanding how the CNN makes a prediction is important for building trust and confidence in the system.

Interpretibility of predictions is then a critical issue — Rudin et al. (2021) stated that interpretable machine learning is about models that are understood by humans, and interpretability can be achieved via separation of information as it traverses through the CNN models. Zhang et al. (2020) developed an interpretable model that provided explicit knowledge representations in the convolutional layers (conv-layers) to explain the patterns that the model used for predictions. Linking middle-layer CNN features with semantic concepts for predictions provided interpretation for the CNN output (Zhou et al., 2015; Simon and Rodner, 2015; Zhang et al., 2019). How mid-level features of a CNN represent specific features of surgical tools and how they can provide hierarchical predictions is the focus of our work. CNNs learn different features of images at different layers, with higher layers extracting more discriminative features (Zeiler and Fergus, 2014). By associating feature maps at different CNN levels to levels in a hierarchical tree, a CNN model could incorporate knowledge of hierarchical categories for better classification accuracy. The model developed by Ferreira et al. (2018) addressed predictions across five categorisation levels: gender, family, category, sub-category and attribute. The levels constituted a hierarchical structure, which was incorporated in the model for better predictions. The benefit of this hierarchical and interpretable approach for surgical tool management is that end users can then make rational, well reasoned decision on whether they can trust the information presented to them (Rudin et al., 2021).

Wang, Ramanan, and Hebert (2017) discussed an approach to fine tuning that used wider or deeper layers of a network, and demonstrated that this significantly outperformed the traditional approaches which used pre-trained weights for fine-tuning. Going deeper was accomplished by constructing new top or adaptation layers, thereby permitting novel compositions without needing modifications to the pre-trained layers for a new task. Shermin et al. (2019) showed that increasing

TABLE 4.1: Surgical Datasets

Characteristic	CATARACTS	Cholec80	Surgical Tools
Size or Instances	50 videos	80 Videos	18300 images
Database Focus	Cataract Surgeries	Cholecystectomy Surgeries	Orthopaedics and General Surgery
Type of Surgery	Open Surgery	Laparoscopic	Open Surgery
Default Task	Detection	Detection	Classification
Type of Item	Videos	Videos	RGB Images
Number of Classes	21	7	361
Images Background	Tissue	Tissue	Flat colours
Image Acquisition Platform / Device	Toshiba 180I camera and MediCap USB200 recorder	Not Specified	Canon D-80 Camera and Logitech 922 Pro Stream Webcam
Image Illumination	Microscope Illumination	Fibre-optic in-cavity	Natural Light, LED, Fluorescent
Distance to Object	V.Close - Microscope	Close - in-cavity	30-cms to 60-cms
Annotations	Binary	Bounding Boxes	Multiple level
Dataset Organisation	500,000 frames each in Training and Test Sets	86,304 & 98,194 frames in Training and Test Set	14,640 images in Training and 3,660 in Validation set
Structure	Flat	Flat	Hierarchical
Image Resolution	1920x1080 pixels	Not Specified	600 x 400 pixels

network depth beyond pre-trained layers improved results for fine-grained and coarse classification tasks. We build on these approaches in our multi-level predictor.

4.2 Surgical Tool Dataset Overview

Kohli, Summers, and Geis (2017) and Maier-Hein et al. (2020) discussed the problems faced by the machine learning community stemming from a lack of data for medical image evaluation, which significantly impairs research in this area. There is just not enough high quality, well annotated data, representative of the particular surgery — a shortfall that needs to be addressed. Most medical datasets are one-off solutions for specific research projects, with limited coverage and restricted in numbers of images or data points (Maier-Hein et al., 2020). To address this, we plan to create and curate a surgical tool dataset with tens of thousands of tool images across all surgical specialities with high quality annotations and reliable ground-truth information. Since surgery is organised along specialities, each with its own categories, a hierarchical classification of surgical tools would be extremely valuable. We therefore developed our initial surgical dataset with a hierarchical structure based on the surgical speciality, pack, set and tool. We captured RGB images of surgical tools using a DSLR camera and a webcam and tried to provide consideration to achieving viewpoint invariant object detection with different backgrounds, illumination, pose, occlusion and intra-class variations captured in the images. We focused on two specialities – Orthopaedics and General Surgery — of the 14 specialities reported by the American College of Surgeons (ACS, 2021). The former offers a wide range of instruments and implants, while the latter covers the most common surgical tools. We propose to

TABLE 4.2: Surgery Knowledge Base (Excerpt)

Speciality	Pack	Set	Tool
Orthopaedics	VA Clavicle Plating Set	LCP Clavicle Plates	Clavicle Plate 3.5 8 Hole Right
Orthopaedics	Trimed Wrist Fixation System	Fixation Fragment Specific	Dorsal Buttress Pin 26mm
General Surgery	Cutting & Dissecting	Scissors	9 Metzenbaum Scissors
General Surgery	Clamping & Occluding	Forceps	6 Babcock Tissue Forceps

add the other specialities in a phased manner, and will make the dataset publicly available to facilitate research in this area.

CNNs have been successfully used for the detection, segmentation and recognition of surgical tools (LeCun, Bengio, and Hinton, 2015). However, the datasets currently available for surgical tool detection present very small instrument sets; to illustrate this, the Cholec80, EndoVis 2017 and m2cai16-tool datasets have seven instruments, the CATARACTS dataset has 21 instruments, the NeuroID dataset has eight instruments and the LapGyn4 Tool Dataset has three instruments (Al Hajj, Lamard, Conze, et al., 2019; Twinanda, Shehata, Mutter, et al., 2017). While designing CNNs to recognise seven or eight instruments for research purposes may be justifiable, this is nowhere nearly adequate enough for real work conditions. Any model trained using this data is unlikely to be usable anywhere else, not even in the same hospital six months later. We needed to develop a new dataset for our work as these surgical tool datasets did not offer a sufficiently large variety or number of tools for analysis, nor were they arranged hierarchically. A comparison of our dataset with CATARACTS (Al Hajj, Lamard, Conze, et al., 2019) and Cholec80 (Twinanda, Shehata, Mutter, et al., 2017), two important publicly available datasets, is presented in Table 4.1.

4.2.1 Surgery Knowledge Base

Setti (2018) points out that most public benchmark datasets only provide images and label annotations, but providing additional prior knowledge can boost performance of CNNs. To complement the dataset, we developed a more comprehensive surgery knowledge-base (Table 4.2) as an attribute-matrix which makes rich information available to the training regime. This proved to be a convenient and useful data structure that captures rich information of class attributes — or the nameable properties of classes — and makes it readily available for computational reasoning (Lampert, Nickisch, and Harmeling, 2014). We developed the knowledge representation structure for 18,300 images to provide rich, multi-level and comprehensive information about each image. The attribute matrix data structure proved to be easy to work with, simple to change and update, and it also provided computational efficiencies.

4.3 Experimental Method

We implemented our project in Tensorflow v-2.4.1 and Keras v-2.4.3. Our architecture consists of a ResNet50V2 network (He, Zhang, Ren, et al., 2016) which we trained on the Surgical Tool training dataset, by replacing the top layer with a dropout and dense layer with 361 outputs. We initially did not use the knowledge base annotations, only

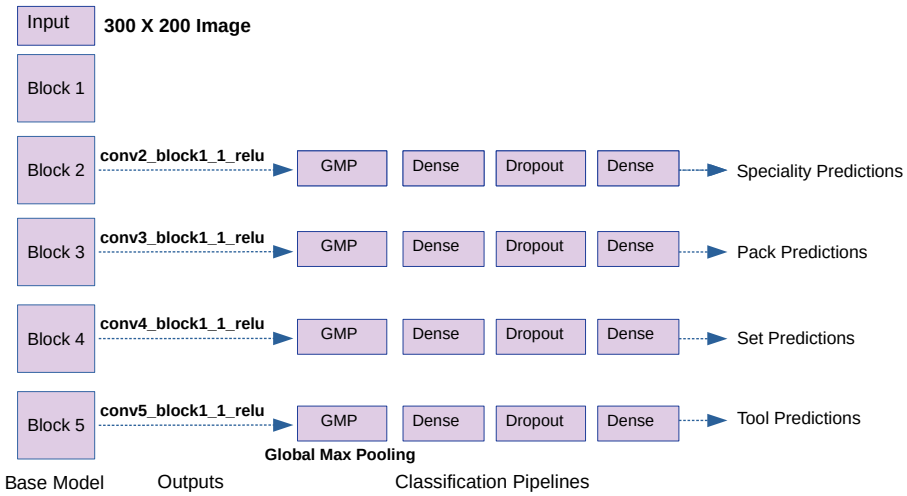


FIGURE 4.2: Resnet50V2 Architecture with Multiple Outputs

TABLE 4.3: Results - Val accuracy with output at different layers

All Outputs at:	Total Parameters	Parameters Trained	Speciality	Pack	Set	Tool
Conv2_block1_1_relu	700,570	686,490	0.956	0.356	0.258	0.091
Conv3_block1_1_relu	1,210,266	948,634	0.989	0.621	0.507	0.231
Conv4_block1_1_relu	3,060,634	1,472,922	0.997	0.927	0.851	0.663
Conv5_block1_1_relu	11,625,370	2,521,498	0.999	0.975	0.945	0.890

the tool labels and trained with the configuration in Table 4.4 with early stopping on validation categorical accuracy. We were able to obtain good predictions from this model with accuracy score at 93.51%, but only at the tool level. We then used this pre-trained architecture with surgical tool weights as our base model, froze the base model, and added separate classification pipelines, one for each prediction of interest - speciality, set, pack and tool (See Fig. 4.2). We relied on the knowledge base annotations which provided data for two specialities, twelve packs, thirty-five sets and 361 possible tools, and used it to create data-frames for the training and validation data. Each image was associated with the relevant annotations for each output, in the form of columns of text values or categorical variables representing the multiple classes for each output. This multi-task framework effectively shared knowledge of the different attribute categories for each image or visual representation. We developed a custom data handler for the training data (x_set) and for the labels for each of the four outputs (y_cat, y_pack, y_set, y_tool), and used one hot encoding to represent the categorical variables in our model. We then implemented training and validation data generators based on our custom data handler to provide batches of data to the model. Our model was compiled with one input (image) and four outputs.

We tested outputs at different layers to evaluate the impact of changing the depth of the network, with the results in Table 4.3. In each experiment, parameters available and actually trained were controlled by adjusting the numbers of layers. An operation within a block in ResNet50V2 consisted of applying convolution, batch normalisation and activation to an input; we obtained our outputs after the first operation in each

TABLE 4.4: Training Configuration

Parameter	Optimiser	Learning Rate	Batch Size	Activation	Loss	Metric
Value	Adam	0.001	64	Softmax	Categorical Crossen-tropy	Categorical Accuracy

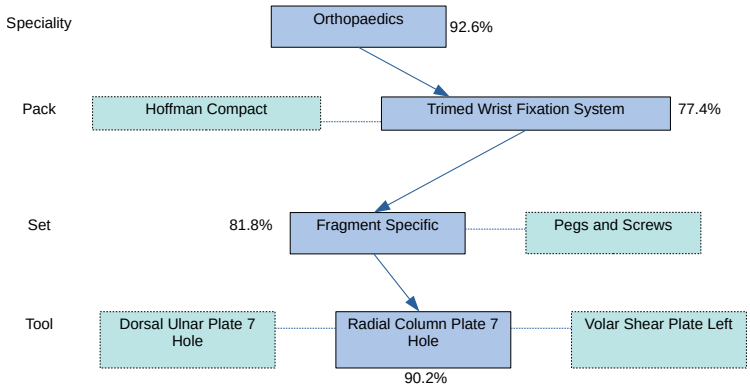


FIGURE 4.3: Interpretable multi-level predictions

block. These outputs were fed to external global max pooling and dense layers. A dropout layer regulated training – we replaced this with a batch normalisation layer but results did not improve. Since this was a multi-class problem, a dense layer with softmax activation was used for the final classification of each prediction, customised to the relevant number of classes. As we expected, better results were obtained by including more layers and by training more parameters – best results were obtained by including all layers up to Block 5. However, it is noteworthy that high accuracy was obtained for specific predictions even early in the model – for example, predictions for speciality were at 95.6% by block 2, for pack and set were at 92.7% and 85.10% at block 4 and for tool at 89% at block 5. Clearly it was possible to obtain accurate predictions for higher level categories using early layers of the model. This is explored further with the objective of improving interpretability for the end user, while reducing the total number of parameters that needed to be trained in the model.

The training set images from the surgery dataset and annotations from the knowledge base were used for training, with real time training data augmentation – including horizontal flip, random contrast and random brightness operations. We used the configuration in Table 4.4, the initial learning rate of 0.001 was decreased to 0.0001 at epoch 45 and to 0.00005 at epoch 75. A dropout rate of 0.2 was imposed. We implemented early stopping on val loss with a patience of 20 epochs. The total parameters in the model were 10,511,258, and parameters trained were 1,407,386 in each of the experiments.

1. ImageNet Training: For an initial baseline experiment, we used a ResNet50V2 model with ImageNet weights and four separate classification outputs were trained, one for each hierarchy – speciality, set, pack and tool.

TABLE 4.5: Architecture Results - Macro score or average for all classes

Level	Metric	ImageNet	Surgical-Tools	Surgical-Tools Depth Adjusted
Speciality	Accuracy score	0.90	0.94	0.94
	Hamming Loss	0.10	0.06	0.06
	f1 Score	0.73	0.84	0.83
	Precision score	0.93	0.95	0.95
	Recall score	0.96	0.99	0.99
Pack	Accuracy score	0.41	0.63	0.77
	Hamming Loss	0.59	0.37	0.23
	f1 Score	0.25	0.53	0.73
	Precision score	0.43	0.67	0.76
	Recall score	0.30	0.55	0.73
Set	Accuracy score	0.31	0.84	0.89
	Hamming Loss	0.69	0.16	0.11
	f1 Score	0.24	0.79	0.84
	Precision score	0.36	0.82	0.85
	Recall score	0.25	0.80	0.87
Tool	Accuracy score	0.20	0.90	0.90
	Hamming Loss	0.80	0.10	0.10
	f1 Score	0.16	0.86	0.86
	Precision score	0.78	0.91	0.91
	Recall score	0.27	0.91	0.90

2. Surgical Tool Training: We used the pre-trained base model with surgical tool weights, and trained the model with its four classification pipelines using the configuration as in Table 4.4 and architecture as in Fig. 4.2.

3. Depth Adjusted Surgical Tool Training: We used the pre-trained model with surgical tool weights as before, but changed the levels within the blocks of the ResNet-50V2 model from which we obtained outputs, thereby adjusting the depth of training. The outputs from Block 5 and 2 were obtained from conv"x"_block1_1, and from Block 3 and 4 were from conv"x"_block4_2. We did this to evaluate the effects of changing depths on the prediction accuracy; this was a minor change within the block but the total number of parameters trained were maintained the same.

4.4 Results and Conclusions

Our results, on a separate test subset of data, are shown in Table 4.5. The test data was images that the model had not seen before, as a sample of 400 random images across all classes had been reserved for testing. Training with ImageNet weights did not provide good results, but the use of surgical tool weights demonstrated that the model had captured relevant information about the dataset and was able to provide good predictions at multiple levels. In this architecture, by extracting multiple predictions along layers from coarse to fine as data traverses the CNN, early layers provided predictions corresponding to specialities while later layers provide finer predictions,

such as tool classifications (Fig. 4.3). It was easy for the CNN to distinguish between our two speciality classes, since General Surgery tools are visually different from orthopaedic tools – as we add more specialities where the visual distinction is not so clear, we may need to train at deeper levels. As the classes increased to 12, 35 and 361 for pack, set and tool respectively, predictions from deeper layers were needed. These hierarchical predictions are expected to provide better interpretability since multiple predictions can be tested and evaluated against each other for consistency or error by the end user. Adjusting the depths of layers used as outputs for predictions improved the results, even within the same block, demonstrating that more features are learned as the data travels through the CNN layers.

We developed a CNN framework that successfully utilised the hierarchical nature of surgical tool classes to provide a comprehensive set of classifications for each tool. This framework was deployed and tested on a new surgical tool dataset and knowledge base. The multi-level prediction system provides a good solution for classification of other types of medical images, if they are hierarchically organised with a large number of classes.

Chapter 5

OctopusNet: Machine Learning for Intelligent Management of Surgical Tools

The work presented in this chapter was published in the journal — Smart Health, Volume 23, DOI 10.1016/j.smhl.2021.100244.

The work conducted and presented on the development of a hierarchical network for hierarchical classification of surgical tools (Chapter 4) was well received at the 4th International Workshop on Interpretability of Machine Intelligence in Medical Image Computing (iMIMIC 2021). Feedback from the conference participants led to a new architecture — which was termed OctopusNet. This network is designed to purposefully share information across prediction hierarchies to improve classification accuracy. The network, as before, provides a degree of interpretability by predicting a set of features for each tool based on multiple classification targets. Important contributions of this work are not just the OctopusNet architecture, but also a novel surgical tool dataset and surgery knowledge base.

5.1 Introduction

A major New Zealand Hospital estimated that lost or misplaced surgical tools resulted in costs of US$350,000 annually. Surgical tools, which include instruments, implants and screws, are organised in sets of related tools and packs of related sets grouped by surgical speciality. Counting these tools pre- and post-surgery resulted in savings of over US$17,000 per month but this practice was too difficult to sustain manually due to the additional staffing requirements and the tedious nature of the work involved in counting each contaminated tool. Surgical set assembly accuracy was around thirty percent, and low packing accuracy required multiple expensive sets to be held in inventory. Sixty packing staff were rostered to work twenty-four hours a day, seven days a week on manual packing of tools and sets. Challenges included high inventory levels, high set assembly errors, lost or misplaced instruments, inconsistent availability of surgical instruments, and non-functional instruments in a tray. Large volumes and varieties of surgical instruments, implants and screws (Figs. 5.1 and 5.2) also posed a formidable challenge (Unit Manager, personal communication, Nov. 2019).

Current technology for surgical instrument recognition and tracking used in commercial settings include bar codes, RFID, colour markers, etching, acoustic tracking, electro-magnetic, ultrasound and nano technology initiatives (Al Hajj, Lamard, Conze, et al., 2019). These solutions offer some advantages but generally only track

instruments and not implants; they also require modification or additions to the tool to be tracked. Computer vision technology does not need any modifications to be made to the instruments or implants, and this is critical both for the sterile status of the instrument and given the fact that these implants are placed inside the human body. This is also a significantly cheaper solution, and is simpler to implement. A marker-less and non-contact solution as provided by computer vision and deep learning would be ideal for accurate and real-time recognition of surgical tools in a hospital.

Surgical tool detection and recognition through computer vision and machine learning has numerous practical applications, and can be invaluable in reducing incidents of lost tools, improving packing accuracy, reducing errors, lowering costs, and providing overall efficiencies. Other applications for surgical tool recognition include not just instrument tracking in hospital inventory management (Ahmadi et al., 2018), but also robotic and computer-assisted surgery (Sarikaya, Corso, and Guru, 2017), instrument position recognition in minimal invasive surgery (Zhao et al., 2017), and pose recognition in surgical training (Leppanen et al., 2018). Current research focuses on the development of algorithms based on, and tested with, small medical datasets involving the actual detection of around 21 types of tools (Al Hajj, Lamard, Conze, et al., 2019). However, there are many thousands of surgical instrument types in circulation and these datasets are not representative of the problem nor realistic for real world conditions. For example, one manufacturer reported that their product line consists of 19,000 surgical instruments, with new types and classes continually being introduced (Sklar, 2016). A new approach is required to handle this volume and variety of surgical tools, as well as to cope with new surgical tools as they are introduced. There are 14 specialities (Table 5.1) – reported by the American College of Surgeons (ACS, 2021). Each speciality has multiple procedures, and the variety and complexity of the tools in each procedure offers a significant challenge. However, the hierarchical nature of surgical tool organisation – grouped by speciality and procedure type, for example – also presents an opportunity to improve tool recognition by the incorporation of prior information about hierarchies and structures of tool classes in the model.

The research question that we address is: **Can we design a convolutional neural network that improves recognition of surgical tools by effectively utilising the hierarchical nature of surgical tool classes?** This is a departure from standard image classification tasks where classes are assumed to be flat and independent of each other; here we use problem-specific knowledge to devise a richer organisation of the classes, and use this rich information to improve the architecture and performance of our CNN. We create a new surgical tool dataset and surgery knowledge base to train the CNN, making it potentially useful for intelligent management of surgical tools in a hospital. Our functional requirement for the CNN is that it provides rich predictions at multiple levels; in the case of surgical tools, predictions for the specific speciality, pack, set and tool levels are required.

5.2 Intelligent Surgical Tool Management

Medical image challenges provide a platform for the development of cutting edge deep learning solutions in medical imaging, and Al Hajj, Lamard, Conze, et al. (2019) highlight the fact that more than 20 challenges were hosted in 2018. Many of the challenges specifically address surgical instrument detection, segmentation and recognition. In these image-based challenges, a specific task is defined, a dataset is

TABLE 5.1: Surgical Specialities

Cardiothoracic surgery	Colon and rectal surgery	General surgery	Gynaecology and obstetrics
Gynaecological oncology	Neurological surgery	Ophthalmic surgery	Oral and maxillofacial surgery
Orthopaedic surgery	Otorhinolaryngology	Paediatric surgery	Plastic & maxillofacial surgery
Urology	Vascular surgery		

FIGURE 5.1: Surgical Tools - Instruments

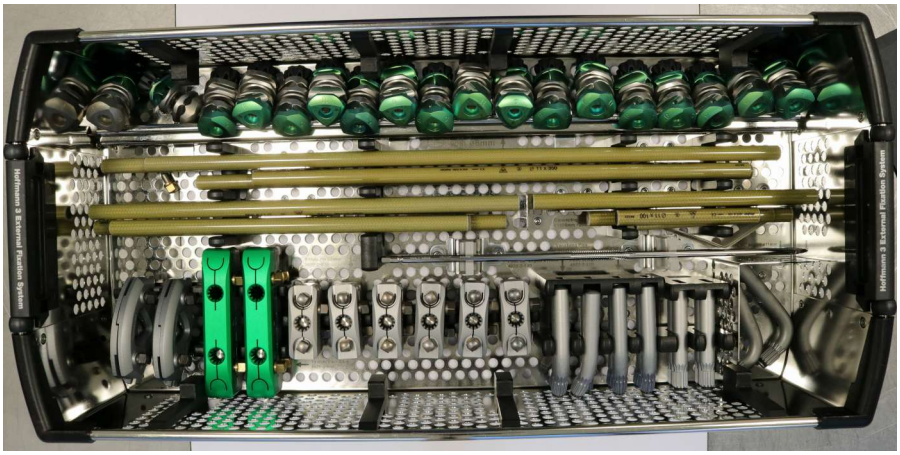


FIGURE 5.2: Surgical Tools - Implants

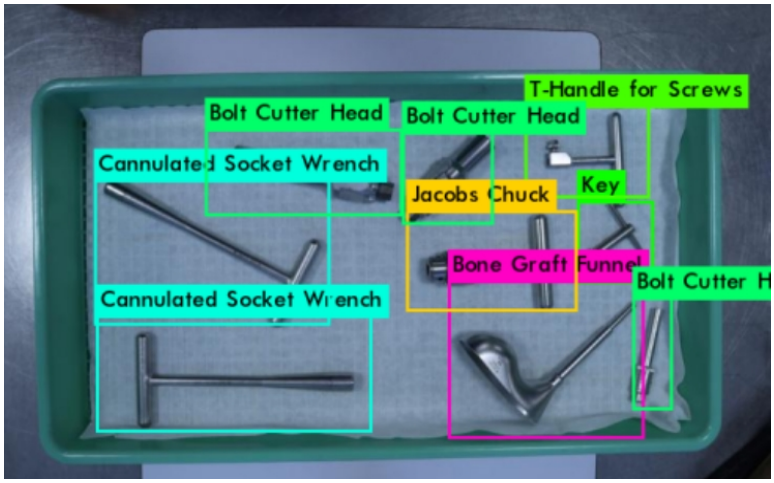


FIGURE 5.3: Identification of Surgical Tools

provided, evaluation procedures are defined, algorithms are developed and applied, and deep learning solutions are tested on a specifically held-out test dataset. One such dataset that is focused on surgical tools is the Cholec80 dataset (Twinanda, Shehata, Mutter, et al., 2017), which has been used to fine tune AlexNet, ResNet-50 and ResNet18 models, pre-trained on ImageNet data (Twinanda, Shehata, Mutter, et al., 2017; Alshirbaji, Jalal, and Moller, 2018; Mondal, Sathish, and Sheet, 2019; Nwoye et al., 2019; Vardazaryan et al., 2018). Wang et al. (2019) developed a deep neural network model, based on DenseNet-121 pre-trained from ImageNet, utilizing both spatial and temporal information from surgical videos for surgical tool presence detection. They used Graph Convolutional Networks (GCNs) and evaluated the model on two datasets: m2cai-tool and Cholec80. Sahu et al. (2020) also used two datasets – Cholec80 and EndoVis15 – to successfully test their Endo-Sim2Real method for instrument segmentation. Using the m2cai16-tool dataset, a subset of Cholec80, researchers have fine-tuned GoogleNet, VGG-16, AlexNet, Inception-v3, YOLO9000 and DenseNet201 models for surgical tool identification (Jo et al., 2019; Lin et al., 2019; Raju, Wang, and Huang, 2016; Sahu et al., 2016; Zia, Castro, and Essa, 2016). Jin et al. (2018) extended the m2cai16-tool dataset by providing labels for 2,532 frames with bounding boxes around the tools, and made a new m2cai16-tool-locations dataset available. They used this dataset to successfully train a Faster R-CNN and VGG-16 model.

Other surgical tool datasets have also been relied upon for research into surgical tool recognition. The CATARACT dataset (Al Hajj, Lamard, Conze, et al., 2019) was used in the EndoVis 2017 Challenge, and a specific aspect of the challenge addressed surgical tools. There were 27 submissions from 14 teams that provided solutions to the challenge; networks relied upon in the solutions included VGG-16, Inception-v3 and v4, SqueezeNet, DenseNets, ResNets, NASNet-A and Inception-ResNet-v2. Almost all the networks were trained on ImageNet (Al Hajj, Lamard, Conze, et al., 2019). Transfer learning techniques, where a CNN model pretrained on general images can be fine-tuned on a surgical instrument database, have therefore been used to achieve state-of-the-art performance for surgical instrument recognition tasks (LeCun, Bengio, and Hinton, 2015).

While most of the recent work on surgical tool analysis has relied on CNNs and deep learning, Al Hajj, Lamard, Conze, et al. (2019) showed that other machine learning approaches using hand crafted and hand engineered features had also been used for this task. However, CNNs are now the predominant approach for computer vision based object recognition and detection, and have been successfully used for the detection, segmentation and recognition of objects and regions in images over the last two decades. CNNs have been deployed for surgical tools detection (LeCun, Bengio, and Hinton, 2015); for example, Yang, Zhao, and Hu (2020) reviewed the use of CNNs for laparoscopic surgery tool detection and tracking. Al Hajj, Lamard, Conze, et al. (2019) also reviewed the CNN architectures used for surgical tool segmentation. However, there are significant problems with the CNN-based approach given the sheer volume and complexity of surgical tools that require to be managed. In many cases, CNNs have been trained on datasets with a relatively small number of classes, such as ImageNet which has 1000 classes, and this is not sufficient for real world tasks where tens of thousands of classes have to be recognised (Liu et al., 2020b). This is particularly true in surgical tool management where the CNNs used in surgical tool classification applications have dealt with very small instrument sets and the currently available benchmark tool datasets offer only 7 to 21 instruments for research (Al Hajj, Lamard, Conze, et al., 2019; Twinanda, Shehata, Mutter, et al., 2017). This highlights the fact that there are significant problems with surgical tool management that have not been addressed in the literature. These problems are detailed and discussed below:

1. Volume and Variety of Tools – Each hospital deploys many thousands of surgical tools, and new tools are constantly being developed and introduced. These tools represent a wide range in terms of shape and size (Figs. 5.1 and 5.2), making accurate surgical tool recognition a complicated and complex task. The fact that these tools are clustered into specific packs and sets with a wide variety of tool types makes the task even harder. Managing this volume and variety is a significant challenge, and limited research has been conducted for large scale, comprehensive and intelligent CNN-based surgical tool management in a hospital.

2. Complexity in Tool Management – Surgical sets, which can contain 200 surgical tools, are currently assembled manually (Mhlaba et al., 2015). This requires technicians to have in-depth knowledge and experience, but packed sets are often found to be incomplete and this can put surgical procedures and patients at risk. Since new tools are continually being introduced, technicians need to constantly train and up-skill to ensure high accuracy in tool management. Managing this complexity with CNNs is a task that has not yet been addressed.

3. Broken surgical instruments – Even if surgical tools can be accurately recognised, it is imperative that broken or damaged tools are immediately flagged. Damaged instruments – for example, needle holders with cracked hinges or jaws – are a potentially serious threat in the operating theatre. They can spread infections but can also fall into open surgical cavities, leading to complications and extended anaesthesia times. Surgeons and nurses cannot physically inspect each surgical instrument since they are working under time pressure, but a CNN based system that captures instrument images and which can indicate possible damage can provide an important functionality.

4. Mission-Critical System – This is a mission-critical system, vital to the functioning of a hospital since management of surgical tools is critical to the surgical

procedure. This issue has been handled by over-stocking inventory and maintaining redundant stocks, but a CNN based system can be a core component of a more efficient solution.

To address the issues enumerated above, accurate and robust identification and classification of a wide range of surgical tools is a critical task. To ensure this, more work needs to be done on surgical tool recognition. In particular, the hierarchical nature of surgical tools have not been adequately explored or utilised in CNNs used for surgical instrument detection, segmentation and recognition. This is a particular focus of our work since it has been demonstrated in prior research that incorporating knowledge of structure and hierarchy in a CNN can lead to improved object recognition and classification accuracy. This is discussed in the next section.

5.3 CNNs and Hierarchical Classification

While training a CNN to recognise surgical tools is relatively straightforward if enough training data is available (Fig. 5.3), it has been shown that classification performance of CNNs could be improved if structural knowledge of the set of classes was incorporated in the model (Srivastava and Salakhutdinov, 2013). In the case of surgical tool recognition, this would be about providing details of which surgical speciality a specific tool belonged to, and additional information of its parent pack and instrument set. Leveraging class structure and hierarchies for improved classification performance was accomplished by Srivastava and Salakhutdinov (2013) via the sharing of knowledge from relevant classes and by focusing on learning only the distinctive class-specific features for each class. Koo, Klabjan, and Utken (2018) exploited the hierarchical structure of classes by embedding deep CNNs into a category hierarchy, but multiple CNNs needed to be trained to classify each class accurately and this was costly and time consuming. Yan, Zhang, Piramuthu, et al. (2015) relied on a CNN for hierarchical representations of images, and recurrent neural network (RNN) or sequence-to-sequence models to evaluate a hierarchical tree of relevant classes.

Why does the provision of hierarchical information lead to better CNN performance? Zeiler and Fergus (2014) pointed out that CNNs learn different features of images at different layers, with higher layers extracting more discriminative features. It was then possible to associate feature maps to different levels in the hierarchical tree. CNN models could then be integrated with knowledge of hierarchical categories and relationships for better classification accuracy. Motamedi et al. (2020) developed a CNN architecture, which they termed "Octopus", in which classification of objects was based on the features extracted from the deepest layer. Since deep layers extract the highest distinctiveness or class-specificity, their Octopus configuration placed kernels of deep layers in class-specific, discriminative, distinctive components, while shallow layer kernels were used to jointly understand all classes. The parameters for each class were kept in distinct, class-specific, dense and loosely-connected components. They then developed an Octopus-based neural network that relied on the Inception architecture as the baseline model, coupled with 100 class-specific branches to classify 100 image classes.

Branch Convolutional Neural Networks (B-CNN) have been developed and used by Hu et al. (2016) and Zhu and Bain (2017); these models relied on branch networks for hierarchical fine-grained visual classifications. The number of prediction branches in the model matched the number of hierarchies in the label tree. A branch-based CNN architecture design would provide multiple coarse to fine predictions along

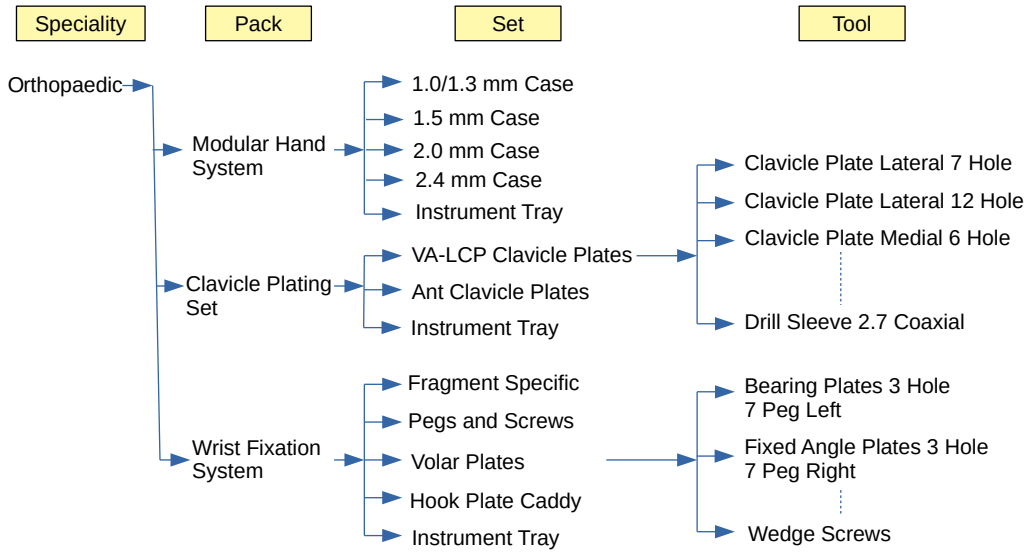
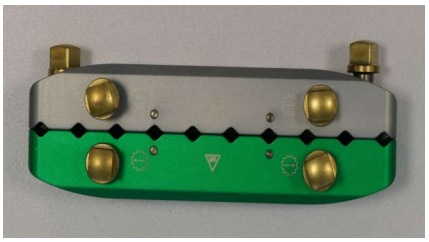


FIGURE 5.4: Sample fragment of the hierarchy in our dataset

layers which correspond to the hierarchical structure of the classes. The classification was achieved in a hierarchical way instead of using the normal flat structure adopted by most datasets and CNNs. This strategy was based on the insight that lower layers capture low level, basic shape features and higher layers extract higher level and more detailed features (Hu et al., 2016; Zhu and Bain, 2017).

The model of Wang et al. (2018) incorporated semantic label relationships to generate better classification results – label activations were propagated bidirectionally and asynchronously to achieve final classification results. This bidirectional structured inference model relied on top-down and bottom-up propagation of activations via a message passing network, and aggregated these messages into final activations for label prediction. A similar approach was used by Inoue, Forster, and Santos (2020) in a unified deep model that provided predictions at different levels in a hierarchy tree, thereby incorporating the label hierarchical structure in the model. This architecture consisted of a Bidirectional Inference Neural Network (BINN), combined with an RNN algorithm that integrated structured information for prediction within the model. The architecture captured intra- and inter-level label relations through two parameters, one addressed two-way label relations between levels, and the other addressed relations within each level. The model developed by Ferreira et al. (2018) addressed predictions across five categorisation levels. In this model, the family, category, and sub-category levels were defined to be mutually exclusive, while attributes could be common across levels. The levels constituted a hierarchical structure, which was incorporated in the model for better classification and predictions.

Semantic hierarchies were used by Inoue, Forster, and Santos (2020) to improve fine-level classification, while also capturing hierarchical information from coarser level outputs. Three variations of hierarchical CNNs were explored – built using regular convolutional and dense layers – with multiple outputs for each hierarchical level. In addition to the B-CNN model, a variant called "Concat-net" was developed. In this model, the last dense layer of each hierarchical level was concatenated with the last dense layer of the previous branch. The concatenations were designed to directly share relevant features and to improve overall hierarchical classification. A third model, "Add-net", was a variation of the "Concat-net" model, but the values of



Surgical Speciality	packType	setType	toolName	toolType	toolShape	toolFeatures
orthopaedics	Hoffman-3 External Fixation System	Hoffman-3 Clamps and Couplings	10 Hole Pin Clamp	Pin clamp	Rectangle metal block	10 hole
Additional Features	**Manufacturer**	**Owner**	**Location**	**Modality**	**Lighting**	**Camera**
Center gear hole	Stryker, USA	XXX Hospital	XXX Dept	RGB	tungsten	Canon D-80 DSLR

FIGURE 5.5: Sample fragment of annotations in our knowledge base

the last dense layers were added instead of concatenating to share information across hierarchical levels. The hierarchical model resulted in an improvement on image classification tasks, with "Add-net" obtaining the best results (Inoue, Forster, and Santos, 2020).

While these approaches are promising, there are some issues that need to be addressed. For example, some of the strategies require different models to be trained for classification at different branches, which is computationally expensive and memory intensive. For example, to predict surgical tools using this approach, one model would have to be trained to recognise specialities, another to recognise packs, a third to recognise sets and a fourth to recognise tools; each model has to be loaded into memory and deployed. Other approaches result in very large models and a significant increase in features, for example by the concatenation of layer outputs. However, there are some approaches that address our functional requirements (Ferreira et al., 2018; Hu et al., 2016; Zhu and Bain, 2017), and we rely on this hierarchical CNN research work as we develop our own solution for the classification of surgical tools.

5.4 Methodology

While significant amount of work has been conducted in this area, Bouget et al. (2017) stated that much more data needed to be made available for algorithm development, and the lack of quality data was a significant handicap for research. Other researchers have also raised concerns about the lack of quality data for research and algorithm development, and have called for the release of more surgical tool datasets into the research community so that better models can be generated (Twinanda, Shehata, Mutter, et al., 2017). These concerns are addressed in our work.

5.4.1 Surgery 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. Given that Sklar, a surgical instrument manufacturer that claimed to offer the largest product line

of instruments, reported that their product line consists of 19,000 surgical instruments (Sklar, 2016), datasets with a wider range of classes are required for research. As part of our work, an important task that we undertook was to develop a comprehensive dataset of surgical tools based on specialities, with a hierarchical structure – speciality, pack, set and tool, as shown in Fig. 5.4. We created an initial dataset of surgical instrument images: over forty-thousand images of surgical tools were captured under different lighting conditions and with different backgrounds. We captured RGB images of surgical tools using a DSLR camera on site in a major hospital under realistic conditions and with the surgical tools currently in use. Image backgrounds in our initial dataset were essentially flat colours, even though different colour backgrounds were used. Discussions with potential end-users highlighted the fact that many more images would have to be included as we further developed our dataset, including much greater occlusions, illumination changes, and the presence of blood, tissue and smoke in the images, which would be more representative of crowded, messy, real-world conditions.

Illumination sources included natural light – direct sunlight and shaded light – LED, halogen and fluorescent lighting, and this accurately reflected the illumination working conditions within the hospital. Distances of the surgical tools to the camera to the object ranged from 60 to 150 cm., and the average class size was 74 images. Images captured included individual object images as well as cluttered, clustered and occluded objects. Our initial focus was on Orthopaedics and General Surgery, two out of the 14 surgical specialities (ACS, 2021) as discussed earlier. We selected these specialities since general surgery instruments are the most commonly used tools across all surgeries and provide instrument volume, while orthopaedics provides variety and complexity given the wide range of procedures, instruments and implants used in orthopaedic surgery. We will add other specialities as we develop this dataset, to reflect the complexities inherent in each of the surgical specialities. In this initial dataset, we covered 2 specialities, 12 packs, 35 sets and 361 different tool types or classes. This dataset was designed to offer a large variety of tools, arranged hierarchically to reflect how surgical tools are organised in real-world conditions.

5.4.2 Surgery Knowledge Base

Marcus (2020) highlights a need for implementation of models with hybrid architecture, rich prior knowledge, and sophisticated reasoning techniques. An important task for us was to make rich information or knowledge accessible for domain inference (Marcus, 2020; Garcez et al., 2019). To achieve this in the surgery domain, we developed a comprehensive surgery knowledge-base to permit better classification, deductive inference and semantic interpretation (Hoehndorf and Queralt-Rosinach, 2017). We set up this surgery knowledge-base as an attribute matrix (Fig. 5.5) which proved to be a convenient and useful data structure capturing rich information readily available for computational reasoning. In addition to only providing the image labels, our potential end-users indicated that additional information about each tool could be useful under real world use conditions, including information as to which speciality the particular tool belonged to, and the pack and set details. We annotated 18,238 individual instruments from the dataset, and the data that we provided for each image includes class information at four levels: speciality, pack, set and tool. We also provide other information that can potentially be useful as we develop our model and architecture further. Such information can be used for the predictions of finer attributes of each tool and to assist in classification of unknown or new classes,

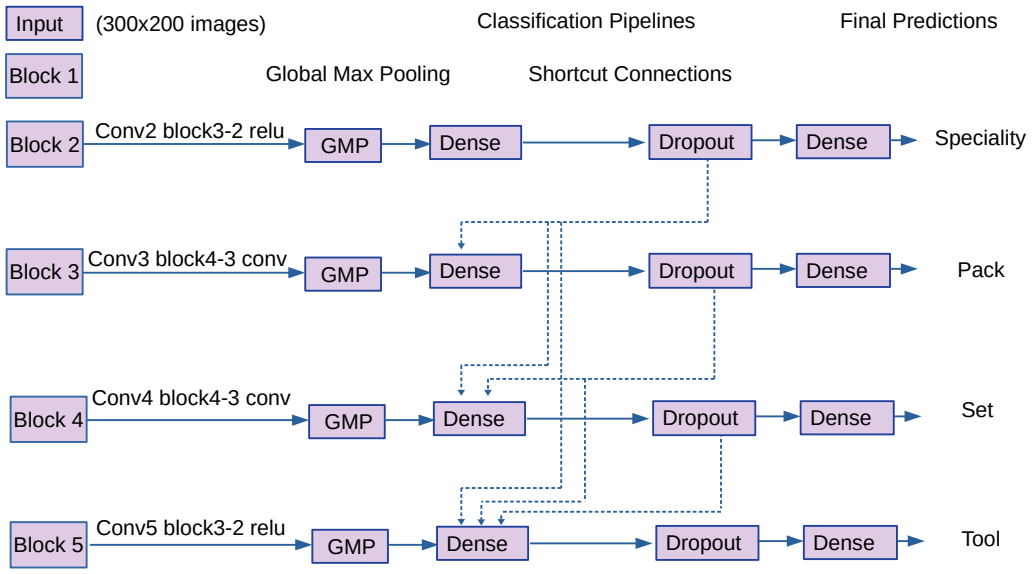


FIGURE 5.6: Architecture with Forward Connections

an issue that is significant given the fact that new tools and classes are continually being introduced in the hospitals.

5.4.3 OctopusNet Architecture

Our knowledge representation annotations were in the form of categorical variables, where text values represented the multiple classes, and these variables needed to be encoded for use in our model. We first created category objects by converting each column in the data-frame to a category. We then used label encoding to replace each of the categorical values with a numeric value between 0 and the number of classes minus 1, and used one hot encoding to represent the categorical variables as binary vectors.

We implemented a custom data generator for the data handling, since our objective was to build a single model that was capable of predicting four distinct outputs for one image input. This data handler was designed to generate batches of data that were provided to our multi-output model; each image fed to the model was also accompanied by its class labels at four levels. A custom method was used to obtain a given batch of data with the training and multi output label data; it was called with the batch size and real time image augmentation was used for the training phase. We then used train and validation data generators based on our data handler to provide batches of data to the model.

Our architecture consisted of a ResNet-50V2 network as the base model around which the rest of the architecture was built. We added separate classification pipelines to the base network, one for each prediction of interest – speciality, set, pack and tool – to create the OctopusNet architecture. The benefit of this architecture was that separate predictions were generated at each level, allowing for better interpretation of the results. Our prototype system was trained on the training images of the surgery dataset and knowledge base, which captured 2 specialities, 12 packs, 35 sets and 361 possible tools. With the ResNet-50V2 CNN as the base model, classification pipelines were created via dense layers with ReLU activation, which acted as message propagation or shortcut connections (He, Zhang, Ren, et al., 2016). These were direct

TABLE 5.2: Training Configuration

Optimiser	Learning Rate	Epochs	Activation	Loss	Metric
Adam	0.001	120	Softmax	Categorical Crossentropy	Categorical Accuracy

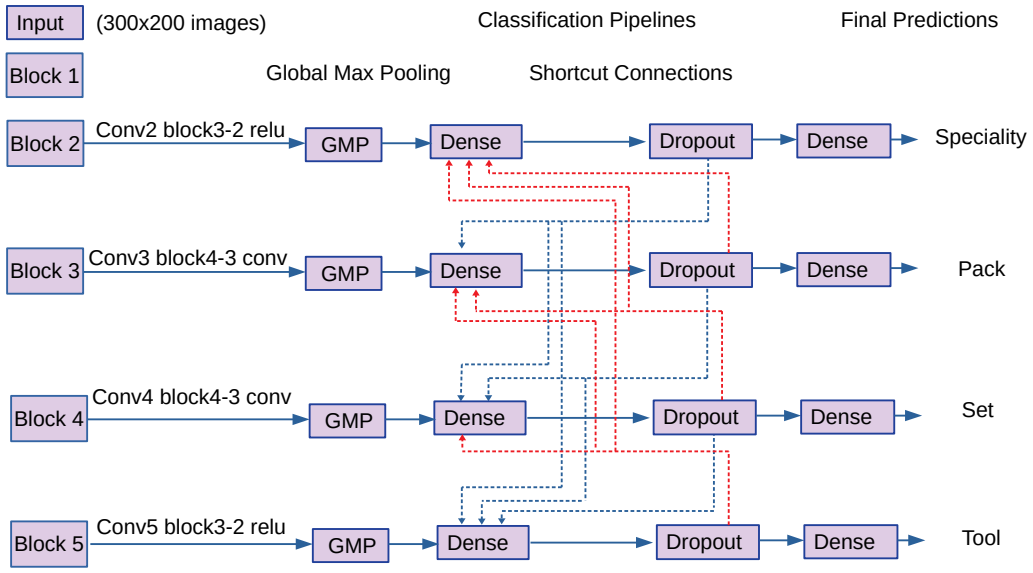


FIGURE 5.7: OctopusNet with Full Connections

path connections that propagated information across levels – for example, speciality latent information is propagated towards the pack, set and tool levels. The outputs of each pipeline were provided to a final dense layer for the classification at each level. Our architecture was designed to permit information transfer between each classification pipeline, since it captures both intra-level and inter-level relations. The speciality level influences the pack level, both speciality and pack level influences the set level and all three higher levels influence the tool level for classification.

We froze the pre-trained ResNet-50V2 base network from the input layer up to the start of Block 5 while training the other layers. We trained the network for 120 epochs and implemented early stopping on validation loss. The model was trained on 14,680 annotated training images, and validated on 3558 images. The surgery knowledge base annotations were relied upon in the training, with their detailed and multi-level manual annotations. A range of augmentations were carried out, including horizontal flip, random contrast and random brightness. We used the training schedule as in Table 5.2, with a batch size of 64 images. We implemented a learning rate schedule with the low initial learning rate; the rate was decreased to 0.0001 at epoch 75 and to 0.00005 at epoch 95. A dropout rate of 0.2 was imposed. We used a separate set of images for testing and the model did not see the test images in training or in validation. Our architecture and training decisions are based on experiments and on previous work (Ferreira et al., 2018; Wang et al., 2018; Inoue, Forster, and Santos, 2020), and we conducted the following three experiments:

1. Surgical Tools: We modified the ResNet-50V2 model by removing the classification block and adding a custom classification block with a dropout and dense layer with 361 outputs. We trained this modified model on the training image

set of the Surgical Tool dataset. We used tool labels and the training configuration as in Table 5.2, and implemented early stopping on validation data categorical accuracy. We then used this model as the base model, and added the classification pipelines to the base. We froze the model up to the start of Block 5, and trained it on the training data. This is a simple approach that provided the benchmark model against which our other experimental architectures were evaluated.

2. Forward Connected: In this architecture, we used the Surgical Tool weights and the multi-level classification architecture as discussed above, but modified the model to include shortcut connections that provided direct paths to propagate information in the top down direction in the label hierarchy. Each level was connected in the forward (top-down) direction to other levels – speciality level predictions provided information to pack, set and tool level; pack levels provided information to set and tool level; set level predictions were passed on to the tool level. Outputs were provided to a dense layer for final classification. The architecture is depicted in Fig. 5.6. Layer-wise addition was used for the connections; a list of tensors were provided as input and the layer-wise addition returned a single tensor of the same shape. By using this functionality, we ensured that the trainable parameters remained at 2,587,034 parameters in all experiments, so that we could accurately estimate the effects of our information propagating connections.

3. Fully Connected: We used the Surgical Tool weights and made direct paths for propagating information through connections in both the top down and bottom up direction in the dataset hierarchy. In addition to the connections in the forward direction (top down) as in the previous experiment, the tool level provided information in the reverse direction to the set, pack and speciality levels; the set level provided information to the pack and speciality level, and the pack level provided information to the speciality level. The outputs were fed through a dense layer to obtain the final prediction. This architecture is depicted in Fig. 5.7 and is the full OctopusNet model. The trainable parameters were maintained at 2,587,034.

5.5 Results and Conclusions

The results in Table 5.3 show that, despite training the same number of parameters in all three cases, the use of shortcut or message passing connections improved overall accuracy. The most significant improvement was in the mid-level predictions, or at the pack and set level. The information provided from the predictions at the category and tool levels greatly improved mid-level predictions, leading to better overall accuracy for the model. This highlighted the importance of using addition layers and connections that added the weights of each layer. The outputs from the connections were added to the other outputs, and these shortcut connections did not add extra parameters or complexity in terms of extra computations (He, Zhang, Ren, et al., 2016). These connections used element-wise summation and an iterative estimation procedure where features were refined through the various layers of the network. Summation of weights can be viewed as a residual correction or delta to the input; this resulted in successively refinement of the feature maps. The ResNet-50V2 Base Block had already learnt a rough estimate of the representation by the prior training on the surgical tool dataset, which was then iteratively refined by the successive

TABLE 5.3: OctopusNet Results - Macro score or average reported for all classes

Level	Metric	Surgical-Tool (ST) Weights	ST Forward Connected (Fig. 5.6)	ST Fully Connected (Fig. 5.7)
Category	Accuracy score	0.99	0.98	1.00
	Hamming Loss	0.01	0.02	0.00
	F1 score	0.98	0.96	1.00
	Precision score	1.00	0.99	1.00
	Recall score	0.99	0.99	1.00
Pack	Accuracy score	0.85	0.92	0.98
	Hamming Loss	0.15	0.08	0.02
	F1 score	0.81	0.87	0.97
	Precision score	0.85	0.88	0.97
	Recall score	0.80	0.86	0.97
Set	Accuracy score	0.86	0.91	0.96
	Hamming Loss	0.14	0.09	0.04
	F1 score	0.79	0.89	0.96
	Precision score	0.79	0.90	0.96
	Recall score	0.81	0.89	0.96
Tool	Accuracy score	0.92	0.93	0.94
	Hamming Loss	0.08	0.07	0.06
	F1 score	0.86	0.88	0.89
	Precision score	0.92	0.93	0.93
	Recall score	0.91	0.93	0.93

layers. These layers cooperate to compute a single level of representation, layers preserved feature identity and iteratively refined and reinforced their estimates of the same features (He, Zhang, Ren, et al., 2016).

Interpretation is potentially enhanced since the technician or end user can be provided with multiple predictions, at different levels. These predictions can be evaluated against each other, and can be used by the end-user to reason about the reliability of the model. The information can also assist in error checking and correction, and to evaluate possible alternatives to the prediction presented. This is important in real world conditions where fine grained tool distinctions may give rise to errors. The provision of rich and relevant information to the users greatly helps in achieving a clear understanding of each particular tool and its position in the surgical dataset hierarchy.

5.5.1 Summary

We addressed the research question by designing a CNN that successfully utilised the hierarchical nature of surgical tool classes to make improved predictions. This was deployed and tested on a new surgical tool dataset and knowledge base. The OctopusNet based system provides a good solution for classification of medical images, if they are hierarchically organised with a large number of classes, with the following benefits:

1. The OctopusNet architecture provides predictions at multiple levels for interpretability and better understanding of the results.

2. We create a powerful knowledge representation data structure that can be extended and modified easily.

3. The system provides for easy deployment in medical settings, and this model is adaptable to different medical images based classification tasks with a hierarchical structure.

The solution is a proof of concept for accurate recognition of surgical tools by utilising the hierarchical nature of the classes, but this is very much a work in progress. Much more work has to be done to achieve the aim of intelligent management of surgical tools in a hospital, thereby reducing incidents of lost tools and packing errors, lowering costs, increasing patient safety and system efficiencies. Maier-Hein et al. (2020) discussed the lack of machine learning 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. The authors recommended creating and providing access to larger, representative and annotated datasets, something that could lead to improved outcomes and success stories in the application of machine learning to surgery. We seek to address this issue in our work, and have set up both a preliminary surgical dataset and a knowledge base data structure with our OctopusNet CNN. These assets will be further developed with an intention of making them publicly available to facilitate research in this area.

5.5.2 Future Work

Our work addressed some of the issues with surgical tool management that we highlighted earlier in this paper, such as managing the volume, variety and complexity of surgical tools. We plan to add to our preliminary dataset, and will comprehensively

capture images across all the surgical specialities. More images with occlusions, reflections, illumination changes, the presence of blood, tissue and smoke will be included in the dataset and also images with different modalities, such as infrared and depth images. Open surgery and laparoscopic surgery images will be sourced, including live surgeries. The attribute matrix will be expanded and better defined, by drawing on medical and surgical tool expertise to clarify terminology and naming conventions. Future work also includes testing the system in practical settings; for example, in the identification and tracking of surgical tool usage during surgery, in the management of misplaced tools, and in the accurate learning of new and unknown tools. This surgical tool dataset and knowledge base can potentially become an important resource for innovative research that successfully addresses the mission critical nature of surgical tool management in a hospital.

Chapter 6

Evaluation of Deep Learning Techniques on a Novel Hierarchical Surgical Tool Dataset

The work in this chapter was presented at the 2021 Australasian Joint Conference on Artificial Intelligence – where it was awarded 2nd Place in the Best Applied Paper Category.

Given significant interest in the surgical tool dataset from the research community, this chapter reports efforts to improve and evaluate the usefulness of the dataset. In addition to 360 surgical tool classes, the dataset was designed with a four level hierarchical structure defined by 2 specialities, 12 packs and 35 sets. To evaluate the performance of this dataset with Deep Learning techniques, the work conducted employed different convolutional neural network training strategies to evaluate image classification and retrieval performance, including the utilisation of prior information in the form of a taxonomic hierarchy tree structure. The work evaluated the effects of image size and the number of images per class on model predictive performance, to see if the dataset could be improved. Experiments with the mapping of image features and class embeddings in semantic space using measures of semantic similarity between classes show that providing prior information results in a significant improvement in image retrieval performance on the dataset, demonstrating the usefulness of the chosen structure. The dataset was then made freely available for public research in this area.

6.1 Introduction

Surgical tool management in hospitals is a difficult, time consuming and costly task; lost, misplaced or unavailable surgical tools were estimated to cost just one New Zealand hospital over NZ$500,000 annually (Unit Manager, personal communication, Nov. 2019). Challenges faced in management of these tools included high inventory levels, multiple surgical tool set assembly errors, high staffing requirements, high costs, inconsistent availability of surgical tools, and non-functional or broken instruments being presented at surgery. Large volumes and varieties of surgical tools (Fig. 6.1) also pose a formidable challenge for management. According to Stockert and Langerman (2014), just one institution can process over 100,000 surgical trays and 2.6 million tools every year. With an average of 38 surgical instruments present per tray, and six trays deployed on average per surgery (Mhlaba et al., 2015), managing this volume and complexity manually under mission-critical conditions is a challenging task. Surgical tool detection and recognition through artificial intelligence (AI) and

FIGURE 6.1: Surgical Set and Tool Examples

TABLE 6.1: Current Tool Datasets

Characteristic	CATARACTS (Al Hajj, Lamard, Conze, et al., 2019)	Cholec80 (Twinanda, Shehata, Mutter, et al., 2017)	EndoVis2017 (Allan et al., 2019)	ROBUST-MIS19 (Ross, Reinke, and Full, 2019)
Size	50 videos	80 Videos	10 Videos	30 Videos
Focus	Cataract Surgeries	Cholecystectomy Surgeries	Abdominal (Porcine)	Varied Surgeries
Use Case	Detection	Detection	Segmentation	Detection
Classes	21	7	7	2
Annotations	Binary	Bounding Boxes	Masks	Masks
Structure	Flat	Flat	Flat	Flat

machine learning systems can provide a solution that can reduce incidents of lost or misplaced tools, improve packing accuracy, reduce errors, lower costs, and improve overall efficiencies within hospitals. Surgical tool recognition can be used in AI based hospital inventory management systems, and also in robotic and computer-assisted surgery, instrument position recognition, and in surgical monitoring, audit and training (Sarikaya, Corso, and Guru, 2017; Zhao et al., 2017; Leppanen et al., 2018).

Maier-Hein et al. (2020) discussed the lack of success stories in the application of machine learning to surgery, and contrasted it to success in other medical fields, such as radiology and dermatology. This was directly attributed to the lack of quality annotated data, representative of the surgery domain, and the small size and limited representation of currently available datasets were reported to be major problems. One available labelled surgical tool dataset, while useful, provides images of only four tools (Lavado, 2018). Similarly, the currently available surgical tool datasets with a larger number of tools do not offer a sufficiently large range nor are they arranged hierarchically (Table 6.1). Kohli, Summers, and Geis (2017) highlighted the lack of data for medical image evaluation with machine learning, and described

current research as being "data starved" in this area. Current research focuses on convolutional neural networks (CNNs) trained on small medical datasets and the actual detection of less than fifty types of tools (Al Hajj, Lamard, Conze, et al., 2019); however, there are many thousands of surgical instrument types in circulation (Sklar, 2016). Clearly a new approach is required to handle this volume and variety of surgical tools. To help in addressing these challenges, we created a new surgical tool dataset named **HOSPITools**, short for "**H**ierarchically **O**rganised **S**urgical **P**rocedure **I**nstruments and **Tools**". This dataset offers a wide range of tools, and we evaluate its performance with different deep learning methods and techniques.

6.2 Class Hierarchies and Training Strategies

Image features learned by CNNs have been used extensively to classify images, or to retrieve images that are visually similar to a query image (Barz and Denzler, 2019). While deep CNNs are extremely effective in object classification and recognition, classification of fine-grained classes and discrimination between classes with relatively minor differences is a challenge (Setti, 2018). This is a significant problem for our work, since many surgical tools are visually similar and often differ in minor, subtle and hard to discern ways. An approach that can potentially improve classification or retrieval performance for such fine grained classes is to embed prior knowledge of the classes or class hierarchies into the model (Deng, Berg, and Fei-Fei, 2011). Class hierarchies share knowledge of relationships in the ground truth class label arrangements, as opposed to class labels in a flattened arrangement where every class is assumed independent and unrelated, and incorporating this information into the model can potentially lead to better classification and retrieval performance.

The main challenge, as highlighted by Narayana et al. (2019), lies in mapping images and labels to a shared latent space where embeddings that correspond to a similar semantic (not just visual) concepts lie closer to each other than embeddings corresponding to different semantic concepts. They addressed this problem by first constructing a semantic embedding space based on prior domain knowledge and then projecting image embeddings onto this fixed semantic embedding space. Their model ensured that distance between image embeddings were similar to corresponding class embedding distances in the semantic embedding space (Narayana et al., 2019). Barz and Denzler (2019) computed class embeddings by a deterministic algorithm based on prior domain knowledge encoded in a hierarchy of classes – this was a novel feature level approach that mapped image embeddings to semantic embeddings, and successfully incorporated class information and semantic relationships into a deep learning model. The semantic embeddings of image features were shown to result in a model that was much more invariant against superficial visual differences such as colour and shape (Barz and Denzler, 2019), and we therefore experiment with this method for our project.

The most common loss function used in the training of CNNs is the categorical cross-entropy loss in conjunction with a softmax activation, also known as the softmax loss (Barz and Denzler, 2019; Wen et al., 2016).

$$\mathcal{L}_{CCE} = -\sum_{i=1}^{k} c_i \log\left(\hat{c}_i\right) \tag{6.1}$$

In Equation 6.1, $\hat{c}_i$ represents the probability score for class c_i. This training strategy separates the classes, but it may not be sufficient for fine grained classification tasks

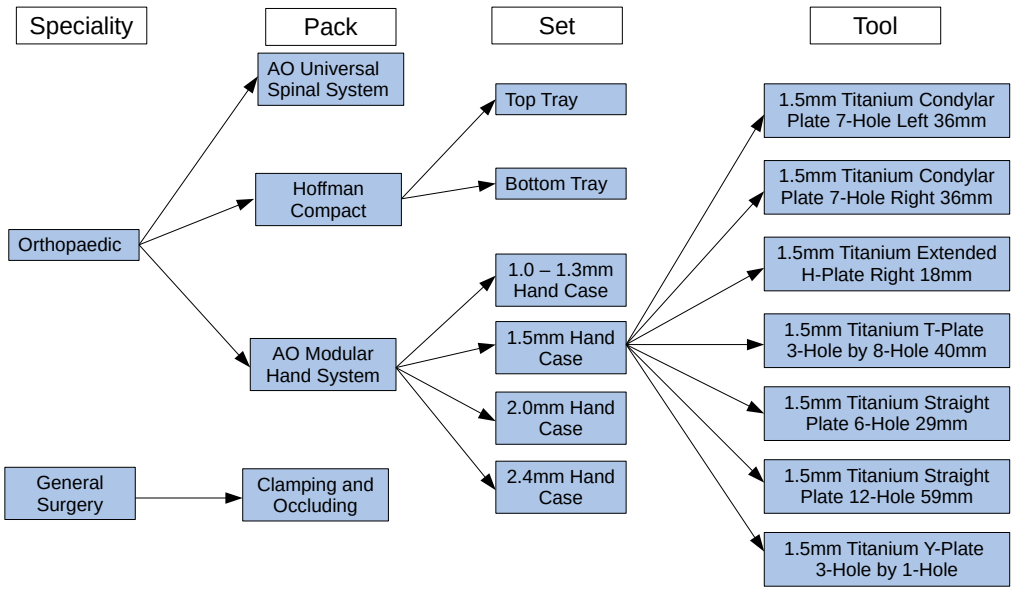


FIGURE 6.2: Surgical Tool Dataset Structure

(Barz and Denzler, 2019). The center-loss was therefore designed to increase the separation of classes while minimizing the distances between samples from the same class, and was defined as (Wen et al., 2016):

$$\mathcal{L}_{center-loss} = \frac{1}{2}\sum_{i=1}^{k} \left\| \boldsymbol{x}_i - \boldsymbol{c}_{y_i} \right\|_2^2 \tag{6.2}$$

In Equation 6.2, $\boldsymbol{x}_i$ represents the center of the i$^{\text{th}}$ class and $\boldsymbol{c}_{y_i}$ the deep feature vectors for each class. A multiple loss training strategy was used where the center-loss was employed to pull the deep features of the same class to their centers, while the softmax loss forced the deep features of different classes apart (Wen et al., 2016). A combination of losses was also employed by Barz and Denzler (2019), who used a classification loss along with an embedding loss designed to maximise the cosine similarity or the inner product between the image features and the embeddings of their classes. This correlation or cosine loss function was defined as:

$$\mathcal{L}_{CORR} = \frac{1}{k}\sum_{i=1}^{k} \left(1 - \psi\left(I_i\right)^{\top} \varphi\left(c_{y_i}\right)\right) \tag{6.3}$$

In Equation 6.3, φ defined the class embedding function, ψ the embedding function for image I, and $^{\top}$ referred to matrix multiplication using the transpose of the embeddings, equivalent to the inner product or the dot product of the embeddings. Another important distance based loss is the mean squared error (MSE) loss, defined for class c_i as:

$$\mathcal{L}_{MSE} = \frac{1}{k}\sum_{i=1}^{k} \left(c_i - \hat{c}_i\right)^2 \tag{6.4}$$

We evaluate our dataset with these training strategies and loss functions.

6.3 Methodology

In this section, we describe the HOSPITools dataset, and we experiment with different strategies to train CNNs using this dataset. We believe that this dataset can be an important resource for AI and machine learning research on surgical tool management, and we use our experience with CNN training strategies to try to improve its structure and organisation.

6.3.1 Surgery Dataset

We developed our surgical dataset based on an hierarchical structure – speciality, pack, set and tool – as shown in Fig. 6.2. We captured RGB images of surgical tools using a DSLR camera, and manually arranged the images hierarchically in the dataset. We took these pictures on site in a major hospital, with the surgical tools currently in use. Image backgrounds were essentially flat colours, even though different backgrounds were used. Illumination sources included natural light – direct sunlight and shaded light – LED, halogen and fluorescent lighting. Distances of the camera to the object ranged from 60 to 150 cms. We focused on two specialities – Orthopaedics and General Surgery – for the initial stages of development of the dataset. The former speciality offers a wide range of instruments, implants and screws, while the latter covers the most common instruments used across all open surgery. We propose to add images of tools used in all 14 surgical specialities reported by the American College of Surgeons (ACS, 2021) in a phased manner as we develop this dataset. Our initial dataset consisted of 15,522 images across all hierarchies, with 11,712 images in the training set and 2,810 images in the validation set. We reserved a further 1,000 images for the test set, which the models did not see during training. While the average class size was 74 images, the range was from 139 images to as low as 10 images. This allowed us to evaluate the performance of the CNN training strategies with low class frequencies, and to explore how the dataset could be optimally structured with minimum images per class required for good performance.

6.3.2 Surgery Hierarchy

While it was relatively straightforward to train a baseline classifier using only the images and labels, some of our other strategies required additional information to be provided to the model. We therefore created a four level hierarchy in the surgery tool dataset, which consisted of speciality (2 classes), pack (12 classes), set (35 classes) and tool (360 classes) levels. The hierarchy was detailed in an indented tree format, which we then converted into "child-parent" tuples, as discussed by Barz and Denzler (2019). Dictionaries mapping class labels to lists of parent class labels and to child class labels in the hierarchy were created, and also a dictionary mapping hypernym identities of each element (class) to depths in the tree. We also developed lists of node identities, commencing with the direct hypernym of the given element and ending with the root node.

We only considered the taxonomic or hierarchical relationship between our classes in our work. The easiest relation is the "is-a" relation, which allows the specification of a hierarchical structure (Barz and Denzler, 2019). Hierarchies, most commonly represented as tree structures, provided us with an effective tool to organise and present the relationships and prior knowledge in our classes. In our tree structure, each class or node has just one parent class and distance was defined in terms of the

length of the shortest path between two classes $\mathbf{c}_i, \mathbf{c}_j$. The dissimilarity of the classes $d_{\mathcal{G}}$ and the semantic similarity $s_{\mathcal{G}}$ was defined as (Barz and Denzler, 2019):

$$\begin{aligned} d_{\mathcal{G}} &= \frac{\text{height}(\text{LCS}(c_i, c_j))}{\text{height}\ (\mathcal{G})} \\ s_{\mathcal{G}}(c_i, c_j) &= 1 - d_{\mathcal{G}}(c_i, c_j) \end{aligned} \tag{6.5}$$

In Equation 6.5, LCS stands for lowest common subsumer – a class c_i was a subsumer of c_j if c_j was a descendant of c_i – and height $(\mathcal{G})$ is the height of the entire hierarchy. Using this, we obtained similarity measures in the range (0, 1), where "1" represented the maximum similarity (no distance) between classes. This information can then be used to train a CNN for image classification and retrieval (Brust and Denzler, 2019), as will be shown in the next section.

6.3.3 CNN Training Strategies

We used the well researched and widely used ResNet-50 (He, Zhang, Ren, et al., 2016) for all our experiments. We computed the channel mean and standard deviation of the images in the training set, and used it to normalise the data. We resized the original 6000×4000 pixel images to 150×100 pixels, and used multiple data augmentation techniques, including flipping, scale augmentation and random cropping, to add diversity to the training data (He, Zhang, Ren, et al., 2016). We evaluated the following experiments for our image classification and retrieval tasks, including hierarchy-based semantic image embeddings, based on prior work by Barz and Denzler (2019):

Baseline Classifier : As a baseline, we used a standard ResNet-50 and the features extracted from the layer before the final classification layer of the network architecture. We used categorical cross entropy as the loss function.

Center-loss : We used the ResNet-50 architecture and trained it with both center-loss and softmax loss, following Wen et al. (2016). We maintained the center-loss weight at 0.1 – this value was used to balance the two loss functions. Wen et al. (2016) experimented with changes of this weight from 0 to 0.1; with the weight at 0, or only using softmax loss, they obtained a poor result but performance was relatively unchanged across other variations of this weight.

MDS Embeddings : We computed embeddings in 360 dimensional space so that the distances of class embeddings corresponded to their semantic dissimilarity (Eq. 6.5) using classical multidimensional scaling (MDS). We used the MSE loss in this distance based approach.

Sphere Embeddings : We calculated a "360-by-360" matrix specifying the distance between each pair of classes, based on the dissimilarity score of the two classes (Eq. 6.5). Following Barz and Denzler (2019), with the first class at the origin, the second class was located at an offset along the first axis by the specified distance. We then placed all remaining classes in an iterative manner at an intersection of the hyperspheres centered at existing classes, with the radii set at the distance of the new class. We used the MSE loss in this training strategy.

Unitsphere Embeddings : The problem statement is: Given a distance matrix D, we wanted to place the set of points on a unit hypersphere which produce the same distance matrix. We used Eq. 6.5 to calculate similarities and the following equation to place class embeddings, where φ defined the class embedding function and $^\top$ referred to matrix multiplication using the transpose of the embeddings, equivalent to the inner product or the dot product of the embeddings (Barz and Denzler, 2019):

$$\begin{aligned} &\varphi\left(c_i\right)^\top \varphi\left(c_j\right) = s_{\mathcal{G}}\left(c_i, c_j\right) \\ &\left\|\varphi\left(c_i\right)\right\| = 1 \end{aligned} \tag{6.6}$$

Equation 6.6 stated that the correlation of class embeddings should equal their similarity. The second function ensured that the L2-norm embeddings were on the unit hypersphere, and the dot product was then used as a substitute for the Euclidean distance (Barz and Denzler, 2020). The network was trained to minimise the difference between image representations and the embeddings of their respective class as per the guidelines of Barz and Denzler (2019) using a combined loss $\mathcal{L}_{\text{CORR+CCE}} = \mathcal{L}_{\text{CORR}} + \lambda\mathcal{L}_{\text{CCE}}$. Since we desired that the embedding loss $\mathcal{L}_{\text{CORR}}$ dominated learning, we set λ to a very low value (0.1) in our experiments (a similar value was used in the center-loss strategy).

We tried two different learning schedules for our training, a standard ResNet training schedule and Stochastic Gradient Descent with Cosine Annealing and Warm Restart (SGDR) (Loshchilov and Hutter, 2017). While we tested these learning rates on each of our strategies, we only present the SGDR results since they are much better. This schedule implemented warm restarts, where in each restart the learning rate was initialized to a new value, scheduled to decrease over the cycle. The initial learning rate at the beginning of each cycle was 0.1, decreasing to a minimum of 10^{-6} using cosine annealing based on number of epochs since the last restart (Loshchilov and Hutter, 2017). The first cycle was set at 12 epochs, the multiplier for cycle length was set at 2, and training was for 5 cycles or 372 epochs (Barz and Denzler, 2019).

6.3.4 Metrics Reported

We report the Accuracy, Top-5 Accuracy, Hierarchical Accuracy and F1-Score for the classification performance. For the retrieval tasks, we report the hierarchical precision of the nearest neighbour search performed on different image embeddings – HP@k for different k values, Average Hierarchical Precision (AHP) and Mean Average Precision (mAP). The Hierarchical Precision at k (HP@k) is a generalization of Precision@k which takes class similarities into account (Deng, Berg, and Fei-Fei, 2011), and we report this for k at 250. This is calculated by the sum of similarities between query image class and retrieved image class over the top k retrieval results, divided by the maximum possible sum of top-k class similarities. Average Hierarchical Precision is defined by the area under the hierarchical precision curve, with the optimum normalized at 1.0. The Mean Average Precision, which does not consider class similarities, is also reported for comparison.

Class similarity is reported by the Wu-Palmer similarity metric ("WUP"), which considered the height and position of classes relative to each other in the tree – classes further from the root with a common parent tend to be more semantically similar. The WUP measure was calculated from equation 6.5, using the $s_{\mathcal{G}}(c_i, c_j)$ definition.

TABLE 6.2: Classification Results

SGDR	Accuracy	Top-5 Accuracy	Hierarchical Accuracy	F1-Score
Classifier	0.84	0.98	0.80	0.81
Center-loss	**0.88**	**1.00**	0.83	**0.86**
MDS	0.83	0.96	0.79	0.80
Spheres	0.85	0.98	0.81	0.82
Unitsphere	**0.88**	0.99	**0.84**	**0.86**

TABLE 6.3: Retrieval Results (WUP)

SGDR	HP@1	HP@10	HP@50	HP@100	AHP	mAP
Classifier	0.84	0.68	0.56	0.54	0.83	0.47
Center-loss	0.91	0.81	0.65	0.60	0.84	0.76
MDS	0.90	0.87	0.87	0.87	0.95	0.73
Spheres	0.90	0.88	0.89	0.89	0.97	0.76
Unitsphere	**0.93**	**0.91**	**0.91**	**0.91**	**0.98**	**0.84**

6.4 Experiments and Results

Classification performance is good across the board, and there is no significant improvement in basic accuracy by including hierarchical information, as shown in Table 6.2. However, the biggest impact of including prior information and in the embedding strategies is found in the retrieval task, as shown in Table 6.3. Retrieval of single images is good for all models tested, but as the number of similar images retrieved increases, there is a definite advantage in terms of the embedding strategies. There is a significant drop in accuracy with increase in the k value with the Classifier and Center-loss models, but embedding with the MDS, Spheres and Unitsphere strategies demonstrates a consistent performance across different k values. Since the number of images per class is low, smaller k values retrieve images from exactly the same category as the query but as k increases, images are retrieved from outside the direct class. This is where the incorporation of semantic information excels, retrieving images from semantically similar classes even at higher k values. Semantic information significantly improves the quality of content-based image retrieval, by retrieving images that are both visually and semantically similar. Incorporating prior knowledge about class similarities by mapping class embeddings in semantic space appears to facilitate better learning by the CNN, thereby leading to better retrieval results. Organising the surgical tool dataset in the form of a hierarchical structure, and providing additional information about the taxonomic or hierarchical relationship between our classes, is therefore conclusively demonstrated to be an approach that leads to better performance, at least for the image retrieval tasks.

6.4.1 Does Size Matter?

The original images were captured at 6000 by 4000 pixels, on the assumption that finer detail could be captured and it would be easier to down-sample the images than to up-sample. Down-sampling was done to improve data handling, storage and processing, and we evaluated the effects of resizing images in the pre-processing

TABLE 6.4: Class Frequency Classification Results

Images per Class	Class	F1-Score
139	Curved Mayo Scissors	0.96
117	7-inch Metzenbaum Scissors	0.94
15	0.76mm Drill Bit with 10mm Stop Mini QC	0.17
18	0.76mm Drill Bit with 12mm Stop Mini QC	0.22
13	Universal Spinal System Holding Sleeve	1
11	Jacobs Chuck	1

pipeline on the CNN performance. We experimented with images of 600 by 400 pixels, with 300 by 200 pixels, with resizing the images to 224 by 224 with padding, and with image size of 150 by 100 pixels. There was no degradation in performance even at the smaller sizes, and so we implemented our training at an image size of 150 by 100 pixels, with random cropping of 100 by 100 pixels during augmentation. Our findings can be contrasted with the work of Sabottke and Spiele (2020), who examined image resolution variations on CNN performance for radio-graphic images. While they did find some performance differences, this was relevant only when finer details needed to be captured for the diagnosis-specific tasks. For our objects of interest, image size variances do not appear to be as significant but this is a promising avenue for future work.

6.4.2 Class Frequencies

The class frequencies for the training set were averaged at 74 images, with a range from 10 to 139 images per class. While additional images were available, we wanted to test performance with different class frequencies. This was difficult to analyse – we obtained good classification results (Unitsphere strategy) even with 11 images per class, while much higher class frequencies did not yield the best results (Table 6.4). Clearly the number of images required for good performance depends on the particular tool and its distinctiveness in the dataset. An initial benchmark – at least for this dataset, for classification tasks, with the prior hierarchy information, and for these types of tools – does appear to be at least 40 images per class but this is not conclusive. As more cluttered images in realistic and messy settings are added, more images will be required to maintain accuracy and predictive performance. We will revisit this as we expand the scope and scale of our dataset.

6.5 Conclusions and Future Work

We developed a new surgical tool dataset – **HOSPITools** – and used it to test different CNN learning strategies. We demonstrated that the hierarchical nature of surgical tool classes could be used to make improved predictions. We also used the training to explore how the dataset should be structured and to evaluate some design parameters. This was a proof of concept for accurate recognition of surgical tools by utilising the hierarchical nature of the classes, and this solution can be used for intelligent management of surgical tools in a hospital.

We will continue to improve the dataset, with a view to making it publicly available for AI and machine learning research. We will address threats to the validity and utility of our work by adding images from more of the 14 surgical specialities, and by including greater coverage and variety in each speciality. We will include images with greater occlusions, reflections, illumination changes, the presence of blood, tissue and smoke, varied backgrounds, and from different modalities such as video, infrared and depth images. Open surgery and laparoscopic surgery images need to be sourced if possible, including live surgeries. If we can do this, then the surgery tool dataset can potentially be a valuable resource for the AI and machine learning communities.

Chapter 7

Making OctopusNet Robust

This chapter evaluates ways of making the OctopusNet better able to cope with changes in illumination and backgrounds, so that it can cope with real world conditions. The work conducted tests the CNN performance with challenging images, and find that performance degrades with changes in backgrounds and illumination. It therefore evaluates methods to make OctopusNet more robust to changes in illumination and backgrounds, develops and uses synthetic data and filter based purposeful augmentation, and achieves improved performance with this augmented dataset.

7.1 Introduction

Maron et al. (2021) highlighted the brittleness in performance of CNNs, stating that small changes in the input image had major adverse effects on the classification performance of the CNN. These image changes reflected the actual variations in images acquired under routine real world conditions, and this brittleness — along with a resultant lack of robustness and reliability — impeded the successful deployment of AI-based systems and tools into routine medical settings. Similarly, Dapello et al. (2020) stated that CNNs struggled to cope with imperceptibly small perturbations in images (adversarial attacks), and had difficultly in recognising objects in corrupted or noisy images. Adversarial training on explicitly crafted perturbed images to counter such attacks was expensive, and could lead to performance degradation. Techniques to address the brittleness in CNN performance were therefore developed by these researchers, which was evaluated in this chapter.

Maron et al. (2021) investigated data augmentation, test time augmentation and anti-aliased networks in search of a solution for this brittleness issue, and used artificial image transformations such as rotations, altered brightness or various zooms along with real images in their training and testing. Extreme forms of artificial data augmentation were used during the training stage with beneficial results. Tremblay et al. (2018) used non-artistically generated synthetic data to train CNNs, and demonstrated that such a trained model provided very good performance. This approach relied on domain randomization to generate synthetic data for training a neural network. The authors hypothesised that the crude images created by their technique was actually beneficial since it forced the CNN to focus on relevant details in the image. Huh et al. (2018) developed a synthetic data generation framework that created visual variations such as motion blur and occlusions in the images, and showed that including synthetic data with visual variations in the training dataset significantly improved real-time performance of object detectors. Jo, Na, and Song (2017) generated training data by synthesizing images of background and relevant objects, and added noise and variable illumination or brightness to the images of objects from different viewpoints. in total, 40 backgrounds and 36 relevant objects

were used to create 25,000 training images with an average of 17.36 images per background/object, and better results were obtained with this synthetic dataset as compared to the results obtained using a real training dataset with 13,000 images of similar objects. Hinterstoisser et al. (2019) created a purely synthetic dataset by using 3D background models to provide background images with realistic shapes and textures. The objects of interest were then rendered on top of these backgrounds. They demonstrated that using this synthetic dataset in training resulted in better performance on a challenging evaluation dataset when compared to CNN models trained with real data. Barros Barbosa et al. (2017) developed a large synthetic dataset by using photo-realistic human body generation software, and used it to train a CNN to recognise multiple discriminative structural attributes of human figures. Their method used synthetic images of human avatars as proxies for real human images. Similarly, Manettas, Nikolakis, and Alexopoulos (2021) developed a synthetic dataset of manufactured parts by using a range of simulation tools. Many synthetic datasets are currently available to train CNNs, including Flying Chairs, FlyingThings3D, SceneNet, SceneNet RGB-D, SYNTHIA and Virtual KITTI (Tremblay et al., 2018), however these are very specific in purpose and object coverage. An potentially useful initiative was to create purposeful and specific synthetic data for experiments, and to evaluate how such synthetic datasets could be used to make OctopusNet more robust to changes in illumination and background conditions.

7.2 Methods

Nowruzi et al. (2019) stated that the two general methods for synthetic data generation were real data augmentation and data generation through simulation; this research adopted the former approach of adding objects to existing frames or backgrounds to create our synthetic surgical tool dataset. The authors also cautioned that the size of a dataset should not be the only consideration for quality or effectiveness, but image diversity, completeness, appearance, object occurrence frequency and distribution should also be considered. This was considered while creating synthetically augmented training and test datasets for the OctopusNet model.

Techniques similar to the works cited in the previous section were relied on to develop a training dataset of synthetic data. Image composition was used to generate new images by composing the isolated object of interest – or specific surgical tool – as a foreground on multiple background images (Figure 7.1). Since the images in the HospiTools dataset consisted of clean objects with well defined coloured backgrounds, the work used standard techniques to cut out the tool from the images. Experiments with edge detection techniques were conducted — including Laplacian, Canny and Sobel gradient edges to detect object edges — and Canny was used as it provided the optimal results for these particular images. Noise was then filtered out with median filters, and the largest contour in the image by area was found. Using the resultant polygon, contour detection techniques were used to define the object of interest and to fill in the holes, and contour smoothing algorithms and Gaussian Blur was used to define the edges. Once this was achieved and a satisfactory approximate contour of the surgical tool was obtained, the GrabCut algorithm (Rother, Kolmogorov, and Blake, 2004) was used with accurate background and foreground differentiation to cut out and save the surgical tool mask. These surgical tool object masks were then composed on a range of different backgrounds, and variations were obtained through various random rotations, random zoom, random translations, random cut-outs and random horizontal flips.

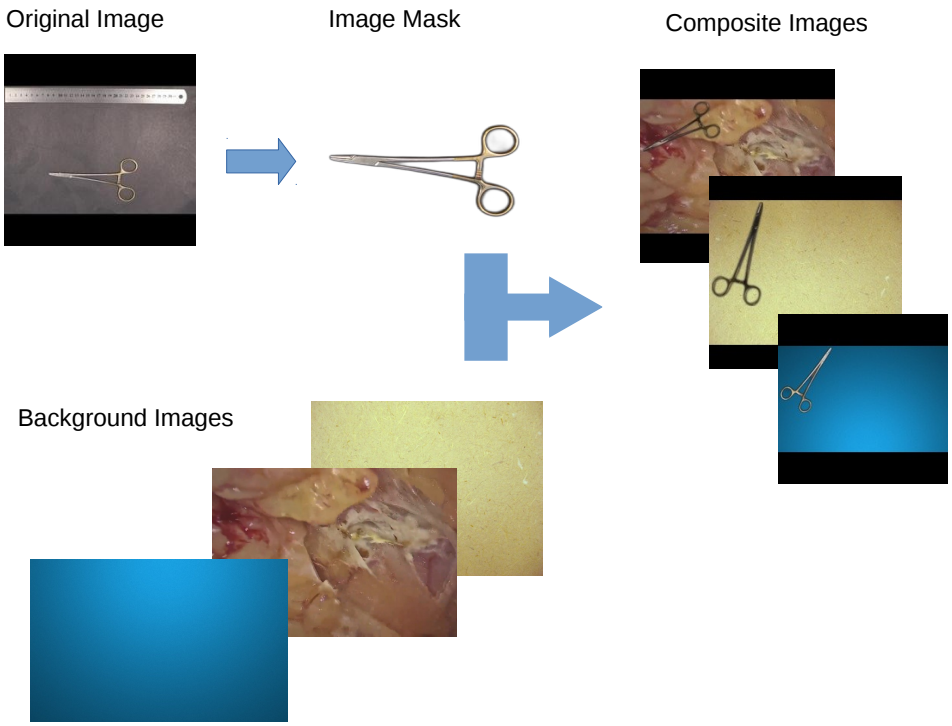


FIGURE 7.1: Example of Synthetic Dataset Composition

Simple techniques were used to remove the backgrounds in the surgical tool images, to isolate the actual tool, and then to superimpose the tool image on various backgrounds, including more realistic medical backgrounds. Background images were created using a DSLR camera, and medical and tissue images from work by Garcia-Peraza-Herrera et al. (2021b) was also used. Multiple transformations were introduced while superimposing the tool object on the background, including flipping, random resizing, random rotation, random cutouts and random noise. A similar approach had been adopted by Tremblay et al. (2018), which was to paste real images (as opposed to synthetic images) of objects on background images. The background could then include images of actual surgeries and was more representative of actual use conditions of the tools. The augmented dataset consisting of 500 images per class for 19 classes was developed using these techniques, and there were a total of 9500 images in the synthetic dataset. Examples of the synthetic dataset are provided in Figure 7.2.

7.2.1 OctopusNet with HospiTools and Synthetic Datasets

The OctopusNet architecture (as presented in Chapter 4 and Chapter 5 of this thesis) was trained on a subset of the HospiTools dataset and on the Synthetic Tool Dataset, with the same training configurations used earlier — as in Table 7.1. During training, the data augmentations of random brightness, random contrast, and random flips were applied. In both cases, early stoppage of training was adopted to avoid overfitting, and the best reported model was retained. Both architectures was trained on a batch size of 16 on an NVIDIA 3060 GPU based machine. Augmentation was used while training on the HospiTools dataset but not while training on the Synthetic Tool

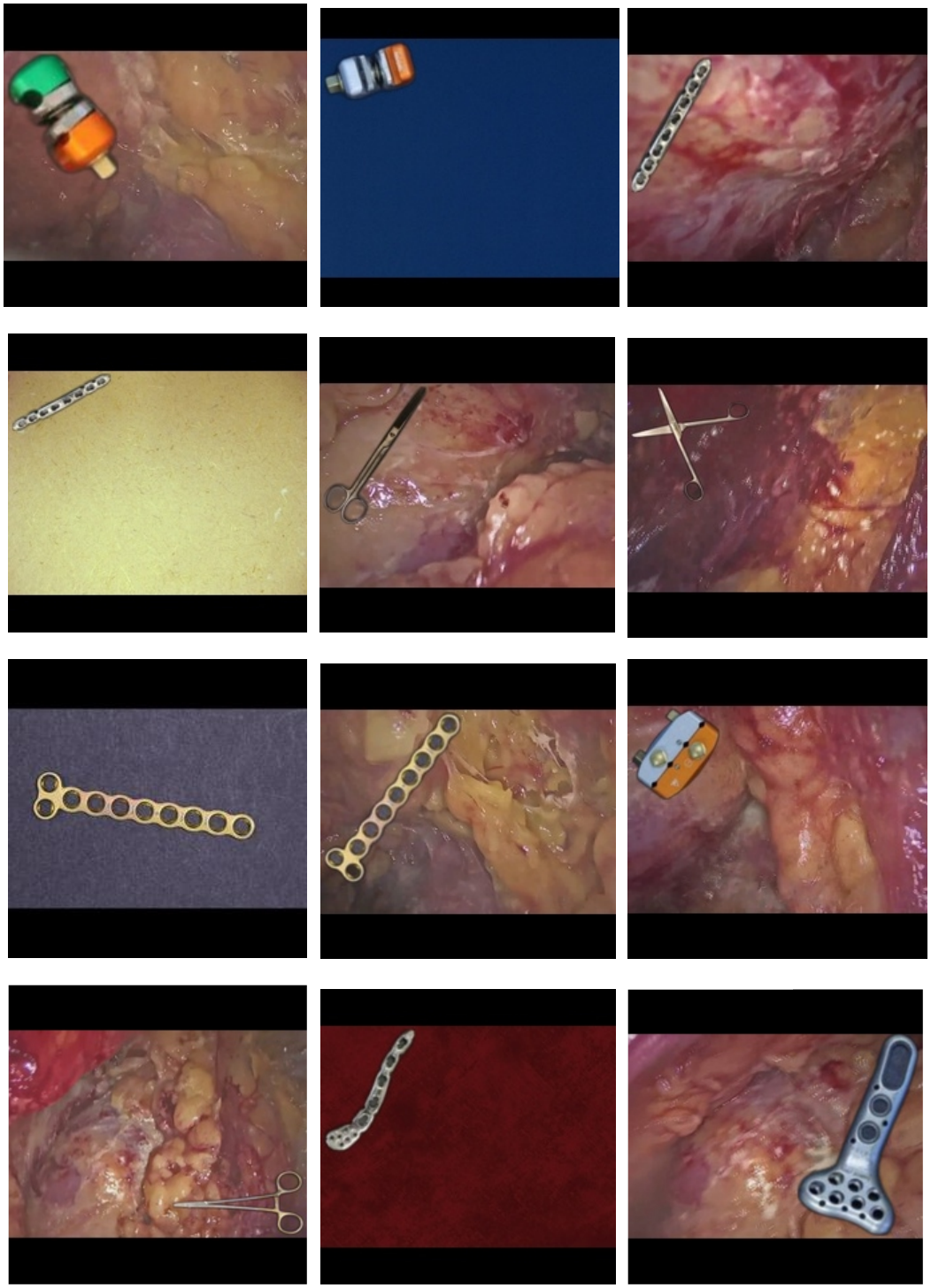


FIGURE 7.2: Synthetic Dataset Example Images

TABLE 7.1: Training Configuration for Synthetic Dataset Experiments

Optimiser	Learning Rate	Epochs	Activation	Loss	Metric
Adam	0.001	100	Softmax	Categorical Crossentropy	Categorical Accuracy

dataset since experiments had found that adding training time augmentation on the synthetic images degraded performance.

The ResNet-50V2 base network was pre-trained on the relevant dataset — a subset of the HospiTools dataset and the new Synthetic Tool dataset — in each training regime. The pre-trained ResNet-50V2 base network was then frozen from the input layer up to the start of Block 5 while training the other layers. This freezing strategy had also been evaluated by Hinterstoisser et al. (2019) and by Tremblay et al. (2018), who evaluated the performance results from freezing the weights of the early network layers (in their cases pre-trained on ImageNet) when training with synthetic data. The network was trained for 100 epochs with early stopping implemented on validation loss. A learning rate schedule was imposed with a low initial learning rate; the rate was decreased to 0.0001 at epoch 45 and to 0.00005 at epoch 65. A dropout rate of 0.3 was imposed. Separate holdout sets of images for testing were used, as discussed below.

7.2.2 OctopusNet with Challenging Test Datasets

To test the performance of the networks, three challenging datasets were created, in addition to the held-out test set of the HospiTools dataset. A "real" dataset was developed by taking images of surgical tools with varying backgrounds and in different illumination conditions. There was significant variety created in these "real" test images, and this included multiple different illumination sources at varying distances, changes of background, and changes in orientation and distance to the object of interest. Challenging backgrounds were introduced, including highly reflective surfaces, light absorbing surfaces, textured and coloured backgrounds, and surfaces with wood grain striations. A further "mixed" dataset was then created which consisted of a range of images from the real and synthetic datasets. This dataset included variations in the image sizes, occlusions in the images, and different class frequencies in the test sets, including imbalanced classes. The class frequencies of these test sets are presented in Table 7.2. Examples of the real test dataset are provided in Figure 7.3, and examples of the mixed test dataset are provided in Figure 7.4. It was anticipated that this would provide a good test of the architecture and of the use of synthetic data for training to increase robustness of the CNN.

7.2.3 OctopusNet with Synthetic Filters Dataset

Since training with synthetic datasets was resulting in good predictive performance boosts, experiments with standard and innovative image creation and augmentation techniques was conducted to further enhance the dataset. Since off-the-shelf image augmentation libraries did not result in good training and performance, a range of image filters were developed, which were applied to the standard HospiTools image dataset. These filters were developed based on the OpenCV library and its functions. For example, the "line" function was used to add vertical, slanted lines to simulate rain, the "blur" function to add dirt and blood effects to patches in the image or to

TABLE 7.2: Class Frequencies - Challenging Test Sets

Class	HospiTools	Synthetic	Mixed	Real
Bearing Plates 3-Hole 7-Peg Right Narrow	73	50	100	0
Lahey Forceps	72	50	168	0
Multi-pin Clamp Grey-Orange	66	50	100	68
Clavicle Plate 3.5-2.7 5-Hole Right	53	50	100	0
7 Inch Metzenbaum Scissors	82	50	156	56
Crile Artery Forceps	79	50	110	10
9 Inch DeBakey Needle Holder	64	50	100	0
6 Inch Mayo Needle Holder	75	50	100	0
2.4mm Titanium T-Plate 2-Hole x 8-Hole 54mm	72	50	100	0
8 Inch Babcock Tissue Forceps	36	50	255	155
Ball and Socket Towel Clips	72	50	100	0
Littlewood Tissue Forceps	74	50	100	0
Allis Tissue Forceps	71	50	180	80
Dressing Scissors	74	50	286	186
2.0mm Titanium Straight Plate 12-Hole 71mm	47	50	100	0
Fixed Angle Plates 3-Hole 7-Peg Left	60	50	100	0
Clavicle Plate Medial 8 Hole	82	50	100	0
Mayo Artery Forceps	79	50	100	0
Pin to Rod Coupling Grey-Orange	105	50	100	0

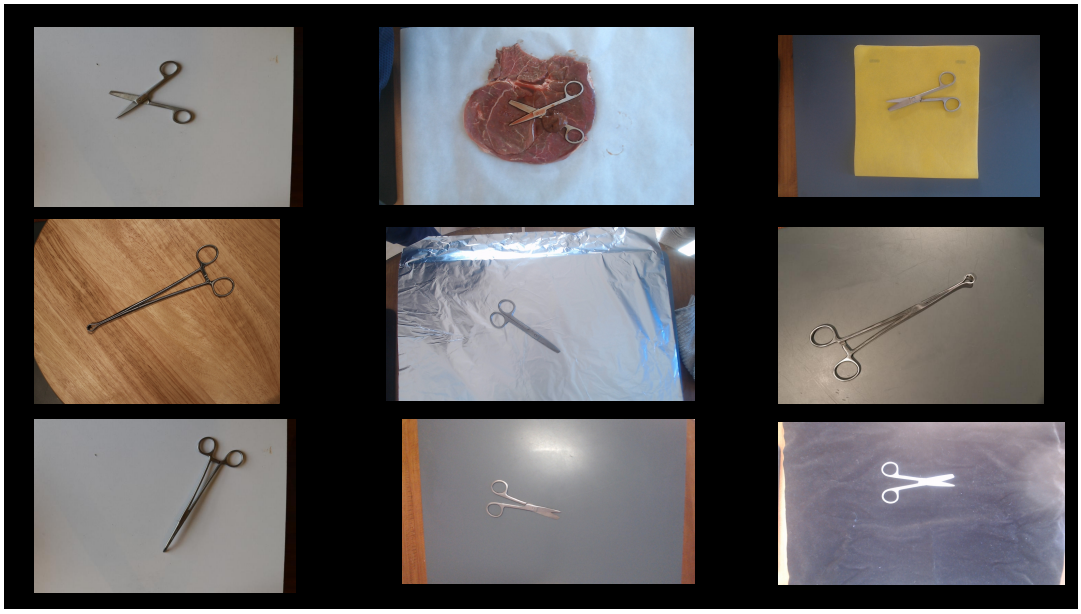


FIGURE 7.3: Real Test Dataset Example Images

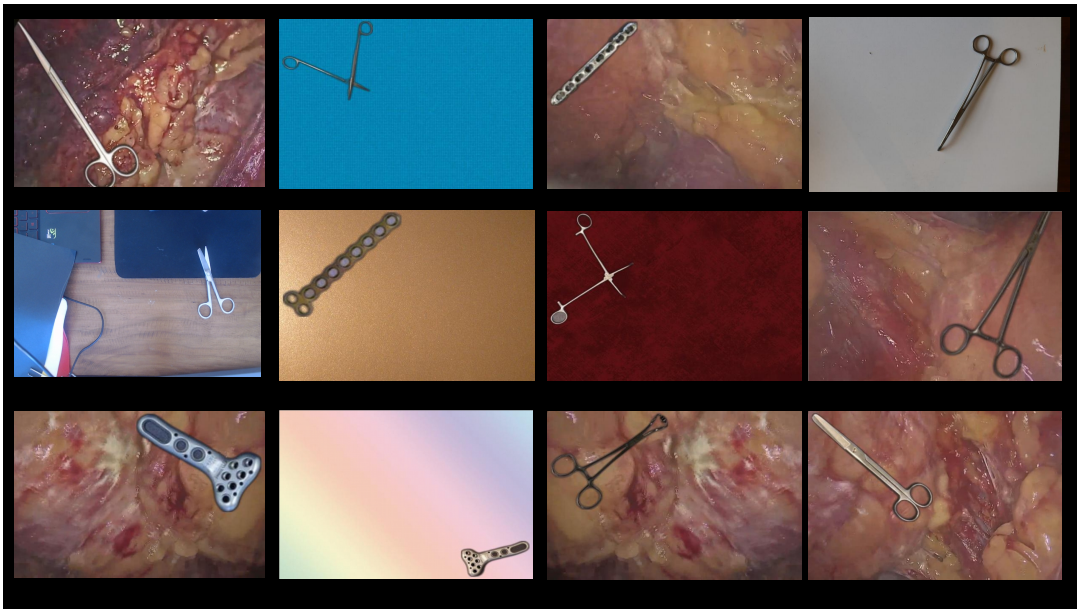


FIGURE 7.4: Mixed Test Dataset Example Images

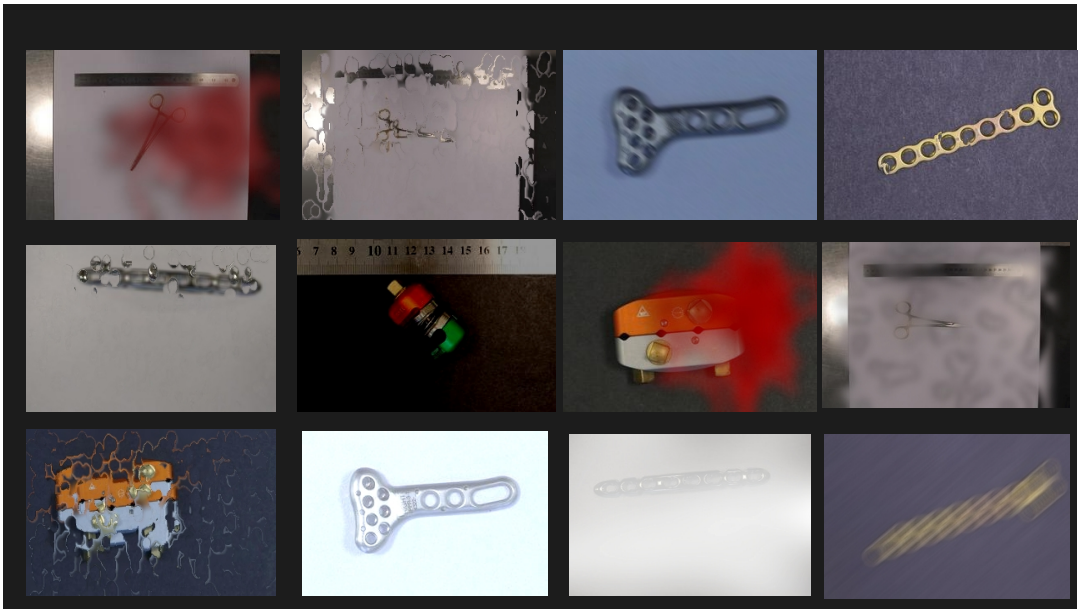


FIGURE 7.5: Synthetic Filters Dataset Example Images

simulate movement, and the brightness or dark effects by changing pixel values in the images.

In all, 19 different filters were developed and utilised to add various effects to the images, and this included three intensities of brightness filters, three degrees of dirt filters, three speeds of motion filters, three densities of focused rain filters and three flare level filters – light, medium and heavy in all cases. In addition, one heavy dark illumination filter, one blood splatter filter, and two unfocused rain moisture filters were applied. For each filter effect, an average of 68 images were created for each of the 32 classes selected for augmentation —- resulting in a total number of 41,534 images in the Synthetic Filter Dataset. This resulted in a heavily augmented dataset as in Fig 7.5 with an average of 1297 images per class. This was purposeful augmentation, to reflect the types of image variations that the CNN might see in real world conditions. The OctopusNet CNN was trained on this filter-based synthetic dataset and its performance was evaluated.

7.2.4 OctopusNet with Synthetic Filters Dataset and Distractor Classes

As a final test, experiments were conducted based on the addition of "Distractor Classes" (Garcia-Peraza-Herrera et al., 2021b) to the Synthetic Filters Dataset during training. This was done to evaluate if distractor classes could improve training, by forcing the CNN to focus on only relevant aspects of the dataset and images. This was an approach used by Garcia-Peraza-Herrera et al. (2021b), who used "flying distractors" or artefacts that were not present in their real images. They hypothesised that these distractors would assist the CNN to focus on the segmentation features to detect surgical tools. To create the distractor classes, two classes with random images were introduced, as shown in Fig. 7.6. The classes included approx. 500 random images each, and meaningless annotations for each class were added to the attribute matrix. OctopusNet was trained using the synthetic dataset and the distractor classes, and its performance was evaluated.

7.3 Results

The results of training OctopusNet on the HospiTools Dataset and testing it on the challenging test sets are presented in Table 7.3, and the results of training OctopusNet on the Synthetic Dataset are presented in Table 7.4. As reported by Tremblay et al. (2018), the work conducted and reported in this section also finds that there is a significantly improved performance from training the network on synthetic data, particularly at the set and tool level accuracy. Results on the mixed dataset are encouraging, and there is significant improvement in the results on the challenging real images — even though much more work is required to get the performance levels to an acceptable standards. Adding variations to synthetic data, changing backgrounds, adding different illumination sources, zooming in and out and creating different perspectives / angles of vision are fairly easy to implement yet are clearly powerful in terms of increasing robustness to issues such as changes in illumination and background, and in reducing brittleness of the model.

The results of training OctopusNet with the filter-based synthetic dataset are presented in Table 7.5. The results were encouraging — while the accuracy results were good at all levels for the Test set of the Filter Dataset, performance was also improved on the HospiTools Test data even though the model had not seen any images from the HospiTools dataset during training. The results with this dataset

TABLE 7.3: Results - OctopusNet Trained on HospiTools Dataset

Level	Metric	HospiTools	Synthetic	Real	Mixed
Speciality	Accuracy	1.00	0.84	0.97	0.87
	Hamming Loss	0.00	0.16	0.03	0.13
	f1 score macro	1.00	0.83	0.87	0.85
	Precision	1.00	0.85	0.82	0.85
	Recall	1.00	0.83	0.93	0.84
Pack	Accuracy	0.97	0.45	0.50	0.46
	Hamming Loss	0.03	0.55	0.50	0.54
	f1 score macro	0.98	0.49	0.70	0.52
	Precision	0.99	0.61	0.70	0.65
	Recall	0.98	0.45	0.74	0.47
Set	Accuracy	**0.99**	**0.64**	**0.65**	**0.64**
	Hamming Loss	0.01	0.36	0.35	0.36
	f1 score macro	0.99	0.39	0.46	0.41
	Precision	0.99	0.50	0.47	0.57
	Recall	1.00	0.37	0.53	0.39
Tool	Accuracy	**0.98**	**0.22**	**0.18**	**0.20**
	Hamming Loss	0.02	0.78	0.82	0.80
	f1 score macro	0.98	0.19	0.29	0.20
	Precision	0.98	0.27	0.33	0.29
	Recall	0.98	0.20	0.37	0.23

TABLE 7.4: Results - OctopusNet Trained on Synthetic Dataset

Level	Metric	HospiTools	Synthetic	Real	Mixed
Speciality	Accuracy	0.70	0.97	0.71	0.91
	Hamming Loss	0.30	0.03	0.29	0.09
	f1 score macro	0.69	0.97	0.55	0.91
	Precision	0.78	0.97	0.58	0.90
	Recall	0.75	0.97	0.85	0.94
Pack	Accuracy	0.71	0.90	0.45	0.80
	Hamming Loss	0.29	0.10	0.55	0.20
	f1 score macro	0.69	0.93	0.47	0.85
	Precision	0.74	0.93	0.46	0.86
	Recall	0.73	0.94	0.68	0.87
Set	Accuracy	**0.84**	**0.99**	**0.67**	**0.92**
	Hamming Loss	0.16	0.01	0.33	0.08
	f1 score macro	0.82	0.99	0.54	0.94
	Precision	0.80	0.99	0.52	0.94
	Recall	0.87	0.98	0.76	0.95
Tool	Accuracy	**0.61**	**0.82**	**0.24**	**0.68**
	Hamming Loss	0.39	0.18	0.76	0.32
	f1 score macro	0.61	0.82	0.28	0.72
	Precision	0.73	0.86	0.30	0.74
	Recall	0.63	0.82	0.47	0.76

TABLE 7.5: Results - OctopusNet With Synthetic Filters

Level	Metric	Synthetic Filters	HospiTools	Real
Speciality	Accuracy	1.00	1.00	0.94
	Hamming Loss	0.00	0.00	0.06
	f1 score macro	1.00	1.00	0.84
	Precision	1.00	1.00	0.78
	Recall	1.00	1.00	0.97
Pack	Accuracy	1.00	1.00	0.58
	Hamming Loss	0.00	0.00	0.42
	f1 score macro	1.00	1.00	0.65
	Precision	1.00	1.00	0.62
	Recall	1.00	1.00	0.81
Set	Accuracy	**1.00**	**1.00**	**0.67**
	Hamming Loss	0.00	0.00	0.33
	f1 score macro	1.00	1.00	0.66
	Precision	1.00	1.00	0.63
	Recall	0.99	1.00	0.84
Tool	Accuracy	**1.00**	**1.00**	**0.27**
	Hamming Loss	0.00	0.00	0.73
	f1 score macro	1.00	1.00	0.54
	Precision	1.00	1.00	0.52
	Recall	0.99	1.00	0.81

were much better than the results obtained with the previous synthetic dataset, and the implications are that purposeful augmentation using a range of filters is a useful technique. Clearly both synthetic images and heavily — but specifically — augmented filter based synthetic images can be used to achieve good performance, and good results are obtainable even in the absence or shortage of actual images. However, the Filter Synthetic Dataset trained CNN performance on the challenging "real" images does not improve from the results we obtained from training with the previous synthetic dataset, even though it is still better than training with normal images.

OctopusNet was trained using the synthetic dataset and the distractor classes, and the test results are presented in Table 7.6. It was found that the addition of Distractor Classes did not improve or degrade the results significantly, but that there was consistently good predictive performance in test results from the unseen HospiTools images.

7.4 Conclusions

The work conducted in this chapter reported that synthetic data is useful for training of the CNN, and specific data augmentation using filters is also useful, particularly in terms of better predictive performance and accuracy at the set and tool level of predictions. It used the first of the two methods for synthetic data creation as discussed by Nowruzi et al. (2019) – real data augmentation, but work could include surgical tool image generation in surgical settings by using simulations that leverage

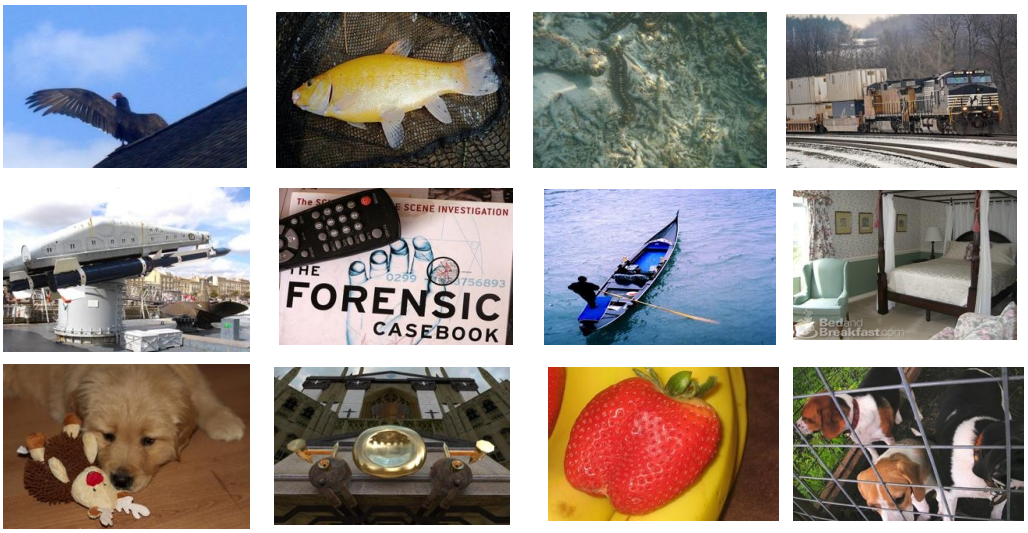


FIGURE 7.6: Distractor Classes Examples

TABLE 7.6: Results - OctopusNet With Synthetic Filters and Distractor Classes

Level	Metric	Synthetic Filters and Distractors	HospiTools	Real
Speciality	Accuracy	1.00	1.00	0.80
	Hamming Loss	0.00	0.00	0.20
	f1 score macro	0.77	0.80	0.51
	Precision	0.81	0.83	0.45
	Recall	0.80	0.83	0.75
Pack	Accuracy	0.99	1.00	0.55
	Hamming Loss	0.01	0.00	0.45
	f1 score macro	0.90	0.91	0.58
	Precision	0.92	0.93	0.55
	Recall	0.92	0.92	0.76
Set	Accuracy	**1.00**	**1.00**	**0.64**
	Hamming Loss	0.00	0.00	0.36
	f1 score macro	0.92	0.93	0.63
	Precision	0.93	0.94	0.64
	Recall	0.93	0.94	0.78
Tool	Accuracy	**0.99**	**1.00**	**0.29**
	Hamming Loss	0.01	0.00	0.71
	f1 score macro	0.96	0.97	0.49
	Precision	0.97	0.98	0.45
	Recall	0.97	0.97	0.78

3-D images and digital twins (Kritzinger et al., 2018; Tao et al., 2019). Future work can build on this foundation of synthetic images by creating additional surgical tool images with much more variety, training on combinations of real and synthetic data, training on synthetic data and fine tuning on real data, and other strategies to improve performance. This reliance on synthetic data is much easier to accomplish, less complicated and less expensive than taking real images in actual hospital conditions, and provides a viable forward path for increasing the deployability of the surgical tool management system in actual hospital and clinic settings.

Chapter 8

Informed Machine Learning

Previous chapters in this thesis (Chapters 4 and 5) have described the dataset, and developed a CNN framework that successfully utilised the hierarchical nature of surgical tool classes to provide a comprehensive set of classifications for each category, sub-category, sub-set and specific tool. To complement the dataset, a surgery knowledge-base was developed as an attribute-matrix which makes relevant and useful information available to the training regime. This proved to be a convenient and useful data structure that captures rich information of class attributes — or the nameable properties of classes — and makes it readily available for computational reasoning (Lampert, Nickisch, and Harmeling, 2014). The work reported in this chapter experiments with using this prior information in training a CNN, since this could be useful in accurate classification of new surgical tools that have not been seen by the CNN. It therefore address informed machine learning, defined in terms of the integration of prior knowledge into the training process of the CNN (Rueden et al., 2021). This explicit integration of knowledge into the machine learning pipeline has also been described as knowledge infused learning (Dash et al., 2022). Machine learning algorithms which integrate domain knowledge have been shown to perform better than purely data-driven machine learning (Deng et al., 2020) — earlier chapters have reported results from experiments using prior knowledge, and the work in this chapter experiments with new ways to incorporate surgical tool knowledge in a CNN using the dataset and knowledge-base.

8.1 Surgical Tool Detection Prototype and Deployment

A specific target of the work in this thesis was deployment and testing in real world conditions, and the intention was to develop a scanning system — Figure 8.1 — and deploy it in critical points within a hospital. This system could be used to scan tools at various points and track the flow of tools within the hospital — Figure 8.2 — as well as to assist in packing of tool. Unfortunately, this was not possible due to the COVID-19 crisis since priorities within hospitals were oriented towards critical and emergency care, and there was no access to test the system in a hospital. However, a prototype system was developed — Figures 8.3 and 8.4 — and was evaluated in simulated conditions, though this was only possible within research labs. This system can correctly identify tools, provide possible alternatives to the end-user, provide basic information about the set that the tool belongs to and other inventory details — Figure 8.5. This is potentially useful, but it still needs to be evaluated in actual conditions.

A further issue is the fact that additional prior information is available which can be incorporated into the model to improve training and predictive performance. The work conducted in this chapter therefore focuses on the use of attributes as

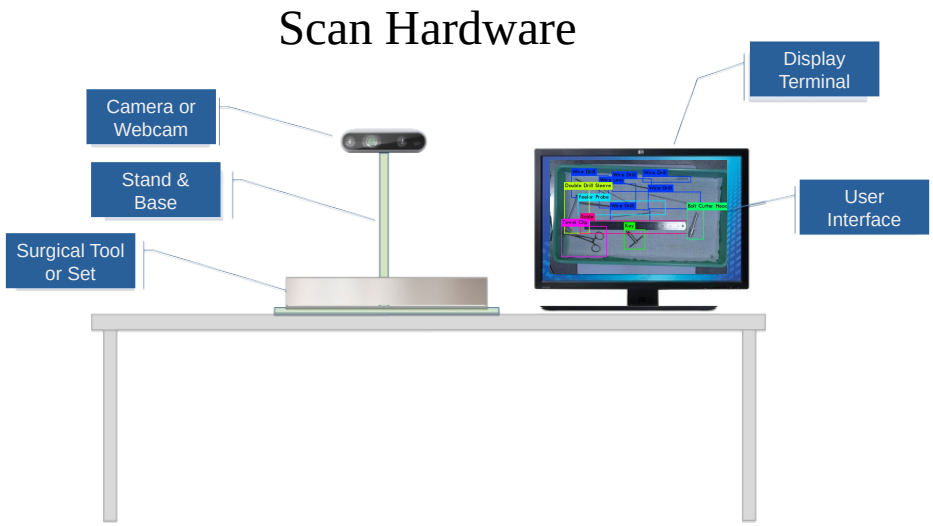


FIGURE 8.1: Example of a Prototype Testing Setup

external sources of knowledge, and uses the attribute matrix as prior information in the training regime of a CNN. It uses text and knowledge graphs to formally represent prior knowledge and on a learning algorithm approach for knowledge integration, implemented as a loss function and regularizer in the training process (Rueden et al., 2021; Dash et al., 2022). The intention is to use both images and attributes in the training process in an effective manner. The utility of such multi-modal representations — where images and attributes form the two modes — in the intelligent management of surgical tools is addressed in the next sections, and the provision of additional information in the training process to improve predictive performance of a CNN is evaluated.

8.2 Image-Text Embeddings

An earlier section (Chapter 6) had evaluated techniques to map images and labels to a shared latent space where prior domain knowledge was used to construct a semantic embedding space, and then image embeddings were projected to this space (Barz and Denzler, 2019; Narayana et al., 2019). The basic idea was that distances between image embeddings were similar to class embedding distances in the semantic embedding space. The approach of Barz and Denzler (2019) was adopted, that relied on domain knowledge encoded in a hierarchy of classes to incorporate class information and semantic relationships into a deep learning model, and the technique achieved good results on the surgical tool dataset. The work conducted in this chapter further explores the use of prior knowledge in the training of a CNN.

A critical concern in computer vision based solution is the transformation of pixel representation of images into more useful representations, or feature extraction. This has been defined in terms of dimension reduction — capturing and retaining relevant information from the original pixel representations in a lower dimension space. Techniques relied on for feature extraction in images include principal component analysis (PCA), projection histograms, Zernike moments, Fourier descriptors, Gabor filters and template matching — to mention just a few important methods (Kumar

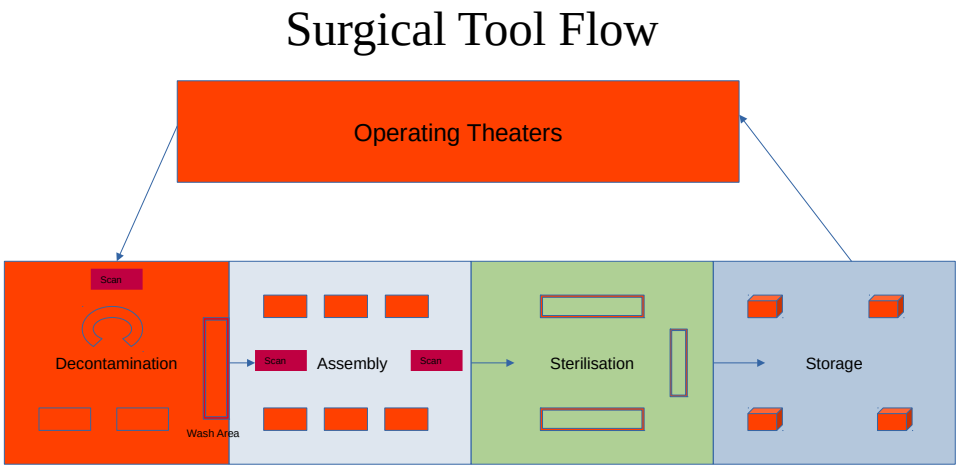


FIGURE 8.2: Example of Prototype Tool Scanning Point

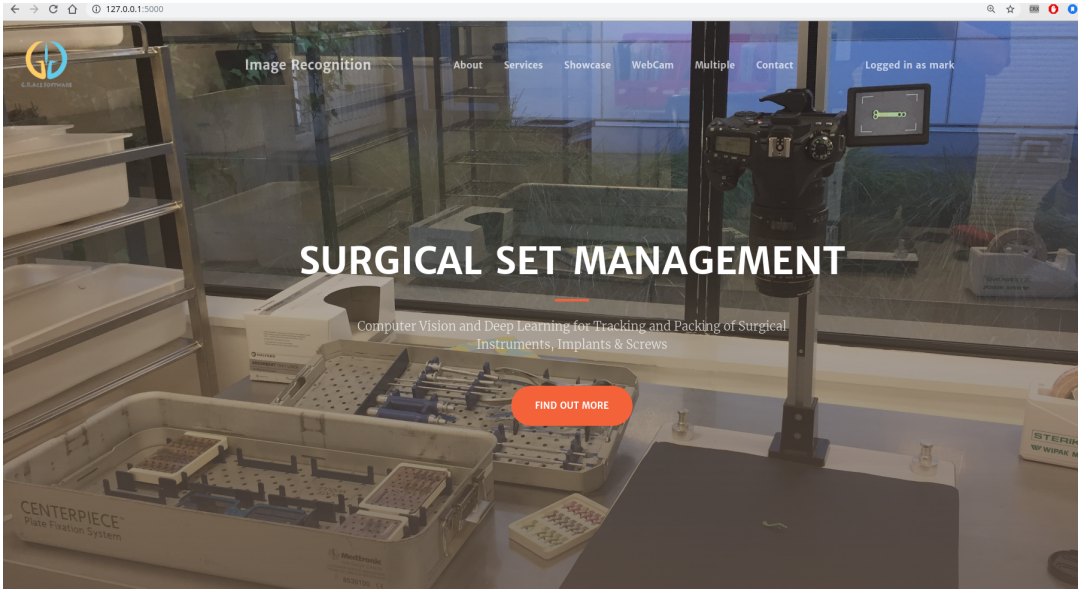


FIGURE 8.3: Example of Prototype System User Interface

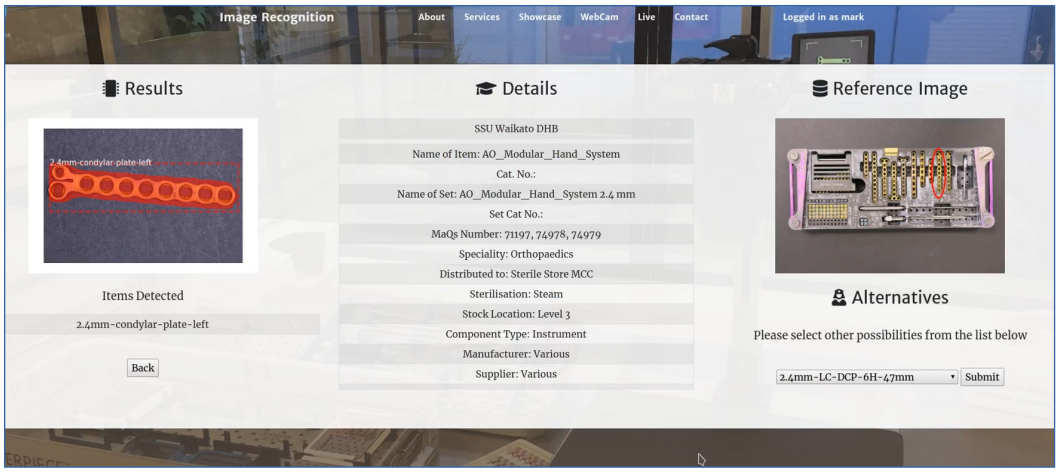


FIGURE 8.4: Informative Inference Results using Prototype System

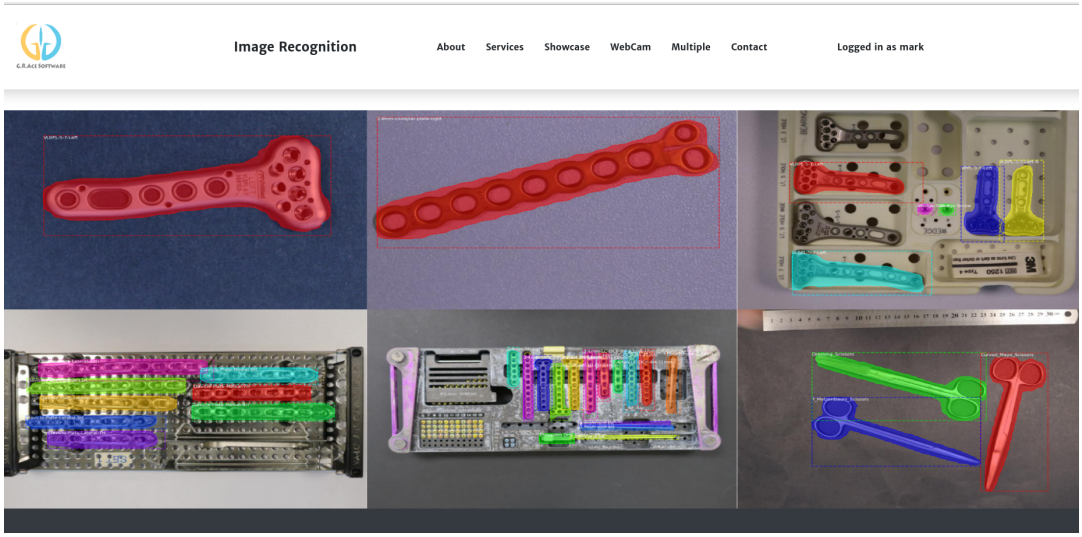


FIGURE 8.5: Detection Examples with Prototype System

and Bhatia, 2014). Akata et al. (2015) stated that good image representations are critical for good performance, and highlighted how features extracted using a CNN are useful for many applications. Feature extraction using CNNs and the subsequent vector representations of images — or embeddings — have been gainfully used for many computer vision tasks (Ueki, 2021). Specially designed embedding networks have been used both for feature extraction and for the organisation of the vector representations or embeddings into low-dimensional output spaces (Frome et al., 2013; Miller et al., 2020; Narayana et al., 2019; Wang, Li, and Lazebnik, 2016). Such networks are trained to learn an embedding space where similar embeddings are closer than dissimilar embeddings, and this can be used for image retrieval and classification tasks (Ueki, 2021).

Defining how embeddings are organized or mapped in the output space is a critical aspect of any system or solution. Feature embeddings should retain user defined concepts of semantic similarity — where images perceived as being similar in some form should be closer in the embedding space than images deemed to be different. This is not a trivial task, since images may have illumination, background, orientation or resolution differences yet are defined to be semantically similar by a user. One widely used technique to retain semantic similarity is metric learning (Weinberger, Blitzer, and Saul, 2005), where the feature representations are mapped or organised in an embedding space such that the L2 distances between embeddings correlate to the similarity between images. This can also be used for text — for example, Akata et al. (2015) used both image and text embeddings, mapped in a joint framework that learnt the compatibility between the different embedding types. Wang, Li, and Lazebnik (2016) sought to learn a lower dimension latent space for image and text embeddings, where vectors from the two modalities could be compared. In their solution, they used a two-branch neural network for learning joint image and text embeddings. In such a learned embedding space, metrics such as the L2-distance can be used to determine the similarity or dissimilarity between embeddings.

In other work using this approach, Frome et al. (2013) developed a deep visual-semantic embedding model (DeViSE) that learned semantic relationships between labels, and mapped images into semantic embedding space. Demirel, Cinbis, and Ikizler (2017) proposed attribute-based zero-shot learning which mapped distinctive attributes in images to a semantic word vector space. Their system evaluated similarities between classes and combination of attribute names to evaluate visual similarity, and this was then used to predict unseen classes based only on these names. Akata et al. (2015) developed a "Structured Joint Embedding" framework that related image embeddings and text embeddings through a compatibility function, and demonstrated that embedding labels in an Euclidean space was an effective technique to capture relationships between classes. In a mapping function approach used for zero-shot learning, Socher et al. (2013) trained a deep neural network on images to obtain rich image representations along with a language model that obtained embedding representations for relevant words. Linear mapping was used to link the image representation space to the embedding space for eight classes that the model was trained on. The system then evaluated if a given test image was from the 8 classes; else the nearest class in the embedding space was used to classify the image. The next section discusses some of the issues and challenges in using multi-modal embeddings.

8.2.1 Multi-modal Learning

Miller et al. (2020) used metadata — text descriptions, titles and tags — along with images in a multi-modal model to improve image classification. Their model used ResNet50 and a Universal Sentence Encoder to process images and text in parallel towers of deep convolutional and sequence networks. Features specific to each modality — image, text descriptors and labels — were extracted, flattened and concatenated via dense layers into a single feature vector, and predictions were obtained from the combined feature vector. A useful approach for training Image-Text CNN models is therefore using neural networks to extract high-level features of both images and text. Those features can then be projected onto a shared visual-semantic or multi-modal embedding space in such a way that correlated embeddings in the shared multi-modal space are closer to each other while uncorrelated embeddings should be far from each other (Wehrmann, Kolling, and Barros, 2019). For example, Gong et al. (2013) discussed how canonical correlation analysis (CCA) can be used to map visual and and textual features into a cross-modal common latent space where their correlation is maximized.

Using multi-modal embeddings — specifically, image and text representations — to share a common latent space could allow exploitation of any complementary information between these modalities, within the enriched common latent space. Enriched in this context refers to the incorporation of additional semantic information in the mapping space, and complementary information could exist between an image and associated textual information that describes the object, scene or concept (Gallo, Calefati, and Nawaz, 2017). However, there are significant challenges that arise in a multi-modal setting because of what has been referred to as a media gap, where inconsistencies in the features from different modalities can make it difficult to exploit complementary knowledge (Narayana et al., 2019). Bayoudh et al. (2022) points out that mono-modal representations — which can be from image, text, audio, video sources — are about linear or nonlinear mappings to high-level semantic representations. There are significant issues in adapting the deep learning model to properly utilise the representation spaces of both the input and output modalities. In this context therefore, the joint or shared embedding space is crucial for exploiting any synergies in the multi-modal data (Bayoudh et al., 2022). The main challenge lies in mapping images and text to a shared latent space where the embeddings corresponding to a similar semantic concept lie closer to each other than the embeddings corresponding to different semantic concepts, irrespective of the modalities (Narayana et al., 2019). Multi-modal learning is therefore about representing a specific object of interest from multiple perspectives or modalities, while maintaining any complementary information and semantic context which can be used to train the network (Bayoudh et al., 2022).

To address this issue, Wang, Li, and Lazebnik (2016) mapped both image and sentence representations — with different dimensions and using different feature extractions techniques — to a joint common dimension space by using a two branch deep CNN framework. Each branch consisted of fully connected layers, a Rectified Linear Unit (ReLU), batch normalization and L2 normalization. Embedding outputs from the image and text layers were L2-normalized and so the inner product between embeddings — equivalent to the Euclidean distance in this space — was used to measure similarity or dissimilarity between embeddings. The work reported in this chapter explores this technique and approach further.

A multi-modal representation needs to be able to leverage any correlation power of each individual mono-modal representation via an aggregation of outputs. This

can be achieved by early fusion, where representative mono-modal features are fused before being classified, or by late fusion, where the features are classified before fusion for a final decision (Bayoudh et al., 2022). Early fusion has also been defined to be at the level of features — for example, by concatenating image and text embeddings into one multi-modal vector — while later fusion is at the level of decisions — for example, image and text embeddings are used to obtain independent decisions and a final decision is calculated based on the weighted product of these two independent decisions (Gallo, Calefati, and Nawaz, 2017). This work maintains a focus on late fusion.

8.3 Prior Domain Knowledge

In informed machine learning models, knowledge transfer is achieved by utilising prior information from attributes, class hierarchies, vector embeddings of names, or text descriptions of classes (Xian et al., 2018; Pourpanah et al., 2020). A common approach has been to use text data that describes class characteristics to classify object categories. This relies on semantic embeddings to learn a mapping from visual space to semantic space, represented by semantic word vectors (Demirel, Cinbis, and Ikizler, 2017; Socher et al., 2013; Frome et al., 2013). A system can consist of, for example, a pre-trained neural network that generates image features for each object in the image, a model trained on prior knowledge in the form of relevant text that represents words as vectors, and a neural network that projects the object features to the suitable embedding space. The distance between text embeddings corresponding to two words has been shown to be an effective method for measuring the semantic similarity of the corresponding words, and this has been used to trained CNNs for good predictive performance (Narayana et al., 2019).

A significant problem in using a text based approach for surgical tool management is the requirement for a corpus that accurately captures domain information. There are no databases of textual descriptions for each surgical tool, though there are textbooks and manufacturer's manuals that provide a basic terminology. In general terms however, available text is inconsistent, imprecise, sometimes irrelevant, noisy and overloaded with ambiguous words. A further problem is that such text has low semantic expression and cannot accurately capture fine-grained relationships between tool classes (Chen et al., 2021b). A major task therefore, was to create such a database of textual descriptions, and this was created in the form of a surgical tool knowledge-base. This was in the form of an attribute matrix or a set of annotations where each row represents a particular surgical tool, and this attribute set can be expanded to add multiple other descriptors. The annotations capture details such as the surgical speciality, pack and set that the tool is belongs to, and other details such as size, shape, type and special features. Additional annotations include image modalities, illumination sources, geographic location, tool manufacturer and camera type. Multiple other attributes could be usefully added to the attribute matrix; for example, to identify the particular surgeon and his/her preferences so that a constrained set of tools could be used for predictions.

While this is a start on creating a dataset of textual descriptions for surgical tool, work to improve the quality, accuracy and comprehensiveness of the annotations continues to progress. Again, the current pandemic situation and subsequent demands on the workload of medical professionals meant that planned meetings with doctors and surgical tool technicians to evaluate and improve the surgical tool annotations

and textual descriptions had to be cancelled. However, the current annotations provide a basic starting point for experiments and this attribute matrix was therefore used to provide prior information to the model during training.

While text is commonly used in CNN models, other forms of knowledge can also be used. Chen et al. (2021b) defined the external knowledge used to craft such relationships into four kinds — text, attribute, Knowledge Graph (KG) and ontology and rules. Rueden et al. (2021) defined eight categories of knowledge representations in their exhaustive survey of informed machine learning – including logic rules, knowledge graphs, probabilistic relations and human feedback. This work relies on both texts and a knowledge graph approach in the experiments, and the method is described in a later section. Incorporation of knowledge-graph based information into a deep learning model has been termed knowledge-infused learning, and knowledge graphs have been shown to be useful for the structured representations of knowledge (Dash et al., 2022).

Wehrmann, Kolling, and Barros (2019) evaluated the impact of different pre-trained word embeddings in their models — such as GloVe embeddings (Pennington, Socher, and Manning, 2014) and concatenation of randomly initialized word-embeddings and Glove vectors. Their working hypothesis was that their models would benefit from pre-trained embeddings, because such embeddings already incorporated rich semantic representations. Akata et al. (2015) also evaluated five different word embeddings, defined as supervised attributes, unsupervised Word2Vec (Mikolov et al., 2013), GloVe, Bag of Words, and WordNet-derived similarity embeddings. For Word2Vec and GloVe embeddings, they pre-trained the system using Wikipedia text. Miller et al. (2020) discussed Word2Vec, GloVe, Universal Sentence Encoder and BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) in their work.

In work that focused on the medical domain, Rasmy et al. (2021) developed a medically contextualized embedding model which they termed Med-BERT. This was pre-trained on a very large structured electronic health records (EHRs) dataset. This was, however, more oriented to disease prediction studies, including diabetes heart failures and pancreatic cancer conditions. In a similar manner, Li et al. (2021) developed BEHR (or BERT for EHR) which was described as a deep neural sequence model for EHRs, focused on medical diagnosis, medication and measurements tasks. Shang et al. (2019) presented G-BERT, a model based on BERT and Graph Neural Networks (GNNs), for medical code representation and medication recommendation and Huang, Altosaar, and Ranganath (2019) applied BERT to clinical notes to build a model for hospital readmission predictions. While these initiatives targeted the medical domain in general, there is no BERT model trained on surgical datasets. However BERT has been shown in the above works to be useful for medical contexts, and this work therefore relied on BERT and processed the surgical tool annotations accordingly. BERT pre-processing includes marking out paired sentences, which facilitates contextual learning. BERT can also capture semantic data about meanings and relationships of phrases and sentences, permitting richer comparison of relationships between phrases and not just individual words (Miller et al., 2020). This is important, since short sentence descriptors are useful for attribute definitions. Narayana et al. (2019) used BERT embeddings for their word or text representations, and also calculated the TF-IDF of text features to determine the importance for each word in the corpus. This work adopts a similar approach and relies on BERT representations for the surgical tool attributes and text used in information integration in the model.

Attributes are useful as knowledge representation because they can potentially offer good semantic expression, can be tightly and succinctly defined, and can be

class	tool_type	tool_shape	tool_features	tool_size
2.0mm_Titanium_Straight_Plate_12_Hole_71mm	straight_plate_titanium	2.0mm	12_hole	71mm
2.4mm_Titanium_T-Plate_2_Holex8_Hole_54mm	T-Plate_titanium	2.4mm_by_54mm	2_hole	8_hole
6_Mayo_Needle_Holder	needle_holder	Mayo	ring_and_lock	6_inch_gold_handles
7_Metzenbaum_Scissors	delicate_tissue_scissors	Metzenbaum	ring_no_lock	7_inch_tapered_tip
8_Babcock_Tissue_Forceps	tissue_forceps	Babcock	ring_and_lock	8_inch_ring_tip
9_DeBakey_Needle_Holder	needle_holder	Debakey	thumb	9_inch_tapered_tip
Allis_Tissue_Forceps	tissue_forceps	Allis	ring_and_lock	narrow_gap_tip
Ball_&_Socket_Towel_Clips	towel_clips	Ball_and_socket	ring_and_lock	ring_with_ball
Bearing_Plates_3_Hole_7_Peg_Right_N	bearing_plate	3_hole	7_peg	right
Clavicle_Plate_3.5-2.7_5_Hole_Right	clavicle_plate	right	5_hole	3.5_2.7mm
Clavicle_Plate_Medial_8_Hole	clavicle_plate	medial	8_hole	rounded_ends
Crile_Artery_Forceps	artery_forceps	Crile	ring_and_lock	curved_pointed_tip
Dressing_Scissors	general_cutting	Dressing	ring_no_lock	blunt
Fixed_Angle_Plates_3_Hole_7_Peg_Left	fixed_angle_plate	3_hole	7_peg	left
Lahey_Forceps	haemostat	Lahey	ring_and_lock	touching_tip
Littlewood_Tissue_Forceps	tissue_forceps	Littlewood	ring_and_lock	oval_tip
Mayo_Artery_Forceps	artery_forceps	Mayo	ring_and_lock	touching_tip
Multi-pin-clamp-grey-orange	multi_pin_clamp	grey_orange	rectangle_block	4_holes
Pin-to-rod-coupling-grey-orange	pin_to_rod_coupling	grey_orange	pin_holder	rod_holder

FIGURE 8.6: Example of Processed Text for Bert

structured to be less noisy and less ambiguous than plain text sentences. Attributes have also been shown to be able to capture relationships between classes to some extent (Chen et al., 2021b). However, defining attributes in a surgical domain requires considerable expertise to ensure succinctness, accuracy, distinctiveness and reliability. As pointed out by Akata et al. (2015), annotation with large numbers of very specific attributes may lead to better predictive performance but is an expensive and time consuming data gathering task. The current annotations are used as a proof of concept, but it is clear that better attribute definitions are needed for better outcomes. This will be addressed in future work but the current experiments highlight the utility of such information in the training of a CNN.

Knowledge graphs have been stated to be more expressive than text and attributes, and therefore this section explores this further. In relevant work using knowledge graphs, Timofeev et al. (2020) developed a neural graph learning framework that effectively used a knowledge graph structure to regularize training of CNNs. In their structure, an image was treated as a "vertex" and pair of images was treated as an "edge", and the resulting graph was used in the machine learning pipeline. Narayana et al. (2019) used a similar approach to create a semantic graph of classes — where each class was a vertex of the graph and two classes were connected by an edge. The cosine distance between two class embeddings was treated as edge weights, and two semantically similar classes had a lower edge weight compared to two semantically different classes. The research work in this section incorporated a semantic graph in the training of the CNN model, in a similar manner to Narayana et al. (2019) but used distance between attributes in the form of an adjacency matrix that effectively captured the relevant semantic information.

8.4 Methods

The experiments used ResNet50 trained on ImageNet to extract image embeddings. The ResNet50 model was not fine-tuned on our HospiTools dataset as in previous experiments, but preliminary work ensured that good predictive performance was

text	classes	processed_text	clean_classes	mapped_classes
needle_holder	9_DeBakey_Needle_Holder	[CLS] needle holder [SEP]	[CLS] 9 debakey needle holder [SEP]	5
ring_no_lock	7_Metzenbaum_Scissors	[CLS] ring no lock [SEP]	[CLS] 7 metzenbaum scissors [SEP]	3
ring_and_lock	6_Mayo_Needle_Holder	[CLS] ring lock [SEP]	[CLS] 6 mayo needle holder [SEP]	2
ring_and_lock	8_Babcock_Tissue_Forceps	[CLS] ring lock [SEP]	[CLS] 8 babcock tissue forceps [SEP]	4
71mm	2.0mm_Titanium_Straight_Plate_12_Hole_71mm	[CLS] 71mm [SEP]	[CLS] 2.0mm titanium straight plate 12 hole 71mm [SEP]	0
narrow_gap_tip	Allis_Tissue_Forceps	[CLS] narrow gap tip [SEP]	[CLS] allis tissue forceps [SEP]	6
Ball_and_socket	Ball_&_Socket_Towel_Clips	[CLS] ball socket [SEP]	[CLS] ball socket towel clips [SEP]	7
grey_orange	multi-pin-clamp-grey-orange	[CLS] grey orange [SEP]	[CLS] multi pin clamp grey orange [SEP]	17
Babcock	8_Babcock_Tissue_Forceps	[CLS] babcock [SEP]	[CLS] 8 babcock tissue forceps [SEP]	4

FIGURE 8.7: Attribute Matrix Examples

achieved using a standard ResNet50 trained on ImageNet. Further work was then conducted to evaluate improvements, if any, in predictive performance by incorporation of prior information in the machine learning pipeline. Text attributes (as per the example in Table 8.6) and an attribute-based semantic graph — developed in the form of an adjacency matrix — were used, which relied on BERT embeddings for word and sentence representations. The attribute text representation was cleaned, underscores with blank spaces were replaced, irrelevant words, hyphens, slashes and apostrophes were checked for and removed, and starting and ending tags were added — [CLS] and [SEP] — as required for BERT processing (Miller et al., 2020). Relevant details such as numbers and special names were retained. The cleaned text was mapped to identity numbers, and an example of the final processed matrix is presented in Figure 8.7. A BERT Tokenizer was used to convert word segment tokens to identities, and the last 4 layers from the resultant 12 layers were encoded and extracted. As recommended by (Devlin et al., 2019) and Narayana et al. (2019), the embeddings from the last four layers for each word were concatenated, and the word embeddings were averaged. The model that was developed used the attribute embeddings along with image embeddings as training input, and predictions were based off these two modalities.

In a second use of the attribute information, distance information was encoded in a adjacency matrix (also referred to in the literature as a semantic graph) and a distance regularisation loss was used in a set of experiments (Dash et al., 2022; Takeishi and Akimoto, 2018). The attribute embeddings extracted using BERT were used to construct an adjacency matrix or semantic graph. In this graph, each attribute embedding was a graph vertex and the cosine distance between two attribute embeddings was used as the edge weights (Narayana et al., 2019). The cosine distance values between the attributes was used to develop an adjacency matrix, which was used as a reference during training. The adjacency matrix A_{ij} captures the non-negative weights associated with each attribute as the cosine distance to the other attributes in the graph. Visualised as a semantic graph, there were 56 vertices corresponding to 56 different attributes in the matrix. The adjacency matrix is used as a regularization or graph loss function designed to force semantically similar attributes to be closer

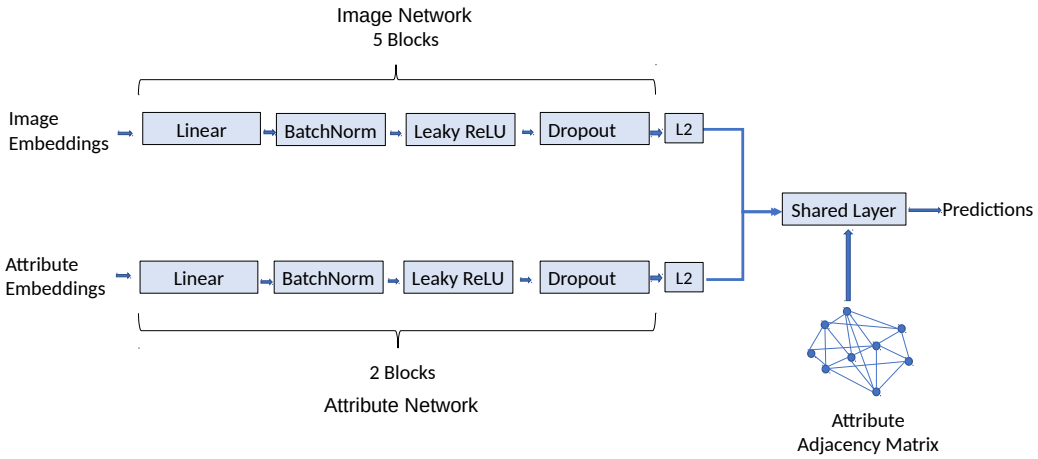


FIGURE 8.8: Prediction Pipeline Network and Shared Layer Architecture

to each other. As in previous experiments, the embeddings were L2 normalized to restrict the universal embedding space to the unit sphere.

Both image and attribute embeddings were relied on for the predictions pipeline. The network was trained on a set of 19 classes from the HospiTools dataset, with on average 95 images per class. General Surgery and Orthopaedics were represented in the classes, and a cross-section of classes from each speciality ere used. The model was tested on a hold-out set of images from the 19 classes. The model was also tested on another set of 25 classes that included both known classes that the model had seen during training, and unknown classes that had not been used in training. This was done to evaluate how the model would cope in Zero Shot Learning conditions, and what attributes could be relevant in predictions.

A ResNet50 model with ImageNet weights was used to obtain the image embeddings and BERT was used to obtain the attribute embeddings. There were two separate networks in the model, implemented in PyTorch — an image network with five blocks and an attribute network with two blocks. A block was formed by a fully connected hidden layer of 512 units, followed by a batch normalisation, a leaky-ReLU and a dropout layer. The image embeddings were resized from 2048 to 512 units and were processed by the image network. The attribute embeddings were resized from 3072 to 512 units using the first linear layers, and processed by the attribute network. The outputs from the respective network were L2 normalised, and classified by a shared fully connected classification layer (Figure 8.8. This architecture was based on work by Wang, Li, and Lazebnik (2016), Narayana et al. (2019) and Miller et al. (2020), and was designed to map image and text embeddings to a shared embedding space. The L2 normalized embeddings from both the images and attributes were classified by this shared hidden layer to provide predictions. The work employed the late fusion approach for multi-modal classification where the classification scores from image and attribute modalities are added based on user-defined weights.

Dash et al. (2022) refer to the incorporation of penalty terms or constraints into the loss function as a standard way of incorporating domain-knowledge into a deep network. This approach was used, and experiments with the use of the following

four losses in the machine learning pipeline were conducted:

- Categorical Cross Entropy Loss : This uses the categorical cross entropy loss with softmax activation to minimise the distance between predictions and labels, to ensure class level similarity, as in previously reported work.

$$\mathcal{L}_{CCE} = -\sum_{i=1}^{k} c_i \log\left(\hat{c}_i\right) \tag{8.1}$$

  In Equation 8.1, $\hat{c}_i$ represents the probability score for class c_i.

- Semantic Similarity Loss: This loss was designed so that embeddings of semantically similar classes were closer than embeddings from semantically different classes (Timofeev et al., 2020). The regularization ensured that the distance between any two embeddings was equal to the edge weight or cosine distance of their corresponding classes in the adjacency matrix. The Mean Square Error or MSE was used for this loss, and the MSE loss was applied on the distance between two embeddings and the cosine distance of their corresponding classes, as under:

$$\mathcal{L}_{SEM} = \frac{1}{k}\sum_{i=1}^{k} \sigma_{ij}\left(\psi\left(c_i\right)^{\top} \varphi\left(c_j\right) - A_{ij}\right)^2 \tag{8.2}$$

  In Equation 8.2, φ and ψ defined the embedding functions for each class, and $^{\top}$ was the dot product. The adjacency matrix of semantic class distances was represented by A_{ij}. A σ was also used, based on a margin which was set to 1 if both A_{ij} and $\psi\left(c_i\right)^{\top} \varphi\left(c_j\right) <$ margin, else $\sigma = 0$. The margin was set at 0.7 in the experiments, so that the loss function was only applied when the cosine distance was less than 0.7 for any two classes or embeddings, and embeddings that were far apart were ignored.

- Center-Loss: A Center-Loss was used between attribute and label embeddings, and also between image and label embeddings. The center-loss was designed to increase the separation of classes while minimizing the distances between embeddings from the same class (Wen et al., 2016).This loss had been used in previous work, reported in Chapter 6, and had achieved good results. The Center-Loss was defined by the following:

$$\mathcal{L}_{center-loss} = \frac{1}{2}\sum_{i=1}^{k} \left\|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\right\|_2^2 \tag{8.3}$$

  In Equation 8.3, $\boldsymbol{x}_i$ represents the center of the i$^{\text{th}}$ class and $\boldsymbol{c}_{y_i}$ the attribute embeddings for the class.

- Multi-Modal Gap Loss: This loss was designed to ensure that image and attribute embeddings from the same class are pulled closer together, or to minimise the distance. This addresses the issue around the modal gap, as discussed earlier.

$$\mathcal{L}_{CORR} = \frac{1}{k}\sum_{i=1}^{k}\left(1 - \psi\left(I_i\right)^{\top} \varphi\left(c_{y_i}\right)\right) \tag{8.4}$$

In Equation 8.4, φ defined the attribute embedding function, ψ the embedding function for image I, and $^\top$ was the dot product.

In these loss functions, the categorical cross entropy loss used predictions and labels, semantic similarity loss used the adjacency matrix, center-loss used either the image or attribute embeddings and the corresponding labels, and the multi-modal gap loss used the image and attribute embeddings in their functions. After experimenting with different losses and loss weights, a final loss based on a combination of three of the above losses was used. This final loss used categorical cross entropy loss with softmax, semantic similarity loss and multi-modal gap loss with specific weights provided to each loss, as under:

$$\mathcal{L}_{Final} = \omega_{cce}\mathcal{L}_{CCE} + \omega_{sem}\mathcal{L}_{SEM} + \omega_{corr}\mathcal{L}_{CORR} \tag{8.5}$$

After experimentation with different values, the weights were set at 0.4 for categorical cross entropy loss with softmax, and at 0.3 for the semantic similarity loss and multi-modal gap loss. A dropout rate of 0.15 was implemented and the network was trained for 60 epochs using RMSProp optimizer. The learning rate was set at 0.001 and momentum set to 0.9, with a batch size of 20 — the final training parameters were set based on prior work by Narayana et al. (2019) and Wang, Li, and Lazebnik (2016), and also based on wide ranging experiments with different parameters.

8.5 Results

Good benchmark results were obtained using only images and labels for training (Table8.1, first column) using the categorical cross entropy loss with softmax. It was noted that there was an improvement in predictive performance using information infusion with the attribute pipeline and multi-modal gap loss added. This further improved when attribute graph regularisation and semantic similarity loss was added. Overall, there was a significant improvement in accuracy and precision performance when late addition fusion was used along with the incorporation of attribute information in the training process. In this case, equal weights were assigned to the mono-modal predictions from image and text, and outputs were added to drive a final prediction. This demonstrates the effectiveness of including prior domain knowledge in the machine learning pipeline.

The work conducted include experiments with concatenation of the mono-modal features before classification instead of addition, as recommended by Miller et al. (2020), but accuracy dropped to 74% on the test set. This work therefore used addition of mono-modal features as recommended by Narayana et al. (2019) in its final predictions. Center-Loss did not improve performance over the benchmark results, with a drop to 69% in accuracy on the test set. It was also noted that adding more attributes — many of which were common across classes such as the speciality, pack and set information — significantly degraded the performance, with a drop to 55% accuracy on the test set. The attributes did not discriminate or define distinctiveness across classes to a sufficient extent, and this resulted in a reduction in performance of the system. This highlights the importance of using distinctive, specific and precise annotations in the knowledge pipeline, even if collecting such data is an expensive proposition (Akata et al., 2015). In this case, less is actually more — introducing additional large numbers of annotations are not as relevant as ensuring a small set of distinctive, precise and specific annotations.

TABLE 8.1: Results - Informed Training

Metric	Image Only	Image and Attributes (Text)	Image, Attributes and Attribute Semantic Graph
Test Set:			
Accuracy	0.70	0.77	0.81
Hamming Loss	0.29	0.23	0.19
F1 Score	0.71	0.79	0.81
Precision	0.72	0.79	0.82
Recall	0.84	0.89	0.87
Known-Unknown Test Set:			
Accuracy	0.30	0.35	0.34
Hamming Loss	0.69	0.64	0.65
F1 Score	0.28	0.30	0.31
Precision	0.30	0.35	0.34
Recall	0.29	0.29	0.32

8.6 Discussion and Future Work

This was a proof of concept for "Informed Machine Learning", where prior domain knowledge is included in the machine learning pipeline. The results demonstrate that prior information can be gainfully used to improve predictive performance of the CNN, but that using the correct set of attributes is a critical issue. This was highlighted by experiments where accuracy dropped when additional attributes were introduced, because these attributes were not discriminative and precise, and therefore introduced noise and ambiguity in the system. The experiments conducted highlighted the utility of such information and also a need to further experiment with techniques to create a relevant, semantically meaningful corpus and set of attributes that can enrich any surgical tool management system. While a start has been made in this research to create a surgical tool description textual dataset, further improvement that were planned as part of the experiments were not possible due to the on-going crisis with the COVID-19 pandemic. This unfortunately led to the cancellation of planned workshops and meetings with health professionals to improve and refine the annotations and textual descriptions of surgical tools. Additional work is therefore needed to define an more comprehensive set of annotations that can be gainfully used by the informed machine learning system, and which can potentially be used to train a Surgical-BERT model for improved performance. While currently only images are used to make a prediction, with an improved set of annotations, text can also be used to predict or identify details of a particular surgical tool. Future work will incorporate knowledge infusion techniques using the adjacency matrix and regularisation functions into the OctopusNet model.

Chapter 9

Conclusions and Future Work

This chapter presents the thesis conclusions and plans for future work for the intelligent management of surgical tools. It demonstrates how the research questions have been addressed, reports problems and shortcomings, and how the research hypothesis has been resolved. The section provides directions for future work.

9.1 The Hypothesis and Research Questions

The research proposal was to develop an applied deep learning system that could recognise and classify surgical tools. The research hypothesis was as follows:

> A hierarchical, informed, robust machine learning based system can be developed for effective management of surgical tools.

9.1.1 Research Questions

The specific research questions are enumerated below, and the work that specifically addresses each question is detailed:

- RQ1 – How can a convolutional neural network be designed for recognition of surgical tools, effectively utilising the hierarchical nature of surgical tool classes? This question was addressed in Chapters 4 and 5.
- RQ2 – How can the design of a CNN be improved for interpretable deep learning for intelligent surgical tool management, by incorporating prior information and knowledge of relationships in the ground truth class label arrangements? This strategy was addressed in Chapter 6.
- RQ3 – How can the robustness of a CNN be improved for recognition of surgical tools under challenging conditions, addressing volume, variety, complexity and illumination / reflection / occlusion issues? This problem was addressed in Chapter 7.
- RQ4 –How can nominal attribute information be included in a Machine Learning model to improve the predictions of a CNN for surgical tool management? This issue was addressed in Chapter 8.

To address the research questions, the thesis describes work conducted to develop a computer vision and deep learning system that could recognise and classify surgical tools. The system needed to cope with a wide variety of tools, with very subtle differences in shapes. It had to work with high volumes, as well as varying illuminations

and backgrounds. Methodology that was adopted included the creation of a surgical tool image dataset, development of a surgery knowledge-base, training CNNs to recognise surgical tools, integration of CNNs with prior knowledge, and deployment of a prototype system. State of the art techniques were developed to cope with volume, variety and vision problems, and algorithms were designed and adapted to address specific surgery tool recognition issues. The system needed to be robust, and synthetic data was used to increase robustness of the network to illumination and background changes. Prior knowledge and information was also relied on to improve predictive performance of a CNN. The specific thesis contributions are detailed in the next section.

9.1.2 Thesis Contributions

- This research proposal focused on designing a deep learning surgical tools system that could perform effectively in the medical domain.
- New datasets and algorithms were developed that addressed management of surgical tools.
- A prototype was developed for computer vision based intelligent management of surgical tools to demonstrate the effectiveness, efficiency and accuracy of the system – however this could not be tested in real world conditions.
- Domain Knowledge was developed in the form of a Surgical Tool dataset and a Surgery Knowledge Base, this was made open source to support further research in this area.
- New architectures were designed to provide rich and relevant information to the end-user, in the form of multiple, hierarchical predictions.
- The dataset was enhanced with synthetic data, using a range of techniques, and the CNN was trained with this synthetically augmented dataset, with improved predictive performance and robustness.
- Experiments with incorporation of prior domain knowledge and information in the training of a CNN were conducted, with good results.
- A contribution is a deeper insight into how to design and train real world deep learning systems for practical applications. The results of this research may be useful for hospitals, District Health Boards and the government, and can contribute to greater efficiencies and cost savings across multiple organisations.

Given the mission critical nature of surgical tool management in a hospital, the actual deployment, integration, adoption and testing of such systems in actual hospital conditions – or MLOps – needs to be addressed (Makinen et al., 2021). The survey conducted and reported in this thesis has shown that most available surgical tool datasets are small sized, specific to a surgical procedure type, privately generated and maintained (Rieke et al., 2020) – future work will evaluate federated learning (Yang et al., 2019b; Zhang et al., 2021) as a strategy to increase access to smaller and private datasets while managing issues of security and ownership. This approach will allow private, smaller datasets to be used for local training, prior to aggregation at a central point for deep learning. Zhang et al. (2021) show how distributed training across data islands address privacy concerns while allowing access to data for ML models. Future work will therefore evaluate the use of federated learning and MLOps for intelligent surgical tool management.

9.2 Conclusions

This thesis addresses a novel issue – the intelligent management of surgical tools in a hospital using deep learning methods. The thesis work conceptualises a radically new way of approaching a fundamental issue, replacing inefficient manual systems with transformative state of the art technology. While the end result is potentially revolutionary, the development path needs to be evolutionary, particularly given the risks inherent in the medical domain. Implementation of this system within the New Zealand health system can realise significant savings in the short term through better efficiencies and lower losses, patient and surgery safety can be improved, and many millions of dollars can be saved in the long term through better inventory management. While this thesis has made a start in development of a hierarchical, robust and informed surgical management system using machine learning, it is acknowledged that much more work needs to be done to implement the system in real world and critical medical conditions.

Bibliography

ACS (2021). *What are the surgical specialties?* https://www.facs.org/ education/resources/ medical-students/faq/specialties. Accessed: 15/2/2021.

Ahmadi, E. et al. (2018). "Inventory management of surgical supplies and sterile instruments in hospitals: a literature review". In: *Health Systems* 2018.8, pp. 134–151. DOI: 10.1080/20476965.2018.1496875.

Akata, Zeynep et al. (2015). "Evaluation of Output Embeddings for Fine-Grained Image Classification." In: *IEEE Computer Vision and Pattern Recognition*. DOI: 10.1109/CVPR.2015.7298911.

Al Hajj, H., M. Lamard, P. H. Conze, et al. (2019). "CATARACTS: Challenge on automatic tool annotation for cataRACT surgery". In: *Medical Image Analysis* 52, pp. 24–41. DOI: 10.1016/j.media.2018.11.008.

Al Hajj, H. et al. (2018). "Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks". In: *Medical Image Analysis* 47, pp. 203–218.

Alfred, Myrtede et al. (2021). "Work systems analysis of sterile processing: assembly". In: *BMJ Quality Safety*. DOI: 10.1136/bmjqs-2019-010740.

Ali, Sharib, Mariia Dmitrieva, and et al. (May 2021). "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy". In: *Medical Image Analysis* 70. DOI: 10.1016/j.media.2021.102002.

Ali Qadir, Hemin et al. (2019). "Polyp Detection and Segmentation using Mask R-CNN: Does a Deeper Feature Extractor CNN Always Perform Better?" In: *13th International Symposium on Medical Information and Communication Technology (ISMICT)*. DOI: 10.1109/ISMICT.2019.8743694.

Allan, M. et al. (2019). "2017 Robotic Instrument Segmentation Challenge". In: *arXiv:1902.06426*.

Allan, M. et al. (2020). "2018 Robotic Scene Segmentation Challenge". In: *ArXiv abs/2001.11190*.

Almubarak, Haidar, Yakoub Bazi, and Naif. Alajlan (2020). "Two-Stage Mask-RCNN Approach for Detecting and Segmenting the Optic Nerve Head, Optic Disc, and Optic Cup in Fundus Images". In: *Applied Sciences*. DOI: 10.3390/app10113833..

Alshirbaji, T. A., N. A. Jalal, and K. Moller (2018). "Surgical tool classification in laparoscopic videos using convolutional neural network". In: *Current Directions in Biomedical Engineering* 4.1, pp. 407–410.

Alshirbaji, T. A. et al. (2020a). "The Effect of Background Pattern on Training a Deep Convolutional Neural Network for Surgical Tool Detection". In: *Proceedings on Automation in Medical Engineering* 1.1, pp. 24–024.

Alshirbaji, T. Abdulbaki et al. (2020b). "The effect of background pattern on training a deep convolutional neural network for surgical tool detection". In: *AUTOMED - Automation in Medical Engineering*.

Alshirbaji, T. Abdulbaki et al. (2021a). "Cross-dataset evaluation of a CNN-based approach for surgical tool detection". In: *AUTOMED 2021*.

Alshirbaji, Tamer Abdulbaki et al. (2021b). "Assessing Generalisation Capabilities of CNN Models for Surgical Tool Classification". In: *Current Directions in Biomedical Engineering*.

Amsterdam, B. van, M. Clarkson, and D. Stoyanov (2021). "Gesture Recognition in Robotic Surgery: a Review". In: *IEEE Transactions on Biomedical Engineering*.

Andersen, Jakob Kristian Holm, Kim Lindberg Schwaner, and Thiusius Rajeeth Savarimuthu (2021). "Real-Time Segmentation of Surgical Tools and Needle Using a Mobile-U-Net". In: *20th International Conference on Advanced Robotics (ICAR)*.

Attanasio, Aleks, Bruno Scaglioni, and et al. (Oct. 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. DOI: 10.1109/LRA.2020.3013914.

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

Bar, O., D. Neimark, and et al. (2020). "Impact of data on generalization of AI for surgical intelligence applications". In: *Scientific Reports* 10. DOI: 10.1038/s41598-020-79173-6.

Barbu, A. et al. (2019). "ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models". In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Barros Barbosa, Igor et al. (2017). "Looking Beyond Appearances: Synthetic Training Data for Deep CNNs in Re-identification". In: *Computer Vision and Image Understanding*. DOI: 10.1016/j.cviu.2017.12.002.

Barz, Bjorn and Joachim Denzler (2019). "Hierarchy-based image embeddings for semantic image retrieval". In: *In IEEE Winter Conference on Applications of Computer Vision (WACV)*. DOI: 10.1109/WACV.2019.00073.

Barz, Bjorn and Joachim Denzler (2020). "Deep Learning on Small Datasets without Pre-Training using Cosine Loss." In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. DOI: 10.1109/WACV45572.2020.9093286.

Bayoudh, Khaled et al. (2022). "A survey on deep multimodal learning for computer vision: advances,trends, applications, and datasets". In: *The Visual Computer*. DOI: 10.1007/s00371-021-02166-7.

Bhatt, Nikita et al. (2018). "Trends in the Use of Laparoscopic Versus Open Paediatric Appendicectomy: A Regional 12-Year Study and a National Survey". In: *World Journal of Surgery 42*.

Bodenstedt, S. et al. (2018). "Real-time image-based instrument classification for laparoscopic surgery". arXiv preprint. arXiv: 1808.00178.

Bouget, D. et al. (2015). "Detecting surgical tools by modelling local appearance and global shape". In: *IEEE transactions on medical imaging* 34.12, pp. 2603–2617.

Bouget, D. et al. (2017). "Vision-based and marker-less surgical tool detection and tracking: a review of the literature". In: *Medical Image Analysis* 35, p. 633.

Brust, Clemens-Alexander and Joachim Denzler (2019). "Not just a matter of semantics: the relationship between visual similarity and semantic similarity". In: *Lecture Notes in Computer Science, vol 11824*. DOI: 10.1007/978-3-030-33676-9_29.

Cao, Zhantao et al. (2019). "An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures." In: *BMC Medical Imaging*. DOI: 10.1186/s12880-019-0349-x..

Ceron, Juan Carlos Angeles et al. (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, Junyi et al. (Aug. 2021). "Deep learning in computer vision: A critical review of emerging techniques and application scenarios". In: *Machine Learning with Applications* 6. DOI: 10.1016/j.mlwa.2021.100134.

Chang, Jia-Ren and Yong-Sheng. Chen (2018). "Pyramid Stereo Matching Network". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5410–5418.

Chen, Hongyu et al. (2021a). "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, Jiaoyan et al. (2021b). "Knowledge-aware Zero-Shot Learning: Survey and Perspective". In: *Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*.

Chen, Liang-Chieh et al. (2016). "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs". In: *arXiv preprint*.

Chen, Liang-Chieh et al. (2017). "Rethinking Atrous Convolution for Semantic Image Segmentation". In: *arXiv:1706.05587v3*.

Chen, Z., Z. Zhao, and X.Oct Cheng (2017). "Surgical instruments tracking based on deep learning with lines detection and spatio-temporal context". In: *In: Proc IEEE CAC. Jinan, China, pp* 2711.

Choi, B. et al. (2017). "Surgical-tools Detection based on Convolutional Neural Network in Laparoscopic Robot-assisted Surgery". In: *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Ed. by New York (ny): Ieee. p. 1756–1759.

Choi, Joonmyeong et al. (2021). "Video Recognition of Simple Mastoidectomy Using Convolutional Neural Nets: Detection and Segmentation of Surgical Tools and Anatomic Regions". In: *Computer Methods and Programs in Biomedicine*. DOI: 10.1016/j.cmpb.2021.106251.

Ciaparrone, G. et al. (2020). "A comparative analysis of multi-backbone Mask R-CNN for surgical tools detection". In: *International Joint Conference on Neural Networks (IJCNN)*. DOI: 10.1109/IJCNN48605.2020.9206854.

Colleoni, E. et al. (2019). "Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers". In: *IEEE Robotics and Automation Letters* 4.3, pp. 2714–2721.

Colleoni, Emanuele, Philip Edwards, and Danail Stoyanov (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)*. DOI: 10.1007/978-3-030-59716-0_67.

Dapello, Joel et al. (2020). "Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations". In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Dash, T. et al. (2022). "A review of some techniques for inclusion of domain-knowledge into deep neural networks". In: *Scientific Reports*. DOI: 10.1038/s41598-021-04590-0.

Demirel, B., R. Cinbis, and N. Ikizler (2017). "Attributes2Classname: A Discriminative Model for Attribute-Based Unsupervised Zero-Shot Learning". In: *ICCV.2017.139* 10.1109, pp. 1241–1250.

Deng, Changyu et al. (2020). "Integrating Machine Learning with Human Knowledge". In: *iScience*. DOI: 10.1016/j.isci.2020.101656.

Deng, J., A. Berg, and L. Fei-Fei (2011). "Hierarchical semantic indexing for large scale image retrieval". In: *CVPR 2011, 785-792*. DOI: 10.1109/CVPR.2011.5995516.

Dergachyova, O. et al. (2016). "Automatic data-driven real-time segmentation and recognition of surgical workflow". In: *International Journal of Computer Assisted Radiology and Surgery* 11.6, pp. 1081–1089.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding". In: *ArXiv abs/1810.04805*.

Du, X. et al. (2018). "Articulated multi-instrument 2-D pose estimation using fully convolutional networks". In: *IEEE Trans Med Imaging* 37, p. 5.

Egger, Jan et al. (2020). "Medical Deep Learning – A systematic Meta-Review". In: *ArXiv abs/2010.14881 (2020)*.

Ferreira, Beatriz et al. (2018). "A unified model with structured output for fashion images classification". In: *AI for Fashion - The third international workshop on Fashion and KDD, London, United Kingdom*, pp. 1–10.

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

Frome, A. et al. (2013). "Devise: A deep visual-semantic embedding model". In: *Advances in Neural Information Processing Systems (NIPS)*.

Gallo, Ignazio, Alessandro Calefati, and Shah Nawaz (2017). "Multimodal Classification Fusion in Real-World Scenarios". In: *International Conference on Document Analysis (ICDAR)*. DOI: 10.1109/ICDAR.2017.326.

Gao, Y et al. (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*.

Garcez, A. et al. (2019). "Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning". In: *FLAP, Vol 6*.

Garcia-Peraza-Herrera, L. et al. (2017). "ToolNet: Holistically-nested real-time segmentation of robotic surgical tools". In: *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on], 5717–5722, IEEE Vancouver,* pp: Canada.

Garcia-Peraza-Herrera, Luis et al. (2021a). "Image Compositing for Segmentation of Surgical Tools Without Manual Annotations". In: *IEEE Transactions on Medical Imaging*.

Garcia-Peraza-Herrera, Luis et al. (2021b). "Image Compositing for Segmentation of Surgical Tools Without Manual Annotations". In: *IEEE transactions on medical imaging*. DOI: 10.1109/TMI.2021.3057884.

Garrow, Carly R. et al. (2021). "Machine Learning for Surgical Phase Recognition: A Systematic Review". In: *Annals of Surgery*.

Gessert, N., M. Schlüter, and A. Schlaefer (2018). "A deep learning approach for pose estimation from volumetric OCT data". In: *Medical image analysis* 46, pp. 162–179.

Girshick, R. (2015). "Fast R-CNN". In: *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448. DOI: 10.1109/ICCV.2015.169..

Girshick, R., J. Donahue, T. Darrell, et al. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Computer Vision and Pattern Recognition. Washington (DC)*. Ed. by S. Fidler and Mortensen E. IEEE Computer Society; p. 580–587.

Gong, Yunchao et al. (2013). "A Multi-View Embedding Space for Modeling Internet Images, Tags, and Their Semantics". In: *International Journal of Computer Vision* 106, pp. 210–233.

Gonzalez, C., L. Bravo-Sanchez, and P. Arbelaez (July 2020). "Isinet: An instance-based approach for surgical instrument segmentation". In: *Medical Image Computing and Computer Assisted Intervention MICCAI 2020*. DOI: 10.1007/978-3-030-59716-0_57.

Grammatikopoulou, M. et al. (2019). "Cadis: Cataract dataset for image segmentation". In: *arXiv:1906.11586*.

Gruijthuijsen, Caspar, Luis Garcia-Peraza-Herrera, and et al. (2021). "Robotic Endoscope Control via Autonomous Instrument Tracking". In: *arXiv:2107.02317*.

Guedon, A.C. et al. (2016). "Where are my instruments? Hazards in delivery of surgical instruments." In: *Surgical endoscopy, 30(7).*

Guo, Yanming et al. (Apr. 2016). "Deep learning for visual understanding: A review". In: *Neurocomputing* 187, pp. 27–48. DOI: 10.1016/j.neucom.2015.09.116.

Hasan, Md. Kamrul et al. (Feb. 2021). "Detection, segmentation, and 3D pose estimation of surgical tools using convolutional neural networks and algebraic geometry". In: *Medical Image Analysis* 70(4). DOI: 10.1016/j.media.2021.101994.

Hasan, S. K. and C. A. Linte (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) (2019)*.

He, K., X. Zhang, S. Ren, et al. (2016). "Deep residual learning for image recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition. Washington (DC)*. IEEE Computer Society, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

He, K. et al. (2017). "Mask R-CNN". In: *Conference on Computer Vision (ICCV)*. Ed. by Ieee International. pp. 2961–2969.

Hein, Eric et al. (July 2018). "Large-scale medical image annotation with crowd-powered algorithms". In: *Journal of Medical Imaging* 5 (3). DOI: 10.1117/1.JMI.5.3.034002.

Hiasa, Yuta et al. (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*.

Hinterstoisser, Stefan et al. (2019). "An Annotation Saved is an Annotation Earned: Using Fully Synthetic Training for Object Detection". In: *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. DOI: 10.1109/ICCVW.2019.00340.

Hoehndorf, R. and N. Queralt-Rosinach (2017). "Data Science and symbolic AI: Synergies, challenges and opportunities". In: *Data Science* 10.3233, pp. 1–12.

Hong, W.-Y. et al. (2020). "CholecSeg8k: A Semantic Segmentation Dataset for Laparoscopic Cholecystectomy Based on Cholec80". In: *ArXiv abs/2012.12453*.

Hossain, M. et al. (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, Yaqing et al. (2022). "Adaptive kernel selection network with attention constraint for surgical instrument classification". In: *Neural Computing and Applications*. DOI: 10.1007/s00521-021-06368-x.

Hu, H. et al. (2016). "Learning Structured Inference Neural Networks with Label Relations". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 2960-2968*.

Hu, X. et al. (2017). "AGNet: Attention-guided network for surgical tool presence detection". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision sSpport. Lecture Notes in Computer Science. , Cham, pp. 186–194*.

Huang, G. et al. (2017). "Densely Connected Convolutional Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

Huang, K., J. Altosaar, and R. Ranganath (2019). "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission". In: *arXiv preprint arXiv:1904.05342*.

Huaulme, A., D. Sarikaya, and et al. (Oct. 2021). "MIcro-surgical anastomose workflow recognition challenge report". In: *Computer Methods and Programs in Biomedicine* 212. DOI: 10.1016/j.cmpb.2021.106452.

Huh, J. et al. (2018). "A Simple Method on Generating Synthetic Data for Training Real-time Object Detection Networks". In: *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1518–1522. DOI: 10.23919/APSIPA.2018.8659778.

Huynh, D. and E. Elhamifar (2020). "Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention". In: *CVPR42600. 2020. 00454* 10.1109, pp. 4482–4492.

Iandola, Forrest et al. (2016). "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size". In: *arxiv:1602.07360*.

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

Ilg, Eddy et al. (2017). "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks". In: *arXiv:1612.01925*.

Inoue, M., C. H. Forster, and A. Carlos dos Santos (2020). "Semantic Hierarchy-based Convolutional Neural Networks for Image Classification". In: *2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, United Kingdom, 2020, pp. 1-8*, pp. 1–8.

Isensee, F. and K. H. Maier-Hein (2020). "OR-UNet: an optimized robust residual u-net for instrument segmentation in endoscopic images." In: *arXiv*.

Islam, M., Y. Li, and H.. Ren (2019). "Learning where to look while tracking instruments in robot-assisted surgery". In: *ArXiv abs/1907.00214*. DOI: 10.1007/978-3-030-32254-0_46.

Islam, M. et al. (Jan. 2021). "ST-MTL: Spatio-temporal multitask learning model to predict scanpath while tracking instruments in robotic surgery". In: *Medical Image Analysis* 67. DOI: 10.1016/j.media.2020.101837.

Jha, Debesh et al. (2021a). "Exploring Deep Learning Methods for Real-Time Surgical Instrument Segmentation in Laparoscopy". In: *arXiv:2107.02319*.

Jha, Debesh et al. (2021b). "Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy". In: *Multi Media Modeling MMM2021 Lecture Notes in Computer Science, vol 12573 Springer, Cham*.

Jin, A. et al. (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*. Washington (DC). Lake Tahoe, 691–699.

Jin, Y., H. Li, Q. Dou, et al. (Jan. 2020). "Multi-task recurrent convolutional network with correlation loss for surgical video analysis". In: *Medical Image Analysis* 59, p. 1. DOI: 10.1016/j.media.2019.101572.

Jin, Y. et al. (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.

Jo, HyunJun, Yong-Ho Na, and Jae-Bok Song (2017). "Data augmentation using synthesized images for object detection". In: *17th International Conference on Control, Automation and Systems (ICCAS)*, pp. 1035–1038. DOI: 10.23919/ICCAS.2017.8204369.

Jo, K. et al. (2019). "Robust Real-Time Detection of Laparoscopic Instruments in Robot Surgery Using Convolutional Neural Networks with Motion Vector Prediction". In: *Applied Sciences* 9.14, p. 2865.

Johnson, J.W. (2018). "Adapting Mask-RCNN for Automatic Nucleus Segmentation". In: *ArXiv, abs/1805.00500.*

Kalavakonda, N. et al. (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. DOI: 10.1109/CVPRW.2019.00076.

Kanakatte, A. et al. (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, Will et al. (2017). "The Kinetics Human Action Video Dataset". In: *arXiv preprint*.

Kayhan, M. et al. (2019). "Deep attention based semi-supervised 2d-pose estimation for surgical instruments". In: *ArXiv, abs/191204618*.

Kletz, S., K. Schoeffmann, and H. Husslein (2019). "Learning the representation of instrument images in laparoscopy videos". In: *Healthcare Technology Letters* 6.6, pp. 197–203.

Kletz, S. Schoeffmann, K. Benois-Pineau J, and Husslein (2019). "Identifying Surgical Instruments in Laparoscopy Using Deep Learning Instance Segmentation". In: *International Conference on Content-Based Multimedia Indexing (CBMI)*. Dublin, Ireland, pp. 1–6.

Kohli, Marc D., Ronald M. Summers, and J. Raymond Geis (2017). "Medical Image Data and Datasets in the Era of Machine Learning – White paper from the 2016 C-MIMI Meeting Dataset Session". In: *Journal of Digital Imaging (2017) 30.*, 392–399. DOI: 10.1007/s10278-017-9976-3.

Kong, Xiaowen et al. (2021). "Accurate instance segmentation of surgical instruments in robotic surgery: model refinement and cross-dataset evaluation". In: *International journal of computer assisted radiology and surgery*. DOI: 10.1007/s11548-021-02438-6.

Koo, J., D. Klabjan, and J. Utken (2018). "Combined convolutional and recurrent neural networks for hierarchical classification of images". In: *arXiv 2018, arXiv:1809.09574*.

Kritzinger, W. et al. (2018). "Digital Twin in manufacturing: A categorical literature review and classification". In: *IFAC-PapersOnLine*.

Krizhevsky, A., I. Sutskever, and G. Hinton (2012). "ImageNet classification with deep convolutional neural networks". In: *Neural Information Processing Systems. Red Hook (NY): Curran Associates Inc.; p. 1097–1105* 2012.

Kugler, D. et al. (2020a). "iPosNet: Instrument Pose Estimation from X-Ray in temporal bone surgery". In: *Int J Comput Assist Radiol Surg. 15(7): 1137-1145.* 3.

Kugler, David et al. (2020b). "AutoSNAP: Automatically Learning NeuralArchitectures for Instrument Pose Estimation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, 23rd International Conference, Lima, Peru,*

Kumar, G. and P. K. Bhatia (2014). "A Detailed Review of Feature Extraction in Image Processing Systems". In: *2014 Fourth International Conference on Advanced Computing and Communication Technologies*, pp. 5–12. DOI: 10.1109/ACCT.2014.74..

Kurmann, T. et al. (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, Thomas et al. (2021). "Mask then classify: multi-instance segmentation for surgical instruments". In: *International Journal of Computer Assisted Radiology and Surgery*. DOI: 10.1007/s11548-021-02404-2.

Laina, I. et al. (2017). "Concurrent segmentation and localization for tracking of surgical instruments". In: *International conference on medical image computing and computer-assisted intervention, . pp. 664–672*.

Lampert, C. H., H. Nickisch, and S. Harmeling (2014). "Attribute-Based Classification for Zero-Shot Visual Object Categorization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 3, pp. 453-465*.

Lavado, Diana Martins (2018). "Sorting surgical tools from a cluttered tray - object detection and occlusion reasoning". MA thesis. University of Coimbra, Portugal.

Law, H. and J. Deng (2020). "CornerNet: Detecting Objects as Paired Keypoints". In: *Int J Comput Vis 128, 642–656*.

Law, H., K. Ghani, and J. Deng (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.

LeCun, Y., Y. Bengio, and G. Hinton (2015). "Deep Learning". In: *Nature.* 10.1038, pp. 436–44.

Lee, E. J. et al. (2019a). "Segmentation of surgical instruments in laparoscopic videos: training dataset generation and deep-learning-based framework". In: *In Medical Imaging Image-Guided Procedures, Robotic Interventions, and Modeling (Vol. 10951, p. 109511T). International Society for Optics and Photonics* 2019.

Lee, E. J. et al. (2019b). "Weakly supervised segmentation for real-time surgical tool tracking". In: *Healthcare Technology Letters* 6.6, pp. 231–236.

Leibetseder, A. et al. (2018). "Lapgyn4: A Dataset for 4 Automatic Content Analysis Problems in the Domain of Laparoscopic Gynecology". In: *Proceedings of the 9th ACM Multimedia Systems Conference. ACM, , NY, USA, pp 357–362*.

Leppanen, T. et al. (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. DOI: 10.1109/CBMS.2018.00044.

Li, Hanchao et al. (2018). "Pyramid Attention Network for Semantic Segmentation". In: *British Machine Vision Conference (BMVC), Newcastle upon Tyne*.

Li, Yikuan et al. (2021). "BEHRT: Transformer for Electronic Health Records". In: *Scientific Reports*.

Lin, T. Y. et al. (2014). "Microsoft coco: Common objects in context". In: *Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol 8693.* Ed. by D. Fleet et al. Springer, Cham.

Lin, T. Y. et al. (2017). "Focal loss for dense object detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence. PP. 1-1. 10.1109/TPAMI.2018.2858826.*

Lin, X. G. et al. (2019). "Presence Detection of Surgical Tool Via Densely Connected Convolutional Networks". In: *DEStech Transactions on Computer Science and Engineering*. 2019 International Conference on Artificial Intelligence and Computing Science (ICAICS 2019), pp. 245 –253.

Litjens, G. et al. (2017). "A survey on deep learning in medical image analysis". In: *Med Image Anal (Supplement C), 60–88* 42.

Liu, Li et al. (2020a). "Deep Learning for Generic Object Detection: A Survey". In: *International Journal of Computer Vision volume 128, pages 261–318*.

Liu, S. et al. (2020b). "Hyperbolic Visual Embedding Learning for Zero-Shot Recognition". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA*, pp. 9270–9278.

Liu, Wei et al. (2016). "SSD: Single Shot MultiBox Detector". In: *Computer Vision and Pattern Recognition*. arXiv: 1512.02325v5.

Liu, Y. et al. (2020d). "An Anchor-Free Convolutional Neural Network for Real-Time Surgical Tool Detection in Robot-Assisted Surgery". In: *IEEE Access*, pp. 78193–78201. DOI: 10.1109/ACCESS.2020.2989807.

Liu, Y. et al. (2020c). "An anchor-free convolutional neural network for real-time surgical tool detection in robot-assisted surgery". In: *IEEE Access–78201* 8, p. 78193.

Loncomilla, Patricio and Javier Ruiz-del Solar (2019). "YoloSPoC: Recognition of Multiple Object Instances by Using Yolo-Based Proposals and Deep SPoC-Based Descriptors." In: *RoboCup 2019: Robot World Cup XXIII*. DOI: 10.1007/978-3-030-35699-6_12.

Loshchilov, Ilya and Frank Hutter (2017). "SGDR: Stochastic Gradient Descent with Warm Restarts". In: *5th International Conference on Learning Representations*.

Lu, Jingpei et al. (2020). "SuPer Deep: A Surgical Perception Framework for Robotic Tissue Manipulation using Deep Learning for Feature Extraction." In: *ArXiv abs/2003.03472*.

Luengo, Imanol et al. (2021). "2020 CATARACTS Semantic Segmentation Challenge". In: *arXiv:2110.10965*.

Maier-Hein, L. et al. (2020). "Surgical Data Science - from Concepts to Clinical Translation". In: *ArXiv, abs/2011.02284*.

Maier-Hein, L. et al. (2021). "Heidelberg colorectal data set for surgical data science in the sensor operating room". In: *Scientific Data*.

Maier-Hein, Lena et al. (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. DOI: 10.1007/978-3-319-10470-6_55.

Makinen, Sasu et al. (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)*.

Manettas, Christos, Nikolaos Nikolakis, and Kosmas Alexopoulos (2021). "Synthetic datasets for Deep Learning in computer-vision assisted tasks in manufacturing". In: *Procedia CIRP*. DOI: 10.1016/j.procir.2021.10.038.

Marcus, G. (2020). "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence". In: *arXiv:2002.06177*.

Maron, Roman C. et al. (2021). "Robustness of convolutional neural networks in recognition of pigmented skin lesions". In: *European Journal of Cancer* 145, pp. 81–91.

Mathis, Alexander et al. (2018). "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning". In: *Nature Neuroscience volume 21, pages 1281–1289*.

Meeuwsen, F. C. et al. (2019). "Surgical phase modelling in minimal invasive surgery". In: *Surgical Endoscopy* 33.5, pp. 1426–1432.

Meireles, Ozanan R. et al. (Sept. 2021). "SAGES consensus recommendations on an annotation framework for surgical video". In: *Surgical Endoscopy* 35 (9). DOI: 10.1007/s00464-021-08578-9.

Meter, M. and R. Adam (2016). "Costs associated with Instrument sterilization in Gynecologic Surgery". In: *American Journal of Obstetrics and Gynecology* 215, 10/1016/j.ajog.2016.06.019.

Mhlaba, J. M. et al. (2015). "Surgical instrumentation: The true cost of instrument trays and a potential strategy for optimization". In: *Journal of Hospital Administration* 4, p. 6. DOI: 10.5430/jha.v4n6p82.

Mikolov, Tomas et al. (2013). "Distributed representations of words and phrases and their compositionality". In: *26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)*.

Miller, Stuart J. et al. (2020). "Multi-Modal Classification Using Images and Text". In: *SMU Data Science Review*.

Mishra, K., R. Sathish, and D. Sheet (2017). "Learning latent temporal connectionism of deep residual visual abstractions for identifying surgical tools in laparoscopy procedures". In: *IEEE Computer Society; 2017. p. 2233–2240*. Ed. by Mortensen E. DC).

Mohammed, Ahmed et al. (Mar. 2019). "StreoScenNet: surgical stereo robotic scene segmentation". In: *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*. SPIE Medical Imaging. San Diego, California, United States. DOI: 10.1117/12.2512518.

Mondal, S., R. Sathish, and D. Sheet (2019). "Multitask Learning of Temporal Connectionism in Convolutional Networks using a Joint Distribution Loss Function to Simultaneously Identify Tools and Phase in Surgical Videos". In: *ArXiv, abs/1905.08315*.

Motamedi, M. et al. (2020). "Octopus: Context-Aware CNN Inference for IoT Applications". In: *IEEE Embed. Syst. Lett. 12, 1 (March 2020), 1–4.*

Murillo, P. C. U., R. J. Moreno, and J. O. P. Arenas (2017). "Comparison between CNN and Haar classifiers for surgical instrumentation classification". In: *Contemporary Engineering Sciences* 10.28, pp. 1351–1363.

Murillo, P.C.U., J. O. P. Arenas, and R. J. Moreno (2018). "Tree-Structured CNN for the Classification of Surgical Instruments". In: *International Symposium on Intelligent Computing Systems*. 211–216.

N., Matton et al. (2022). "Analysis of cataract surgery instrument identification performance of convolutional and recurrent neuralnetwork ensembles leveraging BigCat". In: *Translational Vision Science and Technology*. DOI: 10.1167/tvst.11.4.1.

Nakawala, H. et al. (2019). ""Deep-Onto" network for surgical workflow and context recognition". In: *International journal of computer assisted radiology and surgery* 4.4, pp. 685–696.

Namazi, B., G. Sankaranarayanan, and V. Devarajan (2019). "LapTool-Net: A Contextual Detector of Surgical Tools in Laparoscopic Videos Based on Recurrent Convolutional Neural Networks". arXiv preprint. arXiv: 1905.08983.

Narayana, P. et al. (2019). "HUSE: Hierarchical Universal Semantic Embeddings." In: *ArXiv, abs/1911.05978.*

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

Ng, Andrew (2021). "MLOps: From Model-centric to Data-centric AI". 2021 - YouTube Video Interview.

Nguyen, Nhat-Duy et al. (2020). "An Evaluation of Deep Learning Methods for Small Object Detection". In: *Journal of Electrical and Computer Engineering*. DOI: 10.1155/2020/3189691.

Ni, Z. L. et al. (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)*. IEEE, pp. 5735–5738.

Nogueira-Rodriguez, Alba et al. (Jan. 2020). "Deep Neural Networks approaches for detecting and classifying colorectal polyps". In: *Neurocomputing* 423. DOI: 10.1016/j.neucom.2020.02.123.

Nowruzi, F.E. et al. (2019). "How much real data do we actually need: Analyzing object detection performance using synthetic and real data." In: *ArXiv, abs/1907.07061*.

Nwoye, C. I. et al. (2019). "Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos". In: *International journal of computer assisted radiology and surgery* 4.6, pp. 1059–1067.

Nwoye, C. I. et al. (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. et al. (2021a). "CholecTriplet2021: A benchmark challenge for surgical action triplet recognition". In: *arXiv:2204.04746*.

Nwoye, C. I. et al. (2021b). "Rendezvous: Attention Mechanisms for the Recognition of Surgical Action Triplets in Endoscopic Videos". In: *Elsevier Journal of Medical Image Analysis 2022*.

Orting, S. N. et al. (2020). "A Survey of Crowdsourcing in Medical Image Analysis". In: *Human Computation Journal* 7 (1), pp. 1–26. DOI: 10.15346/hc.v7i1.1.

Pakhomov, D. et al. (2019). "Deep residual learning for instrument segmentation in robotic surgery". In: *International Workshop on Machine Learning in Medical Imaging*, pp. 566–573.

Pan, S. J. and Q. Yang (2010). "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22, p. 10.

Pang, Shanchen et al. (2019). "A novel YOLOv3-arch model for identifying cholelithiasis and classifying gallstones on CT images". In: *PLOS ONE*. DOI: 10.1371/journal.pone.0217647.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. DOI: 10.3115/v1/D14-1162.

Pissas, T. et al. (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*.

Pourpanah, Farhad et al. (2020). "A Review of Generalized Zero-Shot Learning Methods". In: *ArXiv, abs/2011.08641*.

Prellberg, J. and O. Kramer (2018). "Multi-label classification of surgical tools with convolutional neural networks". In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8.

Qin, F. et al. (2019). "Surgical Instrument Segmentation for Endoscopic Vision with Data Fusion of reduction and Kinematic Pose". In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 9821–9827.

Qin, F. et al. (2020). "Towards Better Surgical Instrument Segmentation in Endoscopic Vision: Multi-Angle Feature Aggregation and Contour Supervision." In: *IEEE Robotics and Automation Letters, 5, 6639-6646*.

Qiu, L., C. Li, and H. Ren (2019). "Real-time surgical instrument tracking in robot-assisted surgery using multi-domain convolutional neural networks". In: *Healthcare Technology Letters* 6.6, pp. 159–164.

Raju, A., S. Wang, and J. Huang (2016). *M2CAI surgical tool detection challenge report*. Tech. rep. Tech. rep., University of Texas at Arlington.

Ramesh, Ajay et al. (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. et al. (2021b). "Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures". In: *International Journal of Computer Assisted Radiology and Surgery* 16, pp. 1111–1119. DOI: 10.1007/s11548-021-02388-z.

Rasmy, L. et al. (2021). "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction." In: *npj Digital Medicine*. DOI: 10.1038/s41746-021-00455-y.

Redmon, J. and A. Farhadi (2017). "YOLO9000: Better, Faster, Stronger". In: *IEEE Conference on Computer Vision and Pattern Recognition. Washington (DC)*. IEEE Computer Society. p. 6517–6525.

Redmon, Joseph et al. (2016). "You Only Look Once: Unified, Real-Time Object Detection". In: *IEEE Conference on Computer Vision and Pattern Recognition. Washington (DC)*. IEEE Computer Society; p. 779–788.

Reinke, A. et al. (2018). "How to exploit weaknesses in biomedical challenge design and organization". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain*.

Ren, S., K. He, R. Girshick, et al. (2017). "Faster R-CNN: towards real-time object detection with region proposal networks". In: *IEEE Trans Pattern Anal Mach Intell :1137–1149* 39, p. 6.

Rieke, Nicola et al. (2020). "The future of digital health with federated learning". In: *Digital Medicine* 3, p. 119. DOI: 10.1038/s41746-020-00323-1.

Rocha, Cristian, Nicolas Padoy, and Benoit Rosa (2019). "Self-supervised surgical tool segmentation using kinematic information." In: *In International Conference on Robotics and Automation (ICRA) (). IEEE* 2019, pp. 8720–8726.

Rodrigues, Mark, Michael Mayo, and Panos Patros (2021a). "Evaluation of Deep Learning Techniques on a Novel Hierarchical Surgical Tool Dataset". In: *2021 Australasian Joint Conference on Artificial Intelligence*.

Rodrigues, Mark, Michael Mayo, and Panos Patros (2021b). "Interpretable deep learning for surgical tool management". In: *4th International Workshop on Interpretability of Machine Intelligence in Medical Image Computing (iMIMIC 2021)*. Ed. by Reyes M. et al. Vol. 12929. Lecture Notes in Computer Science. Springer, Cham. DOI: 10.1007/978-3-030-87444-5_1.

Rodrigues, Mark, Michael Mayo, and Panos Patros (Mar. 2022). "OctopusNet: Machine Learning for Intelligent Management of Surgical Tools". In: *Smart Health* 23. DOI: 10.1016/j.smhl.2021.100244.

Rojas, Edgar, Kyle Couperus, and Juan Wachs (2020). "DAISI: Database for AI Surgical Instruction". In: *ArXiv abs/2004.02809*.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, Vol.9351: 234–241*.

Ross, T., A. Reinke, and P. M. et al. Full (2019). "Robust Medical Instrument Segmentation Challenge". In: *ArXiv preprint*. arXiv: 2003.10299.

Rother, Carsten, Vladimir Kolmogorov, and Andrew Blake (2004). "GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts". In: *Association for Computing Machinery SIGGRAPH Papers*. DOI: 10.1145/1186562.1015720.

Roychowdhury, S. et al. (2017). *Identification of surgical tools using deep neural networks*. Tech. rep. D-Wave Systems Inc.

Rudin, Cynthia et al. (2021). "Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges." In: *ArXiv, abs/2103.11251*.

Rueden, Laura von et al. (2021). "Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems". In: *IEEE Transactions on Knowledge and Data Engineering*. DOI: 10.1109/TKDE.2021.3079836..

Sabottke, C. F. and B. M. Spiele (2020). "The Effect of Image Resolution on Deep Learning in Radiography". In: *Radiology: Artificial Intelligence* 2.1. DOI: 10.1148/ryai.2019190015.

Sahu, M. et al. (2016). *Tool and phase recognition using contextual CNN features*. Tech. rep. Tech. Rep. [cs.CV], Zuse Institute Berlin. arXiv: 1610.08854.

Sahu, M. et al. (2017a). "Addressing multi-label imbalance problem of surgical tool detection using CNN". In: *Int J Comput Assist Radiol Surg* 12, p. 6.

Sahu, M. et al. (2017b). "Surgical Tool Presence Detection for Cataract Procedures". In: *ZIB Report* 2017, pp. 30–11.

Sahu, M. et al. (2020). "Endo-Sim2Real: Consistency learning-based domain adaptation for instrument segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 784–794.

Sahu, Manish, Anirban Mukhopadhyay, and Stefan Zachow (2021). "Simulation-to-Real domain adaptation with teacher-student learning for endoscopic instrument segmentation". In: *International Journal of Computer Assisted Radiology and Surgery*.

Sandler, Mark et al. (2018). "MobileNetV2: Inverted Residuals and Linear Bottlenecks". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sarikaya, D., J. J. Corso, and K. A. Guru (2017). "Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection". In: *IEEE Transactions on Medical Imaging* 36.7, pp. 1542–1549. DOI: 10.1109/TMI.2017.2665671.

Schoeffmann, Klaus, Mario Taschwer, and et al. (2018). "Cataract-101 – Video Dataset of 101 Cataract Surgeries". In: *MMSys'18: 9th ACM Multimedia Sys-tems Conference, June 12–15, 2018, Amsterdam, Netherlands.*

Setti, F. (2018). "To know and to learn – about the integration of knowledge representation and deep learning for fine-grained visual categorization". In: *13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*. DOI: 10.5220/0006651803870392.

Shang, Junyuan et al. (2019). "Pre-training of Graph Augmented Transformers for Medication Recommendation". In: *Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*.

Shermin, Tasfia et al. (2019). "Depth Augmented Networks with Optimal Fine-tuning." In: *ArXiv, abs/1903.10150*.

Shimizu, T. et al. (2021). "Hand Motion-Aware Surgical Tool Localization and Classification from an Egocentric Camera". In: *Journal of Imaging*. DOI: 10.3390/jimaging7020015.

Shvets, A. A. et al. (2018). "Automatic instrument segmentation in robot-assisted surgery using deep learning". In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pp. 624–628.

Silva, Santiago et al. (2019). "Federated Learning in Distributed Medical Databases: Meta-Analysis of Large-Scale Subcortical Brain Data". In: *2019 IEEE 16th International Symposium on Biomedical Imaging*.

Simon, Marcel and Erik Rodner (2015). "Neural Activation Constellations: Unsupervised Part Model Discovery with Convolutional Networks". In: *International Conference on Computer Vision (ICCV)*.

Simonyan, K. and A. Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Washington (DC): IEEE Computer Society;* 2014.

Sklar (2016). *Surgical Instruments: The Introductory Guide*. Sklar Instrument, West Chester, PA.

Socher, R. et al. (2013). "Zero-shot learning through cross-modal transfer". In: *International Conference on Learning Representations (ICLR)*.

Srivastava, N. and R. R. Salakhutdinov (2013). "Discriminative transfer learning with tree-based priors". In: *Advances in Neural Information Processing Systems, 2094–2102, 2013*.

Stockert, E. W. and A. J. Langerman (2014). "Assessing the Magnitude and Costs of Intraoperative Inefficiencies Attributable to Surgical Instrument Trays". In: *Journal of the American College of Surgeons* 219.4, pp. 646–655. DOI: 10.1016/j.jamcollsurg.2014.06.019.

Szegedy, C. et al. (2015). "Going Deeper with Convolutions". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1-9,*

Szegedy, Christian et al. (Feb. 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*, 4278–4284.

Szegedy, Christian et al. (2016b). "Rethinking the Inception Architecture for Computer Vision". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308.

Sznitman, R. et al. (2012). "Data-Driven Visual Tracking In Retinal Microsurgery". In: *MICCAI-2012*.

Takeishi, Naoya and Kosuke Akimoto (2018). "Knowledge-Based Distant Regularization in Learning Probabilistic Models." In: *ArXiv abs/1806.11332*.

Tang, Eric M et al. (2022). "Automated instrument-tracking for 4D video-rate imaging of ophthalmic surgical maneuvers". In: *Biomedical optics express*. DOI: 10.1364/BOE.450814.

Tao, F. et al. (2019). "Digital Twin in Industry: State-of-the-Art". In: *IEEE Transactions on Industrial Informatics*. DOI: 10.1109/TII.2018.2873186.

Timofeev, Aleksei et al. (2020). "Graph-RISE: Graph-Regularized Image Semantic Embedding". In: *The 12th International Conference on Web Search and Data Mining*.

Tremblay, Jonathan et al. (2018). "Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Twinanda, A. P., S. Shehata, D. Mutter, et al. (2017). "EndoNet: A deep architecture for recognition tasks on laparoscopic videos". In: *IEEE Transactions on Medical Imaging* 36, pp. 86 –97. DOI: 10.1109/TMI.2016.2593957.

Twinanda, A. P. et al. (2016). *Single-and Multi-Task Architectures for Tool Presence Detection Challenge at M2cai 2016*. Tech. rep. Tech. Rep. [cs], University of Strasbourg. arXiv: 1610.08851.

Ueki, Kazuya (2021). "Survey of Visual-Semantic Embedding Methods for Zero-Shot Image Retrieval". In: *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 628–634.

Vardazaryan, A. et al. (2018). "Weakly-supervised learning for tool localization in laparoscopic videos". In: *Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Ed. by Intravascular Imaging and Computer Assisted. Cham: Springer, pp. 169–179.

Viola, P. and M. Jones (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, Kauai, HI, USA, 2001, pp. I-I.*

Voulodimos, Athanasios et al. (2018). "Deep Learning for Computer Vision: A Brief Review". In: *Computational Intelligence and Neuroscience*, pp. 1–13. DOI: 10.1155/2018/7068349.

Wagner, Martin, Beat-Peter Muller-Stich, and et al. (2021). "Comparative Validation of Machine Learning Algorithms for Surgical Workflow and Skill Analysis with the HeiChole Benchmark". In: *arXiv preprint arXiv:2109.14956*.

Wang, D. et al. (2018). "Dividing and aggregating network for multi-view action recognition". In: *Proceedings of the European Conference on Computer Vision (ECCV), September 2018*, pp. 451–467.

Wang, L., Y. Li, and S. Lazebnik (2016). "Learning Deep Structure-Preserving Image-Text Embeddings". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5005–5013. DOI: 10.1109/CVPR.2016.541..

Wang, S. et al. (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. Springer, Cham.* 10.1007, pp. 1–36.

Wang, Yu-Xiong, D. Ramanan, and M. Hebert (2017). "Growing a Brain: Fine-Tuning by Increasing Model Capacity". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ward, Thomas M. et al. (2021a). "Challenges in surgical video annotation". In: *Computer Assisted Surgery* 26 (1), pp. 58–68. DOI: 10.1080/24699322.2021.1937320.

Ward, Thomas M. et al. (2021b). "Computer vision in surgery". In: *Surgery* 169.5. DOI: 10.1016/j.surg.2020.10.039.

Wehrmann, Jonatas, Camila Kolling, and Rodrigo Barros (2019). "Adaptive Cross-modal Embeddings for Image-Text Alignment". In: *AAAI Conference on Artificial Intelligence*.

Weinberger, Kilian Q., John Blitzer, and Lawrence Saul (2005). "Distance Metric Learning for Large Margin Nearest Neighbor Classification". In: *Advances in Neural Information Processing Systems 18 (NIPS 2005)*.

Weiss, K., T. Khoshgoftaar, and D. Wang (2016). "A survey of transfer learning". In: *Journal of Big Data* 3.10, pp. 43–6.

Wen, Y. et al. (2016). "A discriminative feature learning approach for deep face recognition". In: *In European Conference on Computer Vision (ECCV)*. DOI: 10.1007/978-3-319-46478-7_31.

Wohlin, Claes (2014). "Guidelines for snowballing in systematic literature studies and a replication in software engineering". In: *ACM International Conference Proceeding Series* 10, p. 1145.

Xian, Yongqin et al. (2018). "Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xiao, Tete et al. (2018). "Unified Perceptual Parsing for Scene Understanding". In: *Computer Vision and Pattern Recognition*. Ed. by V. Ferrari et al. Vol. 11209. Lecture Notes in Computer Science. Springer, Cham. DOI: 10.1007/978-3-030-01228-1_26.

Xue, Yao et al. (2022). "A new weakly supervised strategy for surgical tool detection". In: *Knowledge-Based Systems 239 (2022) 107860*.

Yamazaki, Y. et al. (2020). "Automated surgical instrument detection from laparoscopic gastrectomy video images using an open source convolutional neural

network platform". In: *Journal of the American College of Surgeons* 230 (5). DOI: 10.1016/j.jamcollsurg.2020.01.037.

Yan, Zhicheng, Hao Zhang, Robinson Piramuthu, et al. (2015). "HD-CNN: Hierarchical Deep Convolutional Neural Networks for Large Scale Visual Recognition". In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 2740–2748.

Yang, Congmin, Zijian Zhao, and Sanyuan Hu (2020). "Image-based laparoscopic tool detection and tracking using convolutional neural networks: a review of the literature". In: *Computer Assisted Surgery, 25:1, 15-28*.

Yang, H. et al. (2019a). "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 . , Cham*, pp. 263–271.

Yang, Qiang et al. (2019b). "Federated Machine Learning: Concept and Applications". In: *ACM Transactions on Intelligent Systems and Technology*. DOI: 10.1145/3298981.

Yip, Mighten et al. (2021). "Deep learning-based real-time detection of neurons in brain slices for in vitro physiology." In: *Scientific Reports*. DOI: 10.1038/s41598-021-85695-4.

Yu, F. et al. (2018). "Deep Layer Aggregation". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 2403-2412*.

Zadeh, S. Madad et al. (2020). "SurgAI: deep learning for computerized laparoscopic image understanding in gynaecology". In: *Surg Endosc. 34(12):5377-5383*.

Zeiler, M.D. and R. Fergus (2014). "Visualizing and Understanding Convolutional Networks". In: *European Conference on Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham*, pp. 818–833.

Zhang, Chen et al. (Mar. 2021). "A survey on federated learning". In: *Knowledge-Based Systems* 216. DOI: 10.1016/j.knosys.2021.106775.

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

Zhang, Q. et al. (2020). "Interpretable CNNs for Object Classification". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, Quanshi et al. (2019). "Interpreting CNNs via Decision Trees". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Zhongkai, Benoit Rosa, and Florent Nageotte (2021). "Surgical Tool Segmentation Using GenerativeAdversarial Networks With Unpaired Training Data". In: *IEEE Robotics and Automation Letters*.

Zhao, Hengshuang et al. (2016). "Pyramid Scene Parsing Network". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239. DOI: 10.1109/CVPR.2017.660.

Zhao, Z. et al. (2017). "Tracking-by-detection of surgical instruments in minimally invasive surgery via the convolutional neural network deep learning-based method". In: *Computer Assisted Surgery* 22, pp. 26–35. DOI: 10.1080/24699322.2017.1378777.

Zhao, Z. et al. (2019a). "Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade". In: *Healthcare Technology Letters* 6, p. 6.

Zhao, Z. et al. (2019b). "Real-time tracking of surgical instruments based on spatio-temporal context and deep learning". In: *Computer Assisted Surgery* 24, pp. 20–29.

Zhao, Z. et al. (2019c). "Surgical tool tracking based on two CNNs: from coarse to fine". In: *The Journal of Engineering* 2019.14, pp. 467–472.

Zhou, B. et al. (2015). "Object detectors emerge in deep scene CNNs". In: *International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA,*

Zhou, J. T. et al. (2014). "Hybrid heterogeneous transfer learning through deep learning". In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. Québec, Canada — July 27 - 31, 2014: Québec City, pp. 2213–2219.

Zhou, S. K., D. Rueckert, and G. Fichtinger (2019). *Handbook of medical image computing and computer assisted intervention*. Academic Press.

Zhu, X. and M. Bain (2017). "B-CNN: Branch Convolutional Neural Network for Hierarchical Classification". In: *ArXiv, abs/1709.09890.*

Zhu, X. et al. (2019). "Errors in packaging surgical instruments based on a surgical instrument tracking system: an observational study". In: *BMC Health Services Research, 19:176* 2019.

Zhuang, F. et al. (2019). "A Comprehensive Survey on Transfer Learning". In: *ArXiv, abs/* 1911, p. 02685.

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

Zisimopoulos, O. et al. (2017). "Can surgical simulation be used to train detection and classification of neural networks?" In: *Healthcare Technology Letters* 4.5, pp. 216–222.

Zoph, Barret et al. (June 2018). "Learning Transferable Architectures for Scalable Image Recognition". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/CVPR.2018.00907.