Vision-based Context-Aware Assistance for Minimally Invasive Surgery

by

Jinglu Zhang

National Centre for Computer Animation

Faculty of Media & Communication

Bournemouth University

A thesis submitted in partial fulfilment of the requirements of Bournemouth University for the degree of *Doctor of Philosophy*

Feb. 2021

Copyright Statement

Acknowledgements

Throughout my four years PhD journey in Bournemouth, I have received a great deal of support and assistance.
First of all, I would like to thank my supervisors, Professor Jian Chang, Dr. Xiaosong Yang and Professor Jian Jun Zhang for always helping me and supporting me. I would particularly like to single out my supervisor Professor Jian Chang. From the study to life, your insightful feedback and valuable suggestions has reshaped my attitude towards research and brought my work to a higher level.
I also would like to thank Yinyu Nie, Dr.Long Chen, Dr.Yunfei Fu, Dr. Qian Yu, Dr. Li Wang, Dr. Yanran Li who have been working closely with me for this dissertation. Thank all of you for being patient with me and for the invaluable discussions and novel ideas. Especially, thank Yinyu Nie for making me believe in myself and being supportive all the time, I could not achieve any of this without you.
I would like to thank my colleagues and friends: Zhangcan Ding, Ruibin Wang, my lovely roommate Yao Lyu, Nan Xiang, Mengqing Huang, Dr.Tao Jiang, Dr.Shuang Liu, Micheal Wu, Dr. Alessandro Bruno, Dr. Ehtzaz Chaudhry for creating a such lovely living and study environment. And I also want to thank the lovely staffs from FMC: Sonia Ashby, Sunny Choi , Tanesha Duff, Cansu Kurt Green, and Jan Lewis. You guys made my research life enjoyable and easier. Special mention to my piano teacher and special friend Vianna Renaud, who has brought the piano and classical music to my life.

I would like to thank my friends in UK: my dearest Yukun Wang, Xinlong Wang, Leon Su, Dr.Wenshu Zhang, Ziyuan Liu, Dr.Shaojun Bian, Dr.Feilin Han, Shimeng Sun for supporting me and staying with me. You guys have made this journey meaningful.
I also would like to thank my old friends, my best friend Ang Li, Shuhui Liu, Bo Chen, Ziye Yang, Xin Li, Yandi Wu, Haoran Li, Ruoyu Wang, Xiao Tan, Yuanzhi Yao, Sai Ma, Lu Pan, Xin Jin, Xinru Chen, Zuojun Li, who have cheered me up and comforted me whenever I need. Thank you for being my life long friends.
Last but not least, I would like to thank my parents. I want to express all my love and grateful to you. Thank you for always giving me chance to choose the life I want.

Abstract

Context-aware surgical system is a system that can collect surgical data and analyze the operating environment to guide responses for surgeons at any given time, which improves the efficiency, augment the performance and lowers the risk of minimally invasive surgery (MIS). It allows various applications through the whole patient care pathway, such as medical resources scheduling and report generation. Automatic surgical activities understanding is an essential component for building context-aware surgical system. However, analyzing surgical activities is a challenging task, because the operating environment is considerably complicated. Previous methods either require the additional devices or have limited ability to capture discriminating features from surgical data.

This thesis aims to solve the challenges of surgical activities analysis and provide context-aware assistance for MIS. In our study, we consider the surgical visual data as the only input. Because videos and images own high-dimensional and representative features, and it is much easier to access than other data format, for example, kinematic information or motion trajectory.

Following the granularity of surgical activity in a top-down manner, we first propose an attention-based multi-task framework to assess the expertise level and evaluate six standards for surgeons with different skill level in three fundamental surgical robotic tasks, namely suturing, knot tying and needle passing. Second, we present a symmetric dilated convolution structure embedded with self-attention kernel to jointly detect and segment

fine-grained surgical gestures for surgical videos. In addition, we use the transformer encoder-decoder architecture with reinforcement learning to generate surgical instructions based on images.

Overall, this thesis develops a series of novel deep learning frameworks to extract high-level semantic information from surgical video and image content to assist MIS, pushing the boundaries towards integrated context-aware system through the patient care pathway.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Minimally invasive surgery (MIS) is a surgical procedure where operations can be performed through small incisions (usually 0.5 - 1.5 cm) with the assistance of a 2D video camera and several instruments. Comparing with the traditional open surgery, MIS has few post-operative complications, small incisions on the skin, and a relative shorter recovery period. However, the revolution of the technology has also altered the operation routines. For example, the limited field of view (FOV) and the restricted operating space during the MIS often cause undesirable complications. Therefore, it is indispensable to provide efficient assistances to improve operation quality, performance and safety for whole surgical-care pathway.

There has been a growing interest of building context-aware system (CAS) utilizing available information inside the operation room (OR) to provide clinicians with contextual support at appropriate time (Bardram and Nørskov 2008, Nakawala et al. 2017, van Amsterdam et al. 2020). Applications include remaining time estimation, resource scheduling, decision support, etc. In order to achieve this goal, understanding OR content by analyzing surgical activities in different granularity is the prerequisite. The definition of surgical activity depends on the level of abstractions is kind of vague due to the difference of the subjective cognition. We adopt the notion of activity granularity presented in (Lalys and Jannin 2014) as shown in Fig. 1.1.

Surgical procedure (types of surgery such as cholecystectomy and adrenalectomy) is the coarsest level. Each surgery contains a list of phases, which are the major events occurring during surgery. Taking cholecystectomy as an example, it contains seven phases: *preparation, calot triangle dissection, clipping and cutting, gallbladder dissection, gallbladder packaging, cleaning and coagulation and gallbladder retraction.* A phase can be divided into a sequence of steps to achieve a certain task, for instance, suturing and needle passing. Taking one step further, a surgical step consists of several basic actions, for example, pushing needle through tissue. Then we have motion trajectory with no semantic information, and the finest level is the presence of the person or object.

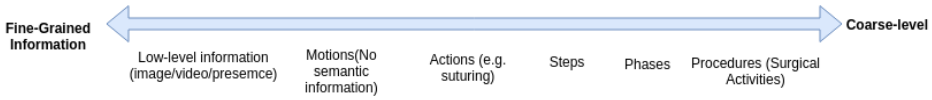


Figure 1.1: Axis shows the granularity of surgical activity from the finest level (left) to the coarsest level (right)

Early methods analyze surgical activities using motion data (Judkins et al. 2009, Bodner et al. 2004) (e.g. instrument usage, trajectory, and total distance traveled), however, additional sensors(Bardram et al. 2011) and human defined motion metrics are always required. Other approaches apply handcrafted visual cues, including pixel values, color, shape, etc., and model the data with statistic models (Hidden Markov and its variants) (Tao et al. 2013, Reiley and Hager 2009, Tao et al. 2012). Results from these methods are plausible and interpretable. Nonetheless, handcrafted features are empirically, and HMM models need manually decompose the surgical action and tune the hard parameters.

In recent years, deep learning algorithms have achieved the great performance in various computer vision tasks, including action recognition (Wang et al. 2016), image classification (Krizhevsky et al. 2012), image segmentation (Ronneberger et al. 2015), etc. Higher level representations can be progressively extracted from the raw input. There has been a growing interest in utilizing deep learning techniques for surgical activities understanding.

For example, in (Wang and Fey 2018), the authors design an end-to-end convolutional neural network for surgical skill assessment using 76 dimensional kinematics data. In another study, Twinanda et al. (2016a) propose EndoNet to jointly detect surgical phase and tool presence on cholecystectomy videos.

Whereas, deep learning methods in surgical activity analysis are still under-explored because of the limited accessibility of medical data and the high-dimensional and irregular surgical content. Accordingly, to solve the surgical data shortage problem, we use the pre-trained models from the open domain to extract the low-level visual feature and apply the transfer learning on the top. Our motivation is to design novel deep learning networks to understand surgical content and provide context-aware assistance for minimally invasive surgery. Figure 1.2 shows a conceptual of vision-based context-aware assistance.

Particularly, we focus on vision-based (surgical videos and images) approaches on account of following reasons. First, owing to the enormous innovations on computer aided surgery (CAS) (Raab 1998) and robotic-assisted minimally invasive surgery (RMIS) such as *da Vinci* (Bodner et al. 2004) system, it is straightforward to record and access to surgical videos and images. Whereas additional devices are required for other types of digital signals. For example, radio-frequency identification (RFID) tracking systems electromagnetic (EM) sensors are often employed to track instrument usage and the movement of clinicians (Parlak and Marsic 2013, Parlak et al. 2011). Second, visual information owns high-dimensional and abundant features than other digital signals. For instance, the kinematics data can only roughly describe how the current action is performed (with position, angle, velocity parameters) without knowing what is actually performed, which is the additional information from the visual data. Several applications can only achieved by analyzing visual data, such as video summarization, video segmentation and indexing, and concept retrieval.

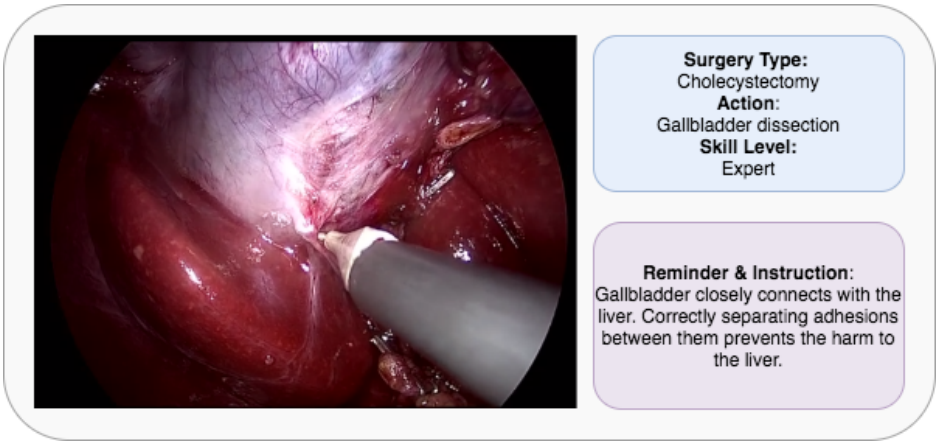


Figure 1.2: The context awareness system automatically recognize the surgery type, the current surgical action, the expertise level of surgeon and generate the instruction

1.2 Main Challenges

Although related technologies (video classification, action recognition, image captioning) have been well studied in general computer vision community, vision-based deep learning methods for surgical activities analysis is still at its early stage. This is mainly because the MIS scene is different with the visual data used in general computer vision research. Specifically, daily scenes often include differentiable foreground (such as human, animal, plant, etc.) and different background. While most of the surgical activities share similar environment due to the similar appearance, color, and texture of human anatomic structure. On the other hand, surgical data often has high level of heterogeneity. Because surgical process is specific to the medical condition, the surgeon and the patient (Lalys and Jannin 2014), the process varies significantly from one to another. In addition, the operation environment for MIS is particularly complicated, the motion of the camera and the frequent usage and movement of the instruments cause the problem of occlusion, blur, and the presence of the smog (see Fig 1.3). Therefore managing, analyzing, and understanding the content of surgical activities is still highly challenge for above reasons. Overcoming these challenges requires the techniques to extract high level discriminative information to represent surgical activities

from the complicated and heterogeneous data.

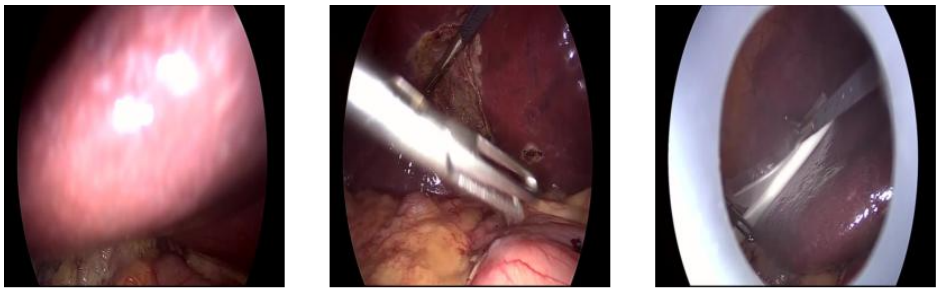


Figure 1.3: Examples of blurry image, the occlusion and the movement of the camera

1.3 Aims and Objectives

This thesis aims to analyze surgical content and design a vision-based context-aware system to provide efficient assistance for MIS. Following the surgical granularity in a top-down manner, from the coarse to fine level, we focus on vision-based surgical skill assessment for a surgical procedure, fine-grained surgical action detection, and single shot surgical instruction generation for context aware surgical activities analysis. Based on our aims, we identify our research questions and objectives as follows:

- **How to identify skill level from a long and untrimmed surgical video?** Videos in general computer vision field usually last for seconds or few minutes, however surgical videos always continue for minutes or even hours. In addition, some surgical actions take few seconds while some take minutes, and they do not equally contribute to the skill determination. The major difficulty is to efficiently model dependencies between the skill level and related frames.

- **How to jointly segment and classify fine-grained surgical gestures?** Fine-grained surgical gestures have low inter-class variance and high intra-class variance based on the fact that a surgery procedure is

performed in an environment with little changes and limited field of view. In addition, the same instrument can be applied in different gestures and one gestures might need multi-instruments. The challenge is to extract spatial features and build the dependency between frames.

- **How to generate surgical instructions from a single image?** Scenes captured in general environments often have discriminate background and characteristics. However, surgical images always has high similarity in appearance due to the constant background and similar color and texture of the anatomic structure and instruments. It is particularly difficult to bridge the visual patterns with informative human linguistic descriptions in a dataset with the limited size.

In order to answer these questions, this thesis aims to achieve three major objectives:

- Our first objective is to objectively assess surgical skill levels taking surgical videos as the only input. In particular, we focus on designing a framework to automatically extract representative spatial-temporal features for skill determination. This objective is explored in Chapter 3.

- The second objective is to jointly segment and classify surgical gestures from the long and untrimmed surgical videos. Surgical gesture recognition requires identify the start-end frames for each fine-grained gestures. We concentrate on capturing both global and local spatial-temporal dependencies. This objective is achieved in Chapter 4.

- Finally, we aim to generate natural human language captions from a single image as the instruction for surgeons. We try to bridge the gap between visual and textual information. Ideally, the algorithm should be able to look at the image, understand current situation, and generate the relevant guidance. This objective is addressed in Chapter 5.

1.4 Contributions

The main contribution of this thesis is to propose novel vision-based deep learning solutions for surgical activities understanding, such that the methods can be applied for context-aware MIS assistance. Our main concern is to bridge semantic gap between low level visual information and high-level medical concept. The detailed contributions in each chapter are summarized as follows:

- In Chapter 3 we propose a novel network architecture to automatically assess surgical skills only using RGB surgical video sequence. Our structure applies a 3D residual network (3D ResNet) to extract spatial-temporal features and involves frame-to-frame relational features through a self-attention module. We evaluate our method on three fundamental robotic surgical tasks (suturing, needle passing, and knot-tying). In addition to the original expertise level prediction task, we also use our framework to evaluate the six assessment standards concurrently. We achieve nearly 100% accuracy for three tasks. In regard to the six standards evaluation, except for the *time and motion*, which we infer is unrelated to the skill determination, predictions of other five targets achieve satisfying accuracy, ranging from 56% to 91%. The results indicates that our architecture is able to obtain representative features by extensively considering the spatial, temporal and relational context from raw video input. Accordingly, this technique introduces applications such as the automatic generation of the comprehensive skill assessment report.

- In Chapter 4 we propose a novel temporal convolutional architecture to automatically detect and segment surgical gestures with corresponding boundaries only using RGB videos. We devise our method with a symmetric dilation structure bridged by a self-attention module to encode and decode the long-term temporal patterns and establish the frame-to-frame relationship accordingly. We validate the effectiveness

of our approach on a fundamental robotic suturing task from the JIGSAWS dataset. The experiment results demonstrate the ability of our method on capturing long-term frame dependencies, which largely outperform the state-of-the-art methods on the frame-wise accuracy up to ∼6 points and the F1@50 score ∼6 points.

- In Chapter 5, inspired by the neural machine translation and imaging captioning tasks in open domain, we introduce a transformer-backboned encoder-decoder network with self-critical reinforcement learning to predict instructions from surgical images. We evaluate the effectiveness of our method on the DAISI dataset, which includes 290 procedures from various medical disciplines. Our approach outperforms the existing baseline over all caption evaluation metrics. The results demonstrate the benefits of the encoder-decoder structure backboned by transformer in handling multimodal context.

- In Chapter 6, we presents an application by designing and implementing a Unity-based laparoscopic cholecystectomy simulator as a starting point in order to share the idea of combining the pre-operative surgical training, intra-operative surgical guidance, and post-operative surgical skill assessment into a whole system. Our design leverages physical simulation and haptic force feedback to offer trainees a realistic visual and tactile experience, respectively. We explore the possibility of using game engine rather than developing from scratch to build the surgical simulator. Based on the results and user feedbacks from a pilot experiment, we conclude that game engine is a viable option for creating a cost-effective, flexible and highly interactive virtual surgery training platform for pedagogical purpose, which can shorten the development time.

1.5 List of Publications

- Zhang, Jinglu, Yinyu Nie, Yao Lyu, Hailin Li, Jian Chang, Xiaosong Yang, and Jian Jun Zhang. "Symmetric Dilated Convolution for Sur-

gical Gesture Recognition." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 409-418. Springer, Cham, 2020.

- Zhang, Jinglu, Yao Lyu, Yukun Wang, Yinyu Nie, Xiaosong Yang, Jianjun Zhang, and Jian Chang. "Development of laparoscopic cholecystectomy simulator based on unity game engine." In Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production, pp. 1-9. 2018.

- Zhang, Jinglu, Jian Chang, Xiaosong Yang, and Jian J. Zhang. "Virtual reality surgery simulation: A survey on patient specific solution." In International Workshop on Next Generation Computer Animation Techniques, pp. 220-233. Springer, Cham, 2017.

- Lyu, Yao, Jinglu Zhang, Jian Chang, Shihui Guo, and Jian Jun Zhang. "Integrating Peridynamics with Material Point Method for Elastoplastic Material Modeling." In Computer Graphics International Conference, pp. 228-239. Springer, Cham, 2019.

- Bruno, Alessandro, Morgan Moore, Jinglu Zhang, Stphane Lancette, Ville P. Ward, and Jian Chang. "Toward a head movementbased system for multilayer digital content exploration." Computer Animation and Virtual Worlds (2020): e1980.

1.6 Outline of Thesis

This thesis is organized as follows:

- Chapter 2 reviews the state-of-the-art approaches for surgical content analysis, including surgical skill assessment, surgical gesture recognition, surgical instruction captioning, and their corresponding vision-based techniques.
- Chapter 3 contains the proposed method for automatic multi-task surgical skills assessment
- Chapter 4 includes fine-grained surgical gesture recognition task with encoder-decoder symmetric dilated architecture
- Chapter 5 describes our method of generating surgical instruction, which use the transformer encoder-decoder with reinforcement learning
- Chapter 6 implement a Unity-based laparoscopic cholecystectomy simulator towards integrating the context awareness surgical system
- Chapter 7 concludes the paper and discuss the possible solutions to improve the current study

Chapter 2

Literature Review

In this chapter, we discuss previous works related to surgical activity analysis. We first review some application scenarios for context-awareness system and the next generation surgery. Then we discuss state-of-the-art approaches in vision-based surgical skill assessment for a surgical procedure, fine-grained surgical action detection, and image based surgical instruction generation. Finally, we review corresponding deep learning based techniques: action recognition, action segmentation, and image captioning.

2.1 Context-aware Assistance for Surgical Applications

2.1.1 Surgical Applications

Surgical data science will build next-generation surgery (Maier-Hein et al. 2017). This emerging scientific field focuses on collecting, managing, analyzing and modelling surgical data to assist the whole patient-care pathway (see Fig. 2.1). Among great varieties technologies of surgical data science, surgical activity understanding is the prerequisite of building context-awareness system. Specifically, such system has many ***intra-operative*** and ***post-operative*** applications.

1. ***Intra-operative applications*** The intra-operative applications of context-aware system involve surgical support for surgeons and resource

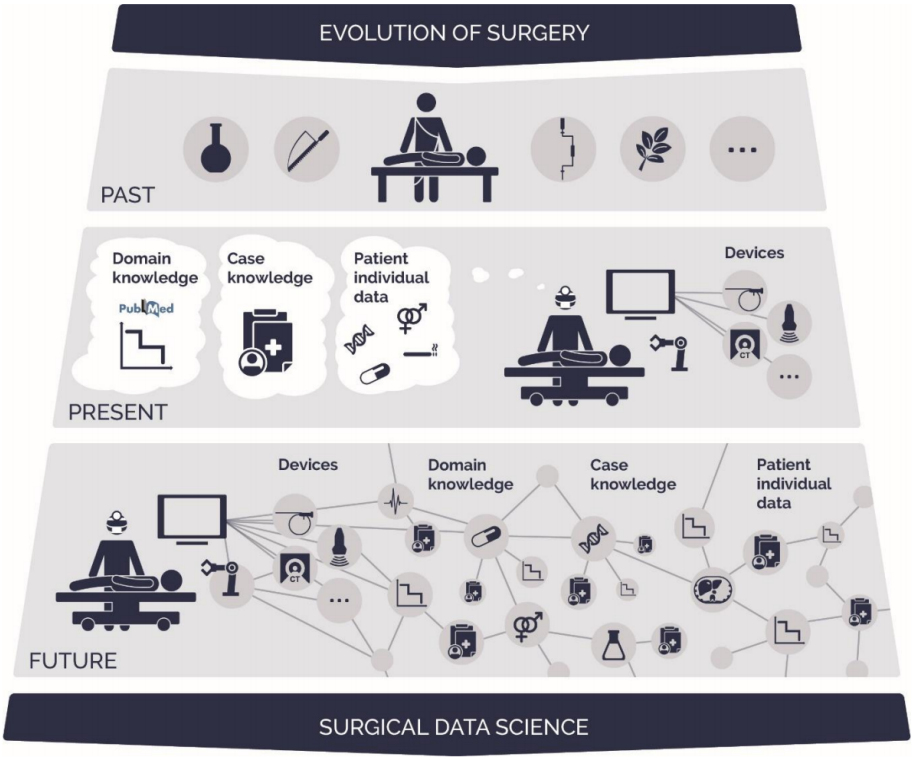


Figure 2.1: Evolution of surgery (Maier-Hein et al. 2017). In the past, surgeons perform the treatment only based on their own experience with minimal devices and tools. At present, abundant information and equipment are accessible. But still, a treatment only rely on domain knowledge. In the future, surgical data science will integrate everything together.

scheduling (Franke et al. 2013). For example, real-time surgical gesture recognition would be able to decide the suitable time to provide useful information (e.g. overlay simulated structure for MIS as shown in Fig. 2.2). Moreover, it can be applied to estimate the remaining time (van Amsterdam et al. 2020) for OR resource management such that minimize the waiting time for patients. In addition, by deeply analyzing surgical workflow, the system can detect rare case or wrong action and provide instructions to handle these cases.

2. ***Post-operative applications*** As for post-operatively, gesture recognition can help with the surgical context indexing and management.

For example, if a novice surgeon want to learn the *gallbladder dissection* procedure from laparoscopic cholecystectomy, he or she needs to manually find the video in database and scroll the video to the wanted part. With the context-aware assistance, the surgeon can just search the name of that procedure. Furthermore, with the help of surgical skill assessment, the system can evaluate the performance, generate the surgical report, and give improvement suggestions for training and education purpose.

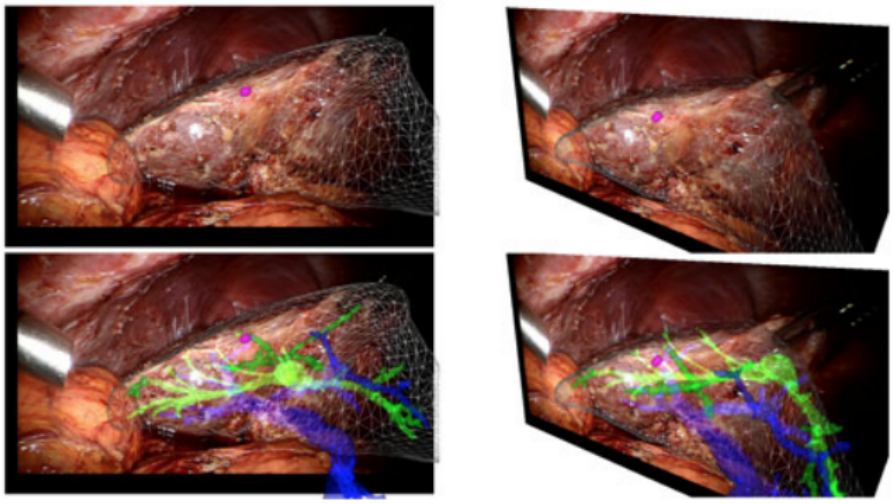


Figure 2.2: Registration of a physically-based liver model during a minimally invasive liver surgery (Haouchine et al. 2014)

2.1.2 Surgical Skill Assessment

Traditionally, the "see one, do one, teach one" approach of Halsted (Kerr and PATRICK 1999) is the mainstream to train a novice surgeon. Besides the long training curve, the huge expenditure in capital and resources, etc., the performance assessment from the experienced surgeon is rather subjective and insufficient. Objective structured assessment of technical skill (OSATS) (Martin et al. 1997), a global rating score system, has been developed to evaluate specific surgical skills for trainees in an objective and systematic

manner. But still, the manual OSATS assessment costs time and human labor resources, because experts are required to observe trainees performing the operation. Under this circumstance, the automatic assessment of dexterous surgical skills while eliminating the subjective factor and implementing the self-guided training becomes crucial. The major challenge of this idea is how to handle and analyze the data such that surgeons can improve their dexterity upon these information.

Before the deep learning era, objective skill assessment techniques mainly fall into two strategies: motion feature based, and latent state representation and interpretation. Number of researches have shown the inseparable correlation between the skill level and surgical motions (Bodner et al. 2004, Datta et al. 2001, Judkins et al. 2009), for instance, the instrument trajectory and orientation, number of actions, and the operation time. Accordingly, the first strategy concentrates on observing different motion features and designing the descriptive metrics to identify surgical skill. In (Judkins et al. 2009), the authors measure the time to task completion, total distance of path, speed, curvature, and relative phase for three laparoscopical training tasks to clearly differentiate the performance between the expert and the novice surgeon. Nevertheless, motion metrics were selected humanly, so defining the optimal metrics is always a controversial problem.

Differ from the first straightforward methodology, the other one tends to transfer the raw motion data into an intermediate representation, and using this latent interpretation to evaluate the skill. It can be further divided into machine learning based and statistic model based approaches. Machine learning based approaches usually work with the Bag-of-Words (BoW) (Bettadapura et al. 2013) expression using motion features (see Fig. 2.3) (Fard et al. 2016), or more robust feature descriptors such as Histogram of Gradients (HOG) (Zhang et al. 2013). Subsequently, different machine learning algorithms (e.g. logistic regression (LG), support vector machine (SVM), k-nearest neighbors (kNN)) are explored for expertise level classification. It is worth noticing that the handcrafted feature selection is reckless, because some valuable features are possible to be neglected. While the statistic model

based approaches focus on decomposing a surgical task into a series of predefined atomic gestures. The statistic modeling algorithms, HMM and its variants (Reiley and Hager 2009, Tao et al. 2012, Zhang and Li 2011, Wang and Mori 2009), are then applied to maximize the likelihood for a given sequence. However, the process of manually decomposing surgical gestures is tedious, not to mention the hard parameters tuning for Markov model.

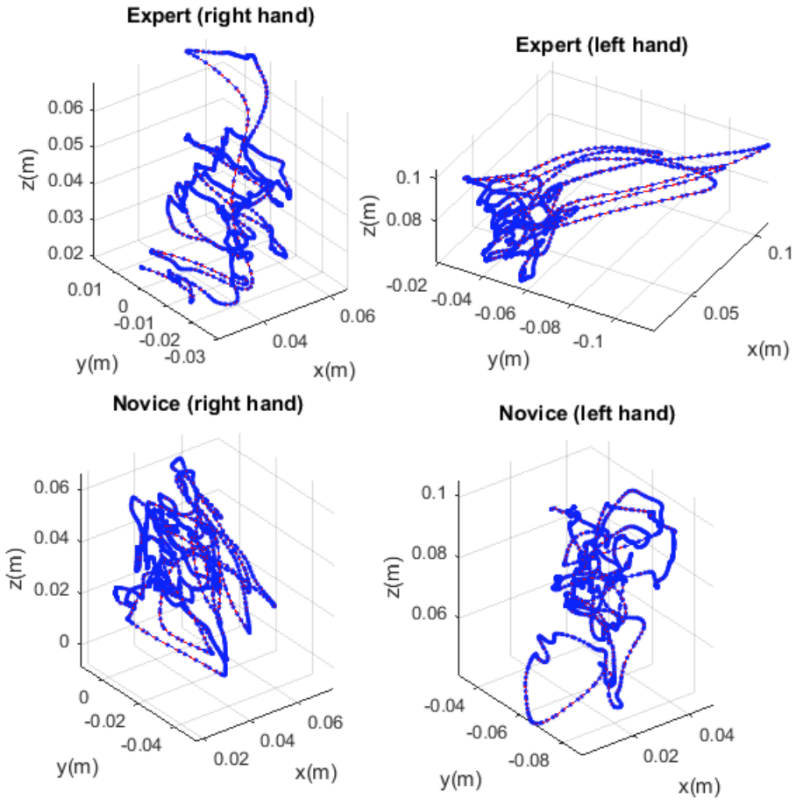


Figure 2.3: Different motion trajectories for expert and novice surgeons (Fard et al. 2016)

A comprehensive review of deep learning in medical image analysis was presented in (Litjens et al. 2017). Here, we only review some related deep learning applications in surgical video analysis and skill determination.

As for the automatic surgical skills assessment, deep learning related algorithms are still under-explored. Notwithstanding the fact that existing deep learning evaluation methods have achieved high accuracy ranging from

93% to 100% (Fawaz et al. 2018, Funke et al. 2019b, Wang and Fey 2018), they either require the kinematics data or fail to capture the long-range temporal dependency. There are other similar skill determination studies not limit to medical application. In Parmar and Tran Morris (2017), Parmar and Morris utilize 3D convolutional neural networks (C3D) and followed by the long short-term memory (LSTM) networks to score Olympic events. In another work, Doughty et al. (2019) take advantages of the Siamese Network (Bertinetto et al. 2016) and customized triple loss function to rank the skills for different tasks (scrambling eggs, braiding hair, suturing, etc.). Among their results, the surgical task gets the lowest final score, it indicates that the architecture designed for the general skills evaluation is insufficient to capture fine-grained surgical motions for different skill levels.

2.1.3 Surgical Action Recognition

As discussed in Chapter 1, surgical activities follow a descending order of granularity. Most of the researches focus on decomposing and classifying coarse level activities (surgical phases), or recognizing fine-grained activities (gestures and tasks). Some of the researches also try to detect and analyze lower level activities, such as key shots and surgical episodes. In this section, we review related techniques for surgical phase recognition and surgical gesture recognition as these two areas have high correlation to our study.

Surgical phase recognition is defined as surgical workflow analysis (SWA). It has many important clinical applications, for instance, operational time scheduling, key phrase indexing, etc. In order to find the semantic concept from low level visual information, the pipeline of surgical activities analysis generally includes: video segmentation (dividing the video into structural units), feature extraction (representing the object or action), feature matching, and action classification (Hu et al. 2011).

Some of the early works depend on tool usage signals to identify current phase (Padoy et al. 2012, Stauder et al. 2014, Bouarfa et al. 2011), however, this requires recording and annotating tool data at each time step, which

is time-consuming and costly. Correspondingly, content-based phase detection has shown growing interest. In (Blum et al. 2010), the authors achieve phase recognition tasks by combining histograms, colour values, and gradient magnitudes etc. features. Then, three different algorithms are applied, namely dynamic time warping (DTW), combinations of hidden Markov models (HMMs), and principle component analysis (PCA).

In addition, more advanced feature extraction and classification approaches based on deep learning have been recently proposed. Twinanda et al. propose Endonet (Twinanda et al. 2016a), which fine-tunes AlexNet (Krizhevsky et al. 2012) architecture for jointly detecting surgical phase and tool presence on laparoscopic cholecystectomy videos (see Figure 2.4). The highest precision reaches 92.2% and 86% for online and offline recognition, respectively. This work has been regarded as the baseline of surgical phase recognition based on deep learning. Later on, Jin et al. (2018) combine a fine-tuned Resnet50 architecture for visual feature extraction and LSTM network to encode temporal information. Lea et al. (2016) also use Spatiotemporal CNN to capture object motion over short time intervals, for offline surgical phase recognition.

Beyond the recognition of coarse phases, some studies apply Hidden Markov Model (HMM) (Tao et al. 2013) and its variants (Lea et al. 2015)to identify the latent state of surgical actions. The latent states transferring among successive actions are subsequently modelled by the transition probability. Although state features in HMMs are interpretable, they only focus on few local frames hence making the model incapable of capturing the global pattern. In addition, some machine learning methods (i.e. Support Vector Machine (SVM) (Twinanda et al. 2016a)) assemble multiple heterogeneous features (color, motion, intensity gradients, etc.) to localize and classify surgical actions. Nonetheless, these features are hand-crafted. Therefore some crucial latent features could be neglected during feature extraction procedure.

More recently, Liu and Jiang (2018) employ deep reinforcement learning algorithm to model the task as a sequential decision-making process and reduce the over-segmentation error. Every time step, the agent look through the video sequence from the beginning and gradually learns a strategical

policy to classify the frame based on the reward. Furthermore, in order to extract discriminative features from kinematics data, van Amsterdam et al. (2020) present a framework to simultaneously recognize the surgical gesture and predict the surgical task progress. The results prove that multi-task architecture improve the performance of surgical gesture recognition without any additional human annotation.

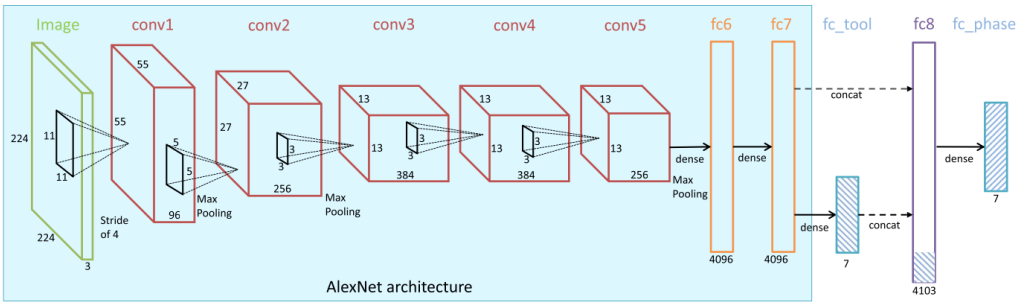


Figure 2.4: Endonet architecture for tool detection and surgical phase recognition (Twinanda et al. 2016a)

2.1.4 Surgical Instruction Generation

Surgical instruction generation deals with the problem of automatically creating guidance from images of surgical procedure. The only prior work for generating surgical instructions from medical images is (Rojas-Muñoz et al. 2020), which is also the baseline of our work in Chapter 5. In their work, the authors present a Database for AI Surgical Instruction (DAISI) and use bi-directional RNN to generate image captions. Correspondingly, we review the most closely related topic, medical report generation. Manual report writing can be subject to error for novice and tedious for experienced physicians. Given a medical image, such as radiology or pathology, automatic report generation aims at describing the ***impression***, ***findings***, ***tags***, etc. for the patient (see Fig. 2.5 as an example).

Medical report generation is a relatively new research field, because the combination of image and human language is particular challenging. On

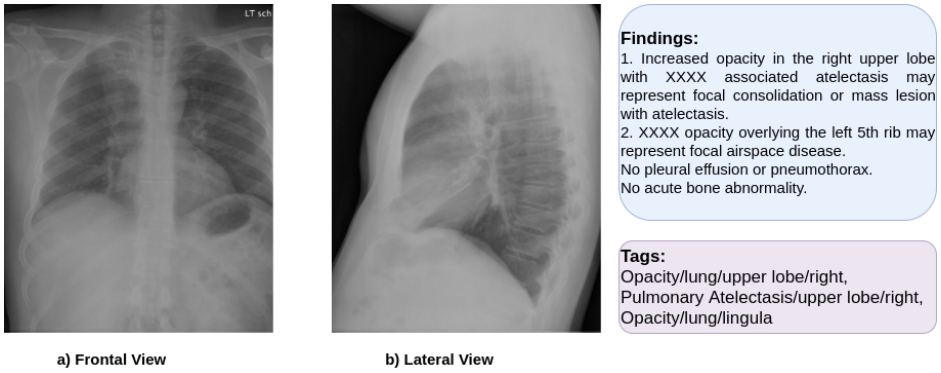


Figure 2.5: The chest x-ray report example has the *findings* section for examinations from different body areas and the *tags* section indicates key clinical information (Demner-Fushman et al. 2016)

the other hand, due to the privacy and sensitiveness of patient data, such resources are not easily accessible, and current public available dataset is rather small or noisy (Pavlopoulos et al. 2019). Some of the early works intend to match semi-template or template texts descriptions to medical images. Taking both optical coherence tomography (OCT) images and textual input, Schlegl et al. (2015) trained a CNN to extract semantic concepts from the textual report and classify images. In (Shin et al. 2016), the authors adopt a CNN-RNN architecture for key words (e.g. locations, disease) prediction from chest x-ray images.

One of the earliest medical report generation works based on natural language is (Jing et al. 2017), where a pipeline of jointly predicting tags and generating long paragraphs with co-attention and hierarchical LSTM is proposed. However, applying conventional image captioning approaches is not sufficient for medical report generation, because radiology report are usually consists of a long paragraph with multiple sentences. In another study, Li et al. (2018) combine the traditional retrieval-based and contemporary learning-based approaches with manually extracted template. They train an agent using deep reinforcement learning, which is updated in regards to sentence-level and word-level rewards. More recently, Chen et al. (2020) improve the transformer model (Vaswani et al. 2017) by designing a

relational memory to record key information of the generation process and a memory-driven layer normalization for transformer decoder.

Despite the challenges, medical reports also have their own discriminating characters. They often share predefined topics and follow similar writing templates, while surgical instruction prediction with natural language has no template to follow.

2.2 Vision-based Techniques

Because our study focuses on using recordings of surgical procedures to assist the surgical skill assessment, surgical gesture recognition and surgical procedure instruction for context-aware MIS. In this section, we review the corresponding vision-based techniques behind the proposed surgical solutions.

2.2.1 Action Recognition

Action recognition has been an active research topic in computer vision community for decades. Generally, it can be comprehended as the extension from single image classification to multi-frames action classification, which aggregates the information for every single frame and identify different actions from videos. Due to huge computational cost and the difficulties of capturing long text, action recognition faces many challenges.

Before deep learning based approaches, conventional computer vision approaches can be broadly divided into three steps: 1) Extracting high dimensional visual features, either densely (Wang et al. 2011) or sparsely with set of interest points (Laptev 2005, Dollár et al. 2005). Among all the handcrafted feature extraction, iDT (improved dense trajectories) (Wang and Schmid 2013), which estimate camera motion by matching feature points between frames using SURF descriptors (Bay et al. 2006) and dense optical flow, achieves state-of-the-art performance; 2) Combining and encoding the extracted features into video-level description (for example, bag of visual words); 3) Training a classifier such as SVF or RF for final prediction.

Soon after, two breakthrough backbones are proposed for action recognition, namely ***single stream*** (Karpathy et al. 2014) and ***two stream*** (Simonyan and Zisserman 2014a) network. ***Single stream*** architecture fuse spatial-temporal features from consecutive video frames using one and only network. Many methods are derived from single stream, including LSTM based (Donahue et al. 2015) and 3D convolution based (Tran et al. 2015, Yao et al. 2015) (see Figure 2.6 for the difference between 2D and 3D convolution). While ***two stream*** architecture are built upon two separate networks with one for spatial features modeling and the other for temporal features modeling. For example, if a task needs to differentiate between wash face and wash hair, the spatial branch are able to capture the spatial character (if it is face or hair) and the temporal branch can capture the duration of the action. Temporal segment networks (Wang et al. 2016) first divide a video into K segments with same lengths. Then they randomly sample snippets from each of the K segments and combine scores of spatial and temporal streams separately by averaging across snippets. Finally, weighted average and softmax are applied over all classes to fuse the final score. Another two-stream based study worth to mention is I3D (Two-Stream Inflated 3D ConvNet) (Carreira and Zisserman 2017), which demonstrate the beneficial of using 2D pre-trained convolutional networks. This research takes advantages of the 2D pre-trained models on Kinetics and apply two different 3D networks for both spatial and temporal streams. Plenty of the action recognition and action detection tasks use the extracted I3D features as the backbone (Hara et al. 2018, He et al. 2019a, Feichtenhofer et al. 2019).

2.2.2 Action Segmentation

Unlike action recognition task, which only has one action class in each video, action segmentation requires jointly detect and segment multi-class actions with their corresponding boundaries from untrimmed videos. Despite the huge progress in recent years, number of intuitive challenges still exists 1) Temporal boundaries are vague, so it is hard to define the exact start/end frame for an action; 2) Still image features can determine the video class

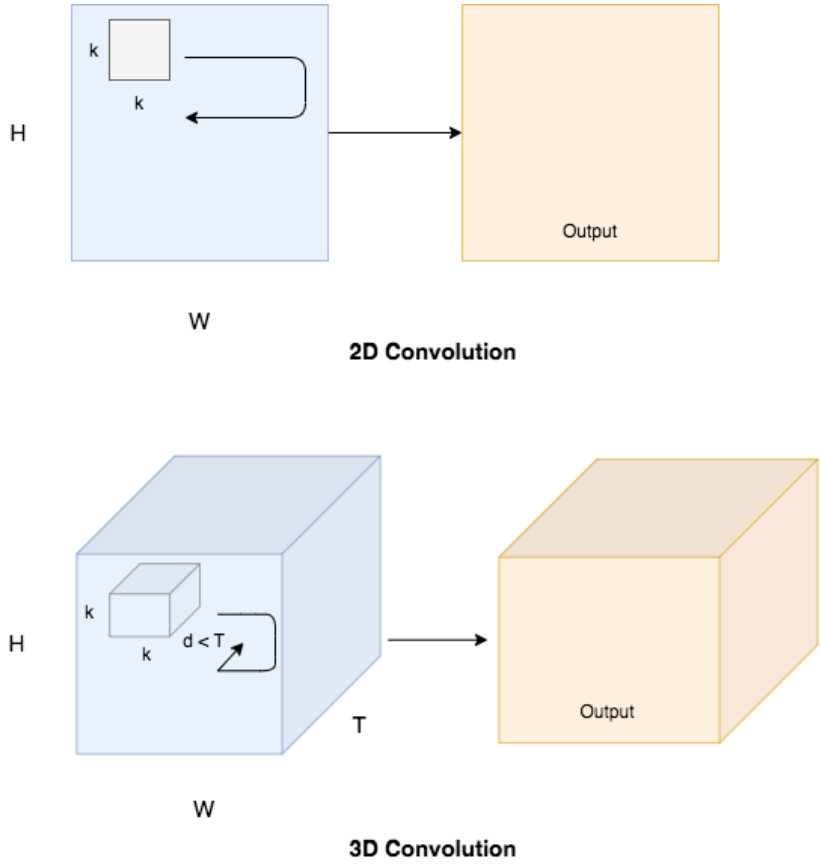


Figure 2.6: Comparison between 2D and 3D convolution. Where k is the kernel size, T stands for the time, and d is the kernel size on time dimension

in action recognition task, although it is better to combine with the temporal features. However, in action detection task, temporal information is indispensable; 3) Time durations between different actions are huge. Some actions only sustain for few seconds, whereas some last for minutes.

Early studies for action segmentation can be divided into two categories: sliding window approaches (Rohrbach et al. 2012, Karaman et al. 2014) with multi-scale temporal windows to segment actions and hybrid approaches (Kuehne et al. 2016, Richard et al. 2017), which use Markov models for coarse temporal modeling above the frame-wise classifiers. Although conventional models works well on modeling temporal dependencies, it takes too much time and computational power to solve the maximization problem over long sequence.

While recent studies solve action segmentation problem in two stages: 1) Extracting the spatial or spatial temporal features by pre-trained CNN models (e.g. Res3D (Hara et al. 2018, Kataoka et al. 2020)); 2) Feeding the features into a one-directional model. In this regards, large number of approaches depend on Recurrent Neural Networks (RNN) (Singh et al. 2016, DiPietro et al. 2016), particularly, the Long Short Term Memory (LSTM) network, because of their notable ability of modeling sequence data in variable length. The gate mechanism of LSTM preserves temporal dependencies and drops irrelevant information during the training stage. However, LSTM-based methods only have limited ability of capturing long-term video context, due to the intrinsic vanishing gradient problem (Pascanu et al. 2013).

From another perspective, inspired by the success of temporal convolution in speech synthesis, Lea et al. (2017) introduce Temporal Convolutional Networks (TCNs) to segment and detect actions by hierarchically convolving, pooling, and upsampling input spatial features using 1-D convolutions and deconvolutions (see Fig. 2.7). The promising experiment results manifest that TCNs are capable of dealing with long-term temporal sequences (Lei and Todorovic 2018, Ding and Xu 2018). Nonetheless, the model handles information among local neighbors, thus showing incapabilities in catching global dependencies. Following this work, Farha and Gall (2019) suggest a multi-stage TCN, in which each stage is composed of several dilation layers, for action segmentation. Their work demonstrates the competence of dilated convolution (Oord et al. 2016) in hierarchically collecting multi-scale temporal information without losing dimensions of data.

2.2.3 Image Captioning

Image captioning, a class of sequence learning problem, is a task of automatically describing visual content for still images with natural human language. As such, it requires machine to understand and model the dependencies between visual and textual information and generate captions. Image captioning is a one-to-many task due to the reason that there are many possible captions correspond to one image. During the generation process, different

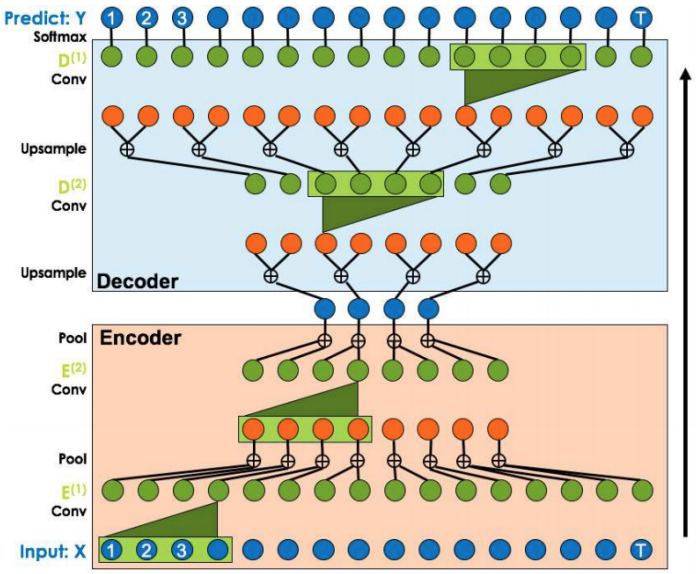


Figure 2.7: An encoder-decoder temporal convolutional network architecture to capture the temporal information from long and untrimmed sequence (Lea et al. 2017)

captions might focus on different parts of the image (Dai et al. 2017). This creates the huge challenge for designing the captioning models as well as evaluating them. The comprehensive review of sequence generation and image captioning can be found in (Staniūtė and Šešok 2019, Lipton et al. 2015).

A typical image captioning system has an encoder-decoder architecture that applies a CNN (convolutional neural network) to extract a high-level feature representation and a RNN (recurrent neural network) with a word vocabulary as the decoder. There are three popular choices of image encoder, VGG Network (Simonyan and Zisserman 2014b), ResNet (He et al. 2016), and Bottom-up features (Faster R-CNN with ResNet pre-trained on ImageNet and attribute features with Visual-Genome data) (Russakovsky et al. 2015, Krishna et al. 2017, Anderson et al. 2018). Few studies also select Google Net (Szegedy et al. 2015, Zhao et al. 2019), DenseNet(Huang et al. 2017, He et al. 2019b, Deng et al. 2020), or the novel Visual Commonsense R-CNN features (Wang et al. 2020) based on casual intervention. As for the language model, mainstream methods tend to choose a RNN network

(such as LSTM) with attention mechanism (Xu et al. 2015, Lu et al. 2017 2018) (Figure 2.8 are some soft-attention visualization results using (Xu et al. 2015)) or transformer based self-attention mechanism without any RNN (Pan et al. 2020, Herdade et al. 2019, Cornia et al. 2020).

From another perspective, sequence generation models are often trained in "Teacher-Forcing" (Bengio et al. 2015), which inputs the ground-truth to maximize the likelihood of next prediction during training and uses previously generated words from the model distribution to predict the next work during test time. In order to alleviate the mismatch between train and test and improve the evaluation performance, Rennie et al. (2017) propose the reinforcement learning based approach to directly optimize the Cider score. Rather than design a baseline to normalize the rewards and reduce the variance, they use the reward obtained by the current model under the inference time at the time to normalize the rewards. Later on, Many variants of self-critical sequence training have also been proposed. (Gao et al. 2019, Zhang et al. 2017b, Liu et al. 2017)

2.2.4 Evaluation Metrics for Image Captioning

Besides the sequence generation task, how to automatically evaluate the generated captions has become increasing important. The key idea is to measure the correlation of generated captions with human judgments. Following most of the image captioning methods, we apply BLEU (Papineni et al. 2002), Rouge-L (Lin 2004), METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam et al. 2015), and SPICE (Anderson et al. 2016) to evaluate our model, while the first three metrics are originated from machine translation and the last two are specifically designed for image captioning.

BLEU (Bilingual Evaluation Understudy) applies the modified precision to compare the similarity between the candidate sentence against one or more reference sentences. It is defined as:

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log p_n) \tag{2.1}$$

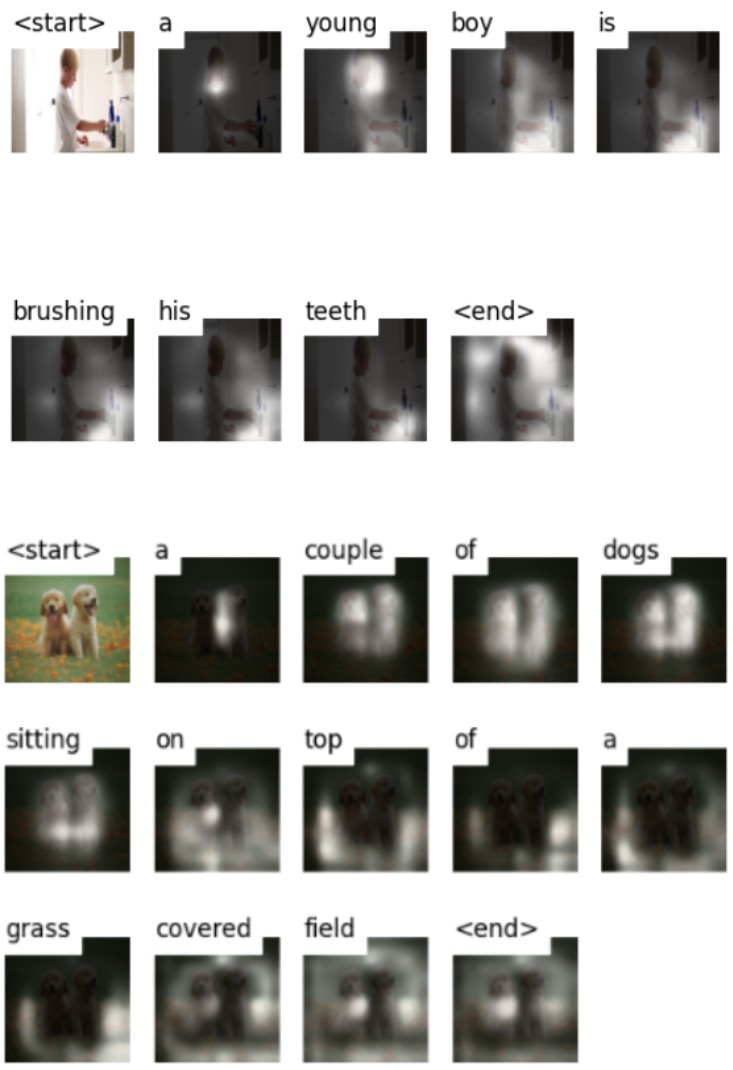


Figure 2.8: Visualization of attention states (Xu et al. 2015)

where n is n-gram and w_n is the weight. The equation stands for the geometric mean of weighted n-gram precision scores p_n multiplied by a brevity

penalty BP. In this regards:

$$p_n = \frac{\sum_i \sum_k min(h_k(c_i), max_{j \in m} h_k(s_{ij}))}{\sum_i \sum_k min(h_k(c_i))} \tag{2.2}$$

where $h_k(c_i)$ is number of times word k appear in generated sentence and $h_k(s_{ij})$ is number of times word k appear in reference sentence. However, the calculation of p_n works well for the short sentence, so the authors a penalty factor to penalize the short sentences.

$$BP = \begin{cases} 1 & \text{if } l_c > l_s \\ e^{1-\frac{l_s}{l_c}} & \text{if } l_c <= l_s \end{cases} \tag{2.3}$$

where l_c is the length of the generated sentence and l_s is the length of the reference sentence.

ROUGE-L ROUGE is originally proposed for summary evaluation and it only focus on recall. ROUGE has many different types, such as ROUGE-N, ROUGE-L and ROUGE-W. Here, we use ROUGE-L, which measures the longest common subsequences (LCS) between a pair of sentences to evaluate the model performance. Suppose X and Y are candidate and reference sentences of length m and n. Then we have:

$$\begin{aligned} P &= \frac{LCS(X,Y)}{m} \\ R &= \frac{LCS(X,Y)}{n} \end{aligned} \tag{2.4}$$

then the weighted harmonic mean of P and R is calculated as:

$$F = \frac{(1+\beta^2)RP}{R+\beta^2 P} \tag{2.5}$$

METEOR is another machine translation evaluation metric, which is claimed to have better correlation with human judgments than BLEU. Considering the Wordnet synonyms and paraphrase matching, it calculates the weighted F score of unigram matches between sentences and a penalty factor Pen for

incorrect word order.

$$\begin{aligned} METEOR &= (1 - Pen) \times F_{means} \\ F_{means} &= \frac{PR}{\alpha P + (1-\alpha)R} \\ P &= \frac{m}{c} \\ R &= \frac{m}{r} \end{aligned} \tag{2.6}$$

where α is an adjustable parameter, m is number of matched unigrams from the candidate sentence, and c and r is the length of candidate sentence and reference sentence, accordingly. And,

$$Pen = \gamma(\frac{ch}{m})^{\beta}, \quad where \quad 0 \leq \gamma \leq 1 \tag{2.7}$$

here ch is number of matching chunks between sentences. Under this circumstance, if most of the matches are continuous, there will be less chunks and lower penalty.

CIDEr is a recent proposed metric for evaluating image captioning based on consensus between candidate description c and the set of reference sentences S. The key idea behind CIDEr is regarding every sentence as a document and calculating its TF-IDF (term frequency-inverse document frequency) weight. Then the cosine similarity is measured between n-grams candidate and reference TF-IDF.

$$CIDEr(c, S) = \frac{1}{M} \sum_{i=1}^{M} \frac{g^n(c) \cdot g^n(S_i)}{\|g^n(c)\| \times \|g^n(S_i)\|} \tag{2.8}$$

where M is the amount of reference sentences and $g^n(.)$ represent the TF-IDF weight of n-gram. CIDEr gives more weights to important words and penalize the common words.

SPICE is also designed for image description similarity evaluation. It takes the semantic meaning into account and parses the candidate and reference sentences into scene graphs. Scene graph can be explained as a semantic representation of the given sentence with semantic tokens such as object class C, attributes A, and relations R. An example of scene-graph is shown in Fig. 2.9

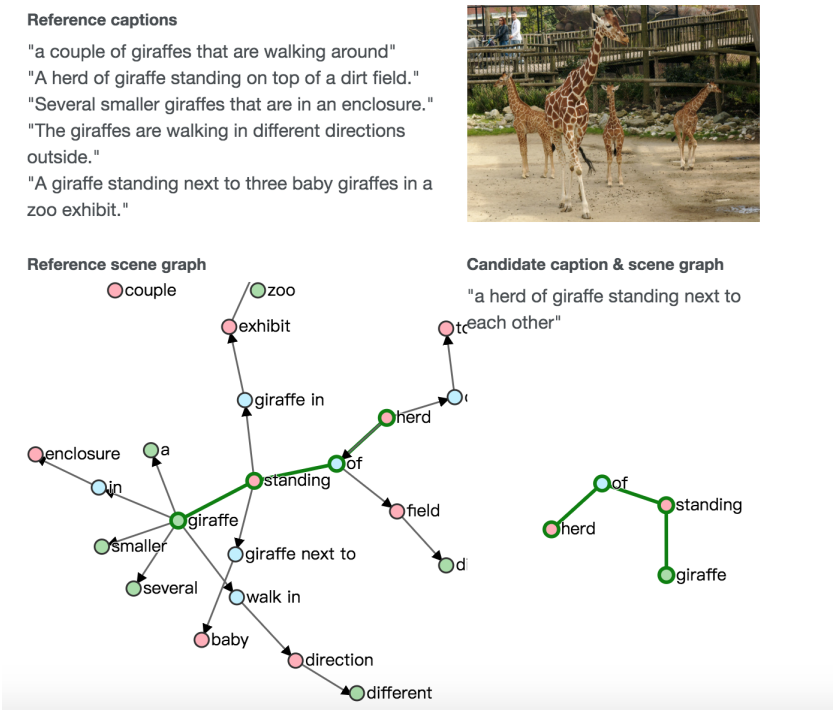


Figure 2.9: An example of scene graph (Anderson et al. 2016)

Chapter 3

Attention based Multi-task Surgical Skill Assessment

3.1 Introduction

The importance of teaching and assessing surgical skills before a trainee becoming a qualified surgeon cannot be overestimated. Despite the fast development of the computational and simulation methods, surgical skills are mainly assessed manually by experienced surgeons who need to monitor or check the whole operation process. Apparently, it is time-consuming and labor-intensive, and lack consistency and reliability. As a result, surgical education and training area would benefit from the idea of automatic surgical skills assessment.

Some recent studies artificially partition a surgical task into several predefined gestures. Then they apply the statistic model, such as Hidden Markov Model (HMM) (Tao et al. 2012) to find the latent structure for the whole process. Although the result is interpretative and promising, this requires huge human labour to label the video. Other studies address the automatic assessment task by extracting and combining the features from motion, speed, surgical instruments usage, etc. (Zia and Essa 2018). However, for one, it is quite unclear how these handcrafted features impact the expertise level of a surgeon. For another, possible significant features might be overlooked during the feature extraction process.

In the recent decade, the outstanding feature learning capability of deep learning algorithms provides possible solutions to extract highly discriminating visual features for surgical video understanding. This prominent technique also benefits the automatic surgical skill assessment task. Fawaz et al. (2018) leverage the 76 dimensional kinematic data with Multivariate Time Series (MTS) to estimate different skill levels for surgical tasks. In order to get the interpretative pattern, rather than the general fully connected layer on the last step, they use Global Average Pooling (GAP) to get Class Activation Map (MAP) (Zhou et al. 2016). Despite nearly 100 percent accuracy, additional devices and tracking systems are demanded to capture the data. Another work treats this task as a video classification problem (Funke et al. 2019b). The authors evenly choose few video snippets (consist of 64 consecutive video frames) and extract their spatial-temporal features by a 3D ConvNet. The ultimate decision is made by the consensus among the selected snippets. Nevertheless, the method cannot capture the long-range temporal information, and they mistakenly assume that all video parts equally contribute to the skill classification.

In this chapter, we propose a novel architecture motivated by the success of the attention mechanism (Bahdanau et al. 2014, Xu et al. 2015), specifically, self-attention model (Vaswani et al. 2017) for automatic surgical skills assessment. Our entire model structure can be seen in Figure 3.1. Taking a whole surgical video into consideration, we first extract its spatial-temporal information by a 3D ResNet (Hara et al. 2018). Next, driven by the expertise level prediction task, the attention network automatically collects the long-term temporal information and builds one-to-one relationships for every frame sequence. Our design is based on three observations and insights: (1) Surgical video data owns high-dimensional and abundant features, and it is much easier to obtain than other data format, for example, kinematics information or motion trajectory; (2) Temporal information is as significant as spatial features in a video sequence; (3) Some video parts or gestures are irrelevant whereas some are critical for skill evaluation. We validate our approach on the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) (Gao et al. 2014), including suturing, knot-tying and needle-passing.

In this regard, we claim our contribution as three-fold:

1. We propose a network architecture for automatic surgical skill assessment, which extensively consider and jointly combine the spatial, temporal and attention information in video frames.

2. Based on the attention mechanism, our model takes a whole video as the input. It intuitively shows superiority in handling long-range temporal signals.

3. To our best knowledge, we are the first to propose the multi-task learning for objective surgical skill assessment.

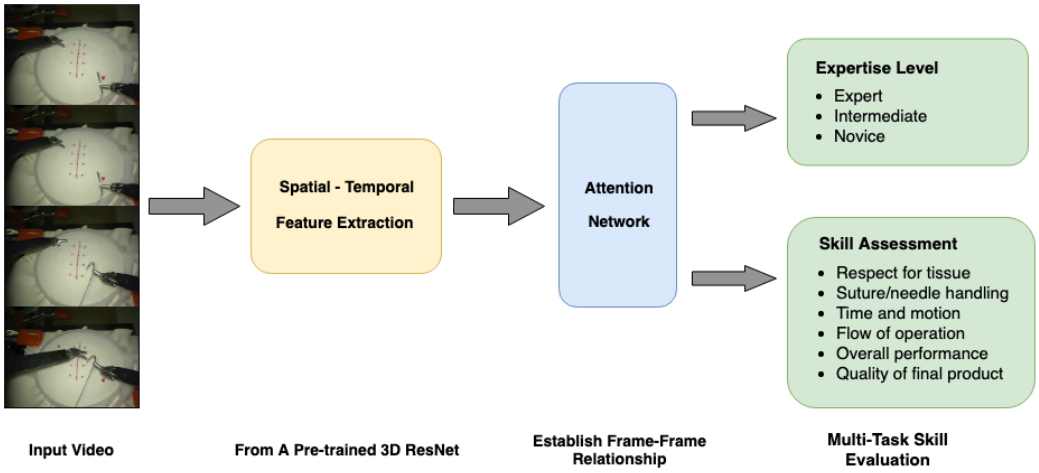


Figure 3.1: The spatial and motion features are extracted from a pre-trained network for an input video. Then the attention network builds the frame-to-frame relationship for input feature sequence. Finally, the outputs are the classification of the expertise level and the concrete skill assessment

3.2 Methodology

When we consider the video analysis, it is naturally to think of the Recurrent Networks (RNN), since RNNs are designed for the sequence with varaince length (DiPietro et al. 2016). However, the conventional RNN and LSTM

are computational inefficient, and learning a long-range dependency is still a challenge. In contrast, the attention model (Vaswani et al. 2017) solves above problems without using any CNN or RNN module. The effect of the attention mechanism can be explained as optionally seeking related information in regard to the intention.
In our work, we first apply a pre-trained 3D residual network to pre-process the input video (see section 3.3.2). Then the attention network permits each frame to look at all other positions in the input sequence to build the vector of importance. These dependencies focus on relevant information when predicting the expertise level and assess the technical skills for a surgical video. Next, we review the key components of our model in details.

3.2.1 Attention Network for Skill Assessment

Figure 3.2 illustrates the technical pipeline of our revised attention model for skill determination. Following the setting of original transformer encoder (Vaswani et al. 2017), the network is composed of N (in our work N = 6 following the original attention paper) identical blocks. Each block is further broken down into two sub-layers: the *Self Multi-head Attention layer* (see 3.2.2) and a simple position-wise fully connected feed-forward network. There is a residual connection and a layer normalization around each sub-layer. The residual and layer normalization operation is expressed as: $y = LayerNorm(x + Sublayer(x))$, where x is the input hidden state for each sub-layer, and y is the corresponding output. Before the attention network, extracted spatial-temporal features are first feeded into a *Positional Encoding* layer (see 3.2.3) to get the absolute position information.

3.2.2 Self Multi-head Attention

Self-attention, sometimes also called as intra-attention, is an attention mechanism (Cheng et al. 2016) to represent an input sequence itself by establishing one-to-all relationships among all positions. The basic self-attention function

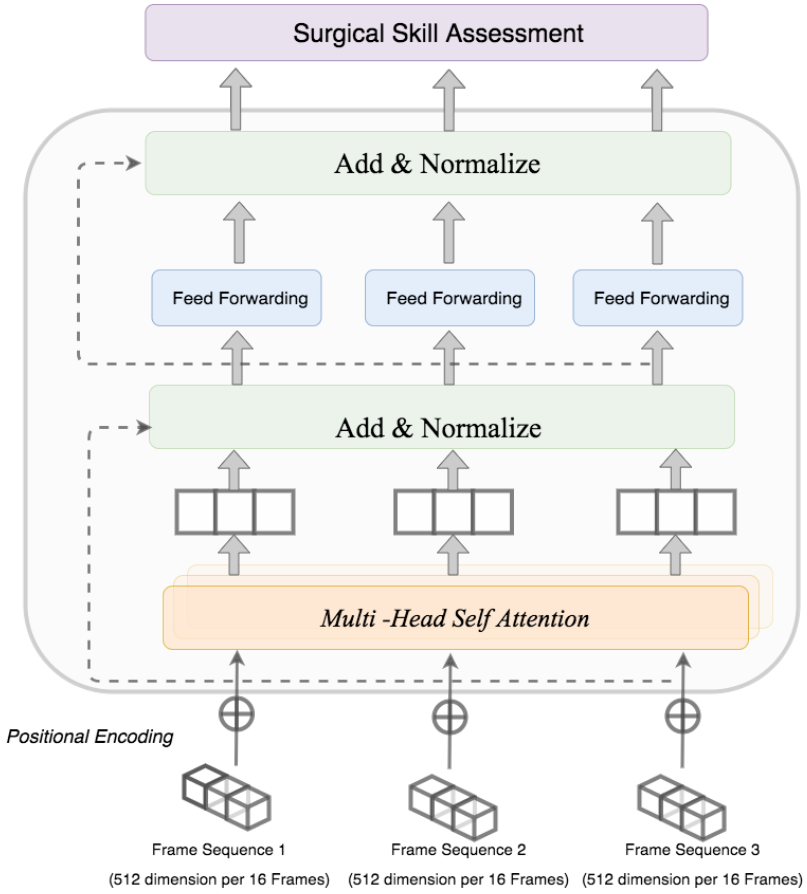


Figure 3.2: Attention network

is called **Scaled Dot-Product Attention**. It is computed as:

$$Attention(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{3.1}$$

where Q stands for queries packed as matrix, K and V are matrices of key-value pairs, and d_k is the dimension of queries and keys. In our case, keys, values and queries are input video features. The dot product is first calculated between query and all the keys to find the similarity, then divided by $\sqrt{d_k}$ to get stable gradients. Finally, a softmax function is applied to get the weighted average mapping from inputs. This is our way of building dependencies between every input frame sequence with all other frames.

In addition, rather than building just one type of correspondence, we use

multi-head attention. The model runs scaled dot-product multiple times in parallel. Using this method, input frames are projected to diverse representation sub-spaces by obtaining information from different positions.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3.2}$$

$$MultiH(Q, K, V) = Concat(h_1, ..., h_h)W^O \tag{3.3}$$

where matrices W_i^Q, W_i^Q, W_i^Q and W^O are projection parameters to be learned during the training phase. Different linear transformations are applied to the queries, values, and keys for each attention "head". In our work, we follow the original settings, 8 parallel heads, from the transformer model (Vaswani et al. 2017).

3.2.3 Positional Encoding

The attention network contains no positional information, since no recurrence nor convolution component are included. Intuitively, we use positional encoding (Vaswani et al. 2017) to explicitly encode the relative and absolute position of the original video sequence and added to the input frame features. The positional encoding is computed as:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{model}}\right) \tag{3.4}$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{model}}\right) \tag{3.5}$$

where pos represents the position and i is the dimension. More specifically, each dimension of the positional encoding is a sine wave with the wavelengths ranging from 2π to $10000 \times 2\pi$. This function enables the relative position easily to be referred due to the fact that any $PE[pos + k]$ can be easily depicted as a linear function of $PE[pos]$.

3.3 Implementation

3.3.1 Dataset Description

The JIGSAWS dataset is captured using da Vinci surgical system from eight surgeons with different levels of skill, that is *expert, intermediate, and novice.* They perform five repetitions for three elementary surgical tasks on a bench-top model: suturing (39 videos), knot-tying (36 videos) and needle-passing (28 videos), which are recognised as standard components of most surgical skills training curricula (Peters et al. 2004) as shown in Figure 3.3. Videos are captured at 30HZ with a resolution of 640×480 pixels. In addition, a gynecologic surgeon with extensive robotic and laparoscopic surgical experience watches each video and assigns a global rating score (GRS) using a modified OSATS for six elements, namely *respect for tissue, suture/needle handling, time and motion, flow of operation, overall performance, and quality of final product.* Due to the limited data and imbalanced label, we divide each of six elements into three levels: score 1-2 as poor performance, score 3-4 as fair performance, and score 5 as good performance.We expect our network automatically predicting the expertise level and evaluating six standards for an input video in parallel.

Figure 3.3: Shortcut of three tasks, from *left* to *right* are: *suturing*, *needle Passing* and *knot-tying* (Gao et al. 2014)

3.3.2 Pre-trained 3D ResNet

Although JIGSAWS is the largest open-source surgical skill assessment dataset, it is still relatively small for training a deep neural network from scratch. In-

spired by the recent success of transfer learning in image classification problem, we take the 3D ResNet (Hara et al. 2018) as a feature extractor. The 3D ResNet has been trained on the Kinetics dataset, which is one of the largest human action dataset including 400 action classes. The output is the spatial and temporal features of 512 dimensions (after global average pooling) for continuous 16 frames.

3.3.3 Implementation and Training Details

The model is trained on a single NVIDIA GeForce GTX 1080 graphics card. Each training batch contains a complete surgical video pre-processed by the 3D ResNet (512 dimension for 16 consecutive frames) as input. The network hyper-parameters settings are described throughout the paper.

Given that three tasks are different, we respectively fit these tasks by training three different models with the same network architecture. The suturing, needle passing and knot-tying tasks are trained 20 epochs with the learning rate at 0.01, 0.001, 0.0001, respectively. Same architecture and other hyper-parameters settings are shared. For the output of each sub-layer, the dropout is performed before each sub-layer, and the probability is set to $P_{drop} = 0.1$. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$, and the standard cross-entropy loss.

3.4 Evaluation

3.4.1 LOSO Evaluation

Originally, authors of JIGSAWS dataset define two cross validation schemes: *leave-one-supertrial-out (LOSO)* and *leave-one-user-out (LOUO)*. The former one (splited into five folds) every time lefts the i^{th} trial from eight surgeons for test and the rest for training. While the latter each time left all the trails from the *ith* surgeon for test and the rest for training. Although LOUO validation is efficient to test if a model works for a new subject, data from JIGSAWS is insufficient to support such a validation. One reason is because of the limited data size and imbalanced label. JIGSAWS only contains

8 subjects in total, including only two experts and two intermediate level surgeons. In one validation round, if we left one expert for testing and the other subjects for training, then we only have one expert in the training set. This is also the case for intermediate level surgeons. On the other hand, the official dataset defined the expertise level of a surgeon in accordance with the robotic surgical experience by hours: experts have more than 100 hours experience, intermediate subjects have 10 to 100 hours experience, and novices have less than 10 hours experience. Nevertheless, some of the intermediate surgeons get higher score than experts for their performance in skill annotation. Therefore, we follow the LOSO scheme as other related studies for three separate tasks.

As the result, the *Micro* average accuracy, *Macro* average recall, and average F_1 score are calculated for the predicted results. The *micrio* and *macro* strategies are defined in (Ahmidi et al. 2017). *Accuracy* is the percentage of correct predictions over total predictions. While *recall* is computed as the ratio between the correct predictions of a specific class and the total instances of this class. On the other hand, *precision* represents how many true positives are actually correct predictions among all the true positives predicted by the model. And finally, F_1 score is a weighted harmonic average between *precision* and *recall*.

We compare our experimental results with deep learning related state-of-the-art algorithms shown in Table 3.1. Notably, the first two approaches ((Fawaz et al. 2018) and (Wang and Fey 2018)) are on the basis of kinematics data from JIGSAWS, whereas (Funke et al. 2019b) and ours are video-based. Moreover, the overall performance of the knot-tying task is worse than the other two, hence we visualize the confusion matrix for knot-tying predictions accumulated five training validations (shown in Figure 3.4). The 3×3 confusion matrix C describes: for $C(i, j)$ where $i, j \in \{0, 1, 2\}$ how many times the class i are classified as class j. Class 0, 1, and 2 represents *expert* level, *intermediate* level, and *novice*, respectively.

Table 3.1: Expertise level prediction results on the suturing, needle passing and knot-tying tasks. The results are averaged over five validation runs for the test set. The result is measured in %

Methods	Suturing			Needle Passing			knot-tying		
	Accu.	*avg.recall*	*avg.F1*	*Accu.*	*avg.recall*	*avg.F1*	*Accu.*	*avg.recall*	*avg.F1*
CNN (Fawaz et al. 2018)	100	100	—	100	100	—	92.1	93.2	—
CNN+SVM (Wang and Fey 2018)	94.1	—	92.3	90.3	—	87.0	86.8	—	—
3D-CNN (Funke et al. 2019b)	100	100	100	100	100	100	95.1	94.2	95.0
Attention (proposed)	100	100	100	100	100	100	94.8	94.2	94.8

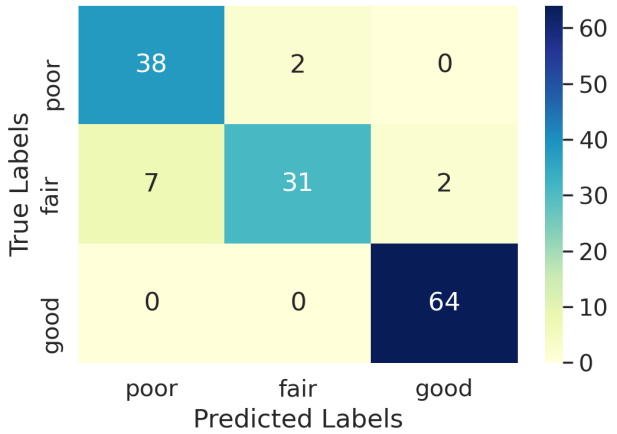


Figure 3.4: Confusion matrix for knot-tying

3.4.2 Multi-Task Learning

Apart from the assessment for expertise level, we also evaluate the six standards (respect for tissue, suture/needle handling, time and motion, flow of operation, overall performance, and quality of final product) in a multi-task manner. We divide each standard into three levels, score 1-2 as poor, score 3-4 as fair and score 5 as good performance. The output from attention model is utilized to predict the expertise level and six elements simultaneously. The standard cross-entropy loss function is also refined to achieve the multi-task prediction as:

$$loss = loss_e + loss_{multi} \tag{3.6}$$

$$loss_{multi} = \frac{1}{m}\sum_{i=1}^{6} loss_i \tag{3.7}$$

where the overall loss is the sum of loss of expertise level $loss_e$ and the mean average loss over six evaluation elements. Following the LOSO cross-validation scheme, the results for three tasks are presented in Table 3.2.

Table 3.2: Multi-task learning for elements of modified global rating score. The result is measured in %

	Expertise Level	Respect for tissue	Suturing needle handing	Time and motion	Flow of operation	Overall performance	Quality of final product
Suturing avg.accuracy	1.00	0.82	0.77	0.26	0.75	0.78	0.85
Needle_Passing avg.accuracy	1.00	0.75	0.82	0.43	0.74	0.83	0.56
Knot_tying avg.accuracy	0.95	0.86	0.77	0.31	0.86	0.91	0.89

3.5 Discussion

Consequently, our approach outperforms the first two ConvNet models, which relies on the 76 dimensional robot kinematics, whereas achieves more or less similar accuracy with the 3D-CNN model. It can be inferred that the pre-trained 3D CNN features together with the attention mechanism are capable of automatically learning complicated spatial temporal video features to assess the expertise level for a given subject sample. Although our model presents slightly lower accuracy (0.3%) and average F_1 score (0.2%) than the 3D-CNN (Funke et al. 2019b) method for the knot-tying task, it is possibly caused by the limited dataset size. Intuitively, we believe our model is more reasonable. The 3D-CNN model evenly extracts several video snippets cross the video and aggregates proposals from each snippet as the final prediction, thus it fails to catch the long-range temporal information. In contrast, our input is the whole surgical video, such that there is no temporal information lost, and the attention mechanism automatically learns relationships between the target and the input frame sequences. Another point is that our approach is much more faster. Both of the two approaches initialize the network input

with a pre-trained 3D ConvNet (Inception-v1 I3D for (Funke et al. 2019b) and 3D ResNet for our network). However, they train the TSN classifier for 1200 epochs while we only train our network no more than 20 epochs to achieve the similar performance.

Among three tasks, the knot-tying shows the lowest accuracy. The confusion matrices from Figure 3.4 mainly mis-classsify the expert and the intermediate level subjects. There are 7 intermediate surgeons mis-classified as expert. On the one hand, knot-tying is regarded as a rather complex task, which is also supported by the results from other studies. On the other hand, the dataset publishers defined the expertise level of a surgeon in accordance with the robotic surgical experience by hours: experts have more than 100 hours, intermediate subjects have 10 to 100 hours, and novices have less than 10 hours experience. Nevertheless, when we check the skill annotation, it is surprisingly to find that some of the intermediate surgeons get higher score than experts for their performance. Our multi-task idea is mostly inspired by this finding.

As for the multi-task predictions, Table 3.2 displays the satisfying predictions among the five individual elements, except the *Time and motion*. It only reaches 26%, 43% and 31% accuracy for three tasks separately, as the random guessing. One possible explanation is that the time and motion is not absolutely related to the skill determination, which need further validation. We further visualize the accumulative confusion metrics for six evaluation standards for suturing task (see Figure 3.5). It can be clearly seen that there is no pattern for *time and motion*, most of the poor and fair performance have been misclassified as good performance. From the result of other five predictions, although we achieve the satisfying classification results, the samples are suffer from data imbalanced problem. Most of the samples are annotated as the inter-mediate level performance such that the network lack the ability of predicting poor and good performance. Nevertheless, we believe that the multi-task method presents the more detailed and complete surgical skill assessment. By doing this, it can provide the meaningful and concrete guidance for a trainee to improve the surgical skills. We provide a baseline and discuss the limitations for current dataset.

In order to prove the efficiency of multi-task learning, we train the network to only predict six evaluation standards without classifying the expertise level. From Table 3.3, it can be seen that jointly training two tasks improve the overall performance.

Table 3.3: Multi-task learning for elements of modified global rating score. The result is measured in %

	Respect for tissue	Suturing needle handing	Time and motion	Flow of operation	Overall performance	Quality of final product
Suturing avg.accuracy	0.8	0.68	0.43	0.79	0.74	0.78
Needle_Passing avg.accuracy	0.69	0.78	0.44	0.76	0.75	0.63
Knot_tying avg.accuracy	0.80	0.79	0.39	0.78	0.88	0.86

3.6 Summary

In this chapter, we have designed an automatic surgical skill assessment framework based on the attention mechanism only using RGB video data. Before feeding a video sequence into the network, we extracted its spatial-temporal feature from the pre-trained 3D ResNet. With considering the inter-relationship between video frames, the suturing and needle passing tasks obtain the accuracy at 100% in testing, and the knot-tying reaches 94.8%. These competitive results denote that attention network allows video frames to focus on the relevant information according to the final target. We also evaluate our framework with not only the expertise level determination, but also in assessing the six elements from OSATS. Among all the predictions, five standards gains satisfied accuracy. This idea opens up new applications for surgical skill assessment such as the comprehensive performance report generation.

In our future work, besides the supplementary study of the multi-task learning, more intensively labeled video data and relevant augmentation algorithms are worthy to be built. Furthermore, rather than only using the RGB video, we will consider the optical flow as motion feature from the video due to its great performance on various video recognition tasks.

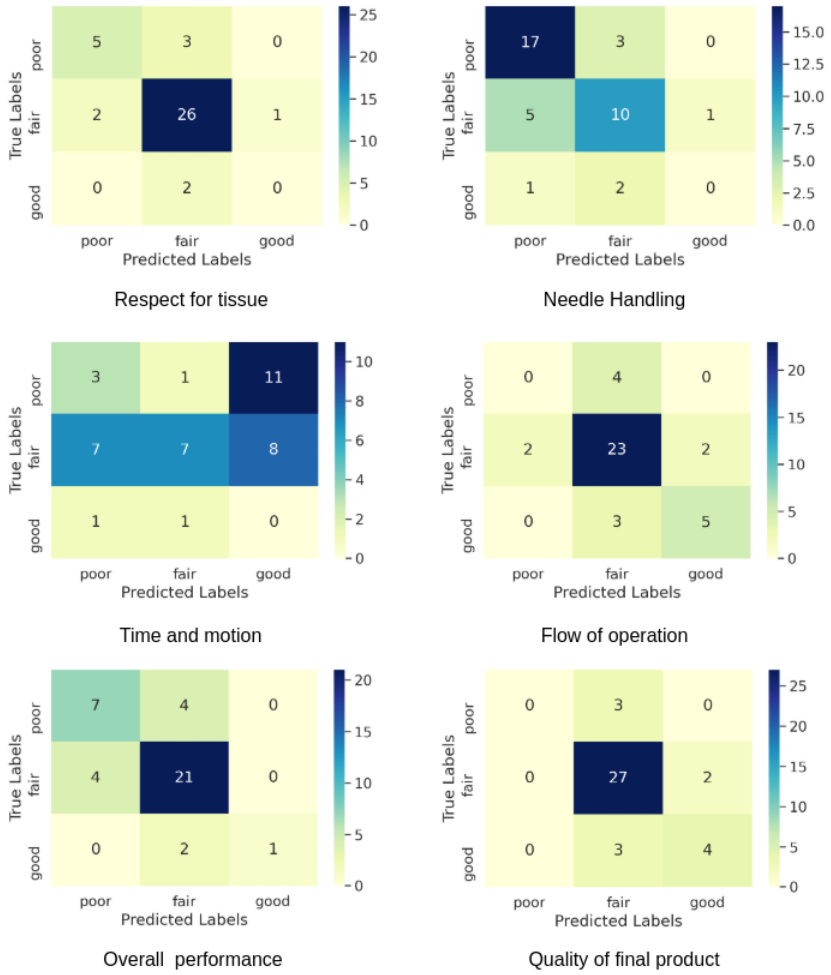


Figure 3.5: Accumulative confusion matrices over five cross-validation runs for suturing task

Chapter 4

Symmetric Dilated Convolution for Surgical Gesture Recognition

4.1 Introduction

Surgical gesture recognition is the process of jointly segmenting and classifying fine-grained surgical actions from surgical videos. It is crucial for surgical video understanding and building the context awareness system towards the next generation surgery (Maier-Hein et al. 2017). Many medical applications are included such as intra-operative computer assistance and objective surgical skill assessment. However, raw surgical videos are normally untrimmed and the operation environment is particularly complicated. Consequently, detecting surgical gestures from these surgical videos with high intra-class variance and low inter-class variance is inherently quite challenging.

Some of the prior works use probabilistic graphic models, for example, HiddenMarkov Models (HMMs) (Varadarajan et al. 2009, Sefati et al. 2015) and Conditional Random Fields (CRFs) (Mavroudi et al. 2018) to model the latent state transition. However, these approaches either require additional sensors to collect kinematics data or have limitations on capturing temporal information from long and untrimmed surgical videos. Recently, various deep learning techniques (recurrent neural networks (DiPietro et al. 2016), deep reinforcement learning (Liu and Jiang 2018), temporal convolutional neu-

ral networks (Lea et al. 2017)) have been applied to capture the long-range temporal patterns. For instance, in order to sequentially capture the video dynamics, Funke et al. (2019a) randomly sample video snippets (16 consecutive frames per snippet) and utilize a 3D Convolutional Neural Network (CNN) to extract the spatial-temporal features. But still, they only consider local continuous information. Because of the huge computational cost and GPU memory expenditure of 3D-CNN, they can only train the network at the clip level rather than inputted with the whole video (Zhang et al. 2020b).

To solve these difficulties, we propose a symmetric dilated convolution structure embedded with self-attention kernel to jointly detect and segment fine-grained surgical gestures. Figure 4.1 is an overview of our framework. Taking the extracted spatial CNN features from (Lea et al. 2017) as input, the encoder captures the long temporal information with a series of 1-D dilated convolutions to enlarge the temporal receptive field, followed by an attention block to establish the one-to-one relationship across all latent representations. Symmetrically, we devise our decoder with another set of dilation layers to map the latent representations back to each frame and predict the frame-wise gesture label. Unlike 3D-CNN learning features from partial sampled clips, our network takes the whole video into consideration. Owing to the symmetric dilated convolution structure with the enclosed self-attention kernel, not only can we learn the long-range temporal information, but also we can process neighbor and global relationship simultaneously.

With the above facts, we claim our contribution as two-fold. First, we propose a symmetric dilation architecture embedded with a self-attention module. It takes into account the long-term temporal patterns and builds frame-to-frame adjacent as well as global dependencies from the surgical video sequence. Second, we validate the effectiveness of our approach on a fundamental robotic suturing task from the JIGSAWS dataset. With the novel network architecture, our approach consistently exceeds the state-of-the-art method both on frame-level and on segmental-level metrics, improving the frame-wise accuracy **∼6 points**, and the F1@50 score **∼6 points**, which largely alleviates the over-segmentation error.

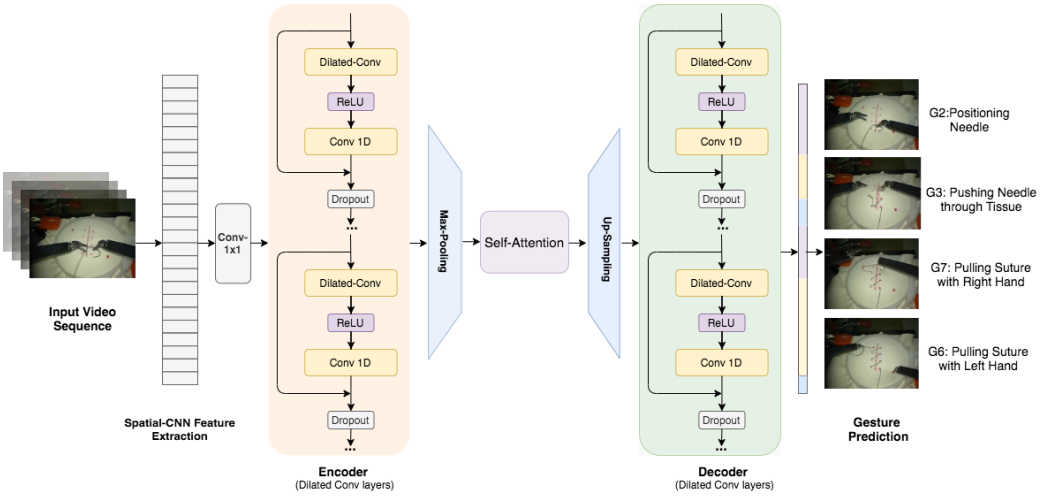


Figure 4.1: Overview of our architecture. Symmetric dilation network takes frame-level spatial-CNN features as input. The architecture can be divided into five steps: 1) 1-D convolution; 2) dilated convolution layers with max-pooling; 3) self-attention; 4) upsampling with dilated convolution layers; 5) frame-wise prediction.

4.2 Methodology

The architecture of our symmetric dilation network for surgical gesture recognition is detailed in this section (see Figure 4.2), which consists of two sub-structures: 1) the symmetric dilated Encoder-Decoder structure to capture long-term frame contents with memory-efficient connections (dilated layers) to aggregate multi-scale temporal information (see section 4.2.1); 2) the self-attention kernel in the middle to deploy the deep frame-to-frame relations to better discriminate the similarities among different frames (see section 4.2.2).

4.2.1 Symmetric Temporal Dilated Convolution

Temporal dilated convolution is a type of convolution applied on the input sequence with a defined sliding gap, which increases the temporal receptive field with less parameters (Oord et al. 2016, Lea et al. 2017, Farha and Gall 2019). In our study, we use blocks of identical dilation layers to capture and aggregate the video dynamics in different time scale. The first layer of

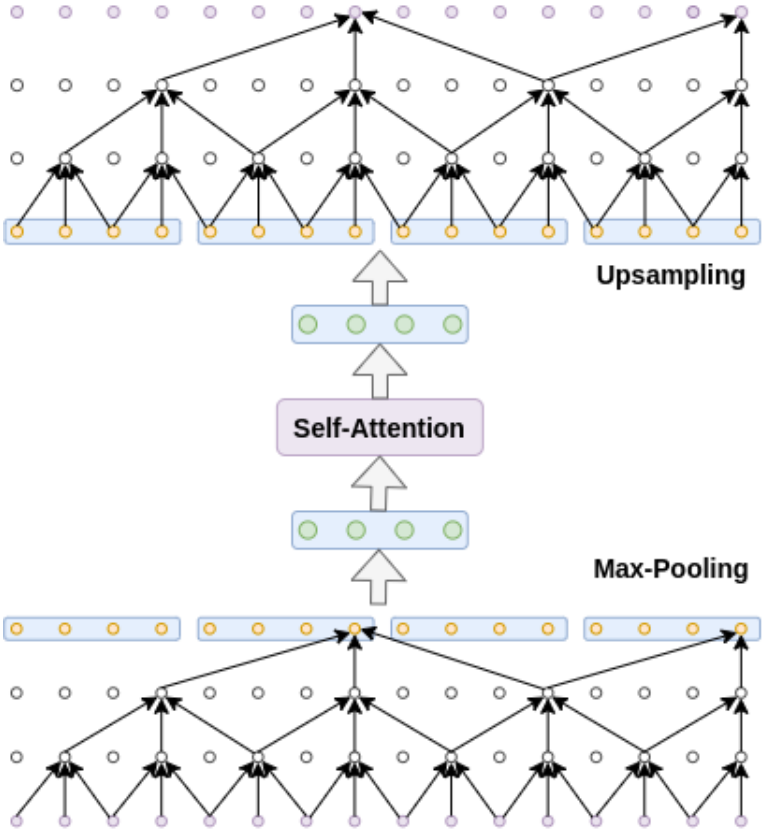


Figure 4.2: Symmetric temporal dilated convolution. With the layer number increasing, the size of the temporal receptive field grows exponentially.

the encoder is a 1×1 convolution to map the dimension of input spatial-CNN features to number of kernels f, followed by l layers of temporal dilated convolutions, where the dilation rates $\{s_l\}$ are set to $s_l = 2^l, l = 0, 1, ..., 9$. Because our target is the off-line recognition, we follow the details in (Farha and Gall 2019) by using acausal mode with kernel size at 3. Furthermore, we apply the non-linear activation function ReLU to each dilation output followed by a residual connection between the layer input and the convolution signal. The temporal dilated procedure can be formulated as follows:

$$\hat{E}_l = \text{ReLU}(W_1 * E_{l-1} + b_1) \tag{4.1}$$

$$E_l = E_{l-1} + W_2 * \hat{E}_l + b_2 \tag{4.2}$$

where E_l is the output of l-th encoder layer, $*$ is the temporal convolutional operation, $W_1 \in \mathbf{R}^{f\times f\times 3}$, $W_2 \in \mathbf{R}^{f\times f\times 1}$ represent the weights of a dilated convolution and the weights of a 1×1 convolution with f convolutional kernels, respectively. $b_1, b_2 \in \mathbf{R}^f$ are denoted as their corresponding biases. In every dilation layer l, the receptive field R grows exponentially to capture the long range temporal pattern, expressed as: $R(l) = 2^{l+1}-1$. By doing this, the temporal information on different scale is hierarchically aggregated while keeps the ordering of sequence. We also employ a 4×1 max-pooling layer behind the encoder dilation block to efficiently reduce the oversegmentation error (see our ablative study results in Table 4.2).

Our symmetric decoder has a similar structure with the encoder block, except that the max-pooling operations are replaced with a 1×4 upsampling. To get the final prediction, we use a 1×1 convolution followed by a softmax activation after the last decoder dilated convolution layer:

$$Y_t = \text{Softmax}(W * D_{L,t} + b) \tag{4.3}$$

where Y_t is the prediction at time t, $D_{L,t}$ is the output from the last decode dilated layer at time t, $W \in \mathbf{R}^{f\times c}$ and $b \in \mathbf{R}^c$, where $c \in [1, C]$ is the surgical gestures classes. Eventually, we use the categorical cross-entropy loss for the classification loss calculation. The encoder-decoder architecture is designed in this way to hierarchically accumulate the spatial features from different temporal span with memory efficient dilated convolutions.

4.2.2 Joint Frame-to-Frame Relation Learning with Self-Attention

The TCNs have shown consistent robustness in handling long temporal sequences with using ***relational features*** among frames. However, current methods (Ding and Xu 2017, Lea et al. 2017) only consider relations in local neighbors, which could undermine their performance in capturing relational features within a longer period. To obtain the global relationship among frames, it is essential to build frame-to-frame relational features with a non-local manner in addition to our encoder-decoder dilated convolutions.

With this insight, we introduce the non-local self-attention module to extract discriminate spatial-temporal features for better prediction.

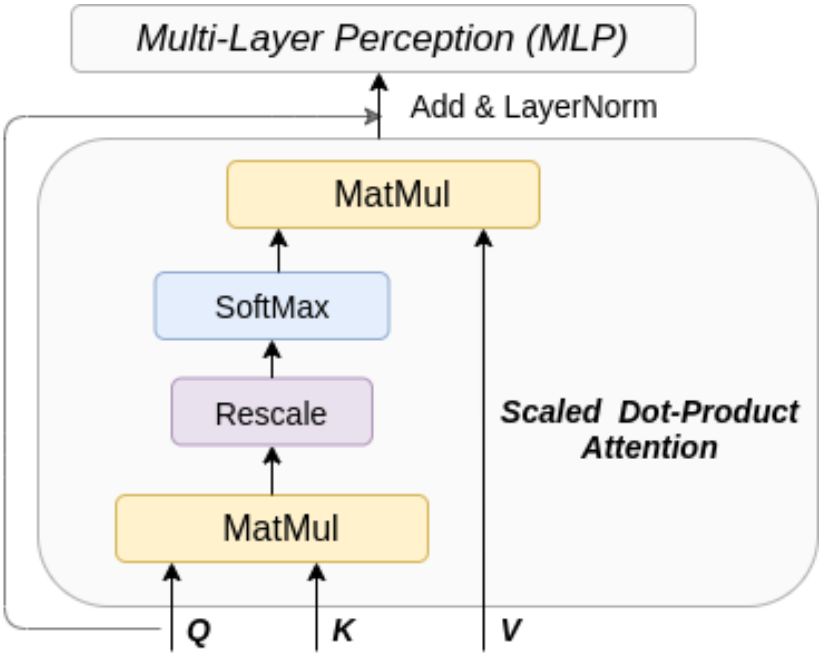


Figure 4.3: Self-attention block

Self-attention, or intra-attention refers to an attention mechanism, which attends every position of the input sequence itself and build one-to-one global dependencies. This idea has been widely used in Natural Language Processing (NLP) (Vaswani et al. 2017), Object Detection and Segmentation (Wang et al. 2018, Hu et al. 2018), etc. The key component of self-attention is called **Scaled Dot-Product Attention** (Vaswani et al. 2017), which is calculated as:

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{4.4}$$

where Q is a packed *Query* matrix, K and V stand for *Key-Value* pairs, and $\sqrt{d_k}$ is the feature dimension of queries and keys. The structure of the self-attention is shown in Figure 4.3. In our work, the input *Queries, Keys, and Values* to the self-attention module are the same, that is the output hidden temporal states from the encoder downsampling. The first step is to take the dot product between the query and the key to calculate the similarity.

This similarity determines the relevance between all other frames from the input sequence to a certain frame. Then, the dot product is rescaled by $\sqrt{d_k}$ to prevent the exploding gradient and followed by a softmax function to normalize the result. The intention of applying the softmax function here is to give relevant frames more focus and drop irrelevant ones. Eventually, the attention matrix is multiplied by the value and summed up. There is a residual connection followed by a layer normalization to feed the result to next two fully connected 1-D concolutional layers (see Figure 4.3). In this manner, frame-to-frame global dependencies are constructed.

4.3 Evaluation

4.3.1 Experiment Settings

Dataset Description: We evaluate our approach on an elementary suturing task from JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) (Gao et al. 2014), a robotic assisted bench-top model collected using *da Vinci* surgical system. To our best knowledge, JIGSAWS is the only public dataset for fine-grained surgical gesture recognition. There are 39 videos performed by eight surgeons with three skill levels. Ten different fine-grained surgical gestures for example, *pushing needle through tissue and oriental needle* for suturing task are manually annotated by an experienced surgeon. We follow the standard *leave-one-user-out* (LOUO), a 8-fold cross validation scheme for evaluation. In each fold, we leave one surgeon out for testing to verify if the recognition model works for an unseen subject. For the network input, we use the 128 dimensional spatial-CNN features extracted by (Lea et al. 2017) with 10 FPS. Given a video sequence $v \in V$ with length T: $v_{1:T} = (v_1, ..., v_T)$, our goal is to assign the corresponding gesture label $g \in \mathbf{G}$ to each frame: $g_{1:T} = (g_1, ..., g_T)$.

Implementation and Training Details: The model is implemented based on Pytorch and trained on a single NVIDIA GeForce GTX 1080 graphics card. For the symmetric dilated convolution, we set the layer number to

10 (see the supplementary material for the hyperparameter tuning experiment) and the channel number to 128 with the kernel size 3 followed by a dropout after each layer. In regard to the attention module, the feature dimension of queries and keys is set to 16. The network is trained for 30 epochs with the learning rate at 0.01. In addition, we apply Adam Optimizer such that $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$.

Evaluation Metrics: We adopt three evaluation metrics in our experiments: **frame-wise accuracy**, **edit score**, and **segmented F1 score**. Frame-wise accuracy is to measure the performance in frame level. However, long gesture segments tend to have more impact than short gesture segments, and the frame-wise accuracy is not sensitive to the oversegmentation error. Therefore, we use the edit score and F1 score to assess the model at segmental level. Edit score is defined as the normalized Levenshtein distance between the prediction and the groundtruth. While F1 score is the harmonic mean of precision and recall with the threshold 10%, 25%, and 50% as defined in (Lea et al. 2017).

4.3.2 Comparison with the State-of-The-Arts

Table 4.1 compares our symmetric dilation network with other state-of-the-art methods. It can be seen that our model achieves the best performance in all three metrics. Among other approaches, the baseline model **Bi-LSTM** reaches the relative lower performance than other methods indicating that the traditional RNN-based method is incapable of handing long video sequence. **Deep Reinforcement Learning (RL)** method trains an intelligent agent with reward mechanism and achieves the high edit 87.96 and F1 score 92.0, but the low frame-wise accuracy at 81.43%, which shows its inadequacy in capturing the global similarities throughout the frames. The latest **3D-CNN** method obtains the fair frame-wise accuracy at 84.3%, but it only obtains 80.0 for the edit score. This reflects that the model based on clip-level is still inefficient in catching long temporal relationship such that it suffers from the oversegmentation error.

While our model reaches the best frame-wise accuracy at 90.1% as well as the highest edit and F1 score at 89.9 and 92.5, respectively. It demonstrates that our model is able to capture the long-range temporal information along with the frame-to-frame global dependencies.

Table 4.1: Comparsion with the most recent and related works for surgical gesture recognition. Acc., Edit, and F1@10, 25, 50 stand for the frame-wise accuracy, segmented edit distance, and F1 score, respectively

JIGSAWS (Suturing)	**Acc.**	**Edit**	**F1@10**	**F1@25**	**F1@50**
Bi-LSTM (Singh et al. 2016)	77.4	66.8	77.8	-	-
ED-TCN (Lea et al. 2017)	80.8	84.7	89.2	-	-
TricorNet (Ding and Xu 2017)	82.9	86.8	-	-	-
RL (Liu and Jiang 2018)	81.43	87.96	92.0	90.5	82.2
3D-CNN (Funke et al. 2019a)	84.3	80.0	87.0	-	-
Symmetric dilation (w. pooling)	**90.1**	**89.9**	**92.5**	**92.0**	**88.2**

4.4 Discussion

4.4.1 Effectiveness of Submodules

To further investigate the functionality of each submodule in our method, we conduct ablative studies with five configurations as follows. As our network consists of a symmetric dilation structure with a self-attention kernel in the middle. We decouple it into a head dilation module, a tail dilation module, and the self-attention kernel to explore their joint effects.

(1) Self-attention module only (baseline)

(2) Baseline + head dilated convolution

(3) Baseline + tail dilated convolution

(4) Baseline + symmetric dilated convolution

(5) Baseline + symmetric dilated convolution + pooling

We apply these settings to segment and classify the surgical gestures and measure their **frame-wise accuracy**, **edit score**, and **segmented F1 score** separately. The experiment results are shown in Table 4.2.

Table 4.2: Ablative experiment results show the effectiveness of each sub-model. Acc., Edit, and F1@{10, 25, 50}, stand for the frame-wise accuracy, segmented edit distance, and F1 score, respectively

JIGSAWS (Suturing)	**Acc.**	**Edit**	**F1@10**	**F1@25**	**F1@50**
Self-attn only	87.8	44.0	54.8	53.5	49.0
Head dilation + attn	90.8	76.9	82.5	81.8	79.3
Tail dilation + attn	90.5	77.9	83.4	83.4	79.7
Symmetric dilation + attn	90.7	83.7	87.7	86.9	83.6
Symmetric dilation (w. pooling)	90.1	**89.9**	**92.5**	**92.0**	**88.2**

(1) only: Self-attention module can achieve promising frame-wise accuracy at 87.8%, but with very low edit distance (44.0) and F1 scores. It can be concluded that attention module is robust for classification tasks while missing the long temporal information.

(1) v.s. (2) and (3): We put the temporal dilated convolution structure before and after the self-attention module and get the similar results. The results have huge improvement in edit score and F1 score with different threshold, increase around 30% in each metric. It states that temporal convolution is capable of catching long temporal patterns.

(4): The obvious improvement on the segmental level evaluation shows that the symmetric encoder-decoder dilation structure helps capture the high-level temporal features.

(5): Max-pooling and upsampling further improve the edit distance and F1 score at segmental level such that smooth the prediction and allievate the oversegmentation problem.

Above controlled experiments verify the indispensability of each component for our proposed architecture. From frame-level view, self-attention mechanism is feasible to build non-local dependencies for accurate classification. And from the segmental-level perspective, symmetric dilation with

pooling is a viable solution for recognizing gestures from long and complicated surgical video data.

In addition, Figure 4.4 and Figure 4.5 are the list of gesture labels and the visualization result of our ablative experiments, respectively. From Figure 4.5, it can be seen that the self-attention can classify most of the frames correctly, but it suffers from the over-segmentation problem. Dilation layers and pooling mechanism further smooth the classification result. Taking a deeper look at the confusion metrics from one validation run (see Figure 4.6), there are 60 frames of G1 has been mis-classified as G2 and 47 frames of G5 has been mis-classified as G3. This mainly because G1 and G2, G5 and G3 are consecutive actions. It is hard to precisely identify the boundary between two continuous gestures.

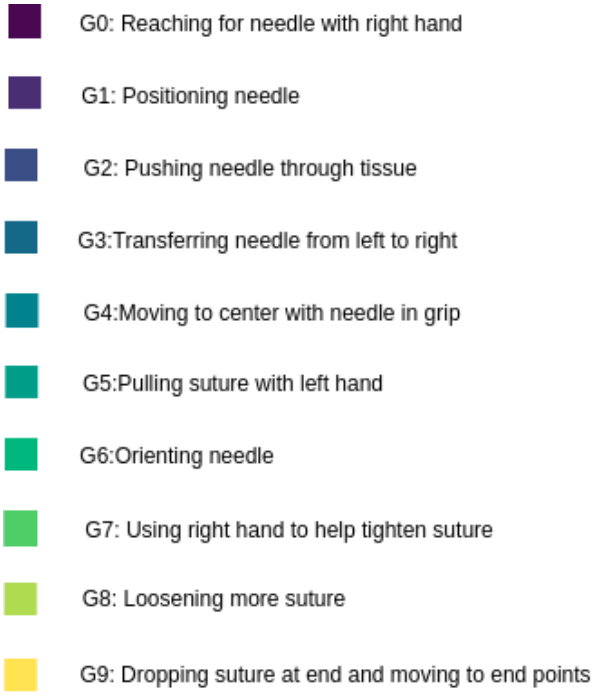


Figure 4.4: List of gestures in suturing task

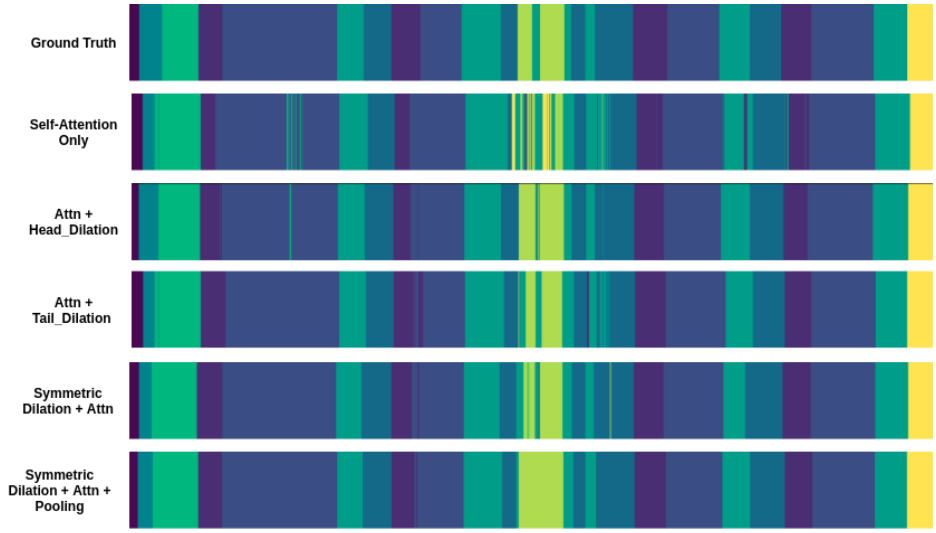


Figure 4.5: Visualization of ablative experiments.(0) ground truth; (1) self-attention module only (baseline); (2) baseline + head dilated convolution; (3) baseline + tail dilated convolution; (4)baseline + symmetric dilated convolution; (5) baseline + symmetric dilated convolution + pooling.

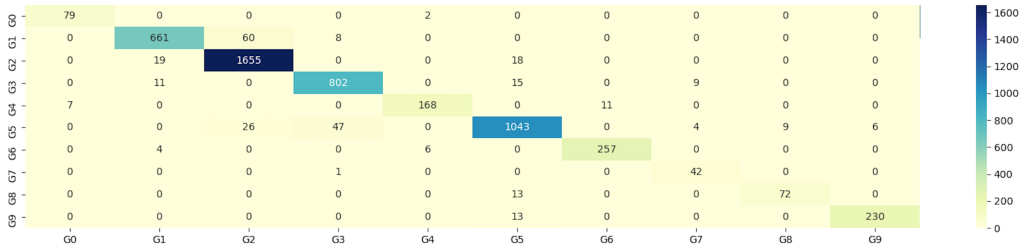


Figure 4.6: Confusion metrics from one validation run.

4.4.2 Effectiveness of Number of Dilation Layers

In our experiments, we fix both encoder and decoder dilated convolution layer number to 10. In Table 4.3 and Figure 4.7, we set the layer number l to 2, 6, 10, 14 both in encoder and decoder block to explore the impact of the receptive field size. Increasing the layer number from 2, 6, 10, improves the performance over all evaluation metrics. This is due to the increase of the receptive field. However, the performances decrease in framewise accuracy, F1 scores and edit score when we increase the layer number from 10 to 14.

When $l = 14$, the receptive filed is 16383, which is much larger than number of frames. It can be demonstrated that large receptive field can capture changes over a wider area, but cause a less accurate perception. Among all the settings, the configuration of 10 symmetric dilation layers achieves the best results.

Table 4.3: Ablative experiment results show the effect of the number of dilation layers (i.e the size of receptive field). Acc., Edit, and F1@{10, 25, 50} stand for the frame-wise accuracy, segmented edit distance, and F1 score, respectively.

JIGSAWS (Suturing)	**Acc.**	**Edit**	**F1@10**	**F1@25**	**F1@50**
2 Layers	89.6	75.0	82.2	81.3	81.3
6 Layers	90.6	88.2	91.5	91.0	87.6
10 Layers	90.1	**89.9**	**92.5**	**92.0**	**88.2**
14 Layers	89.9	86.4	90.6	89.6	86.3

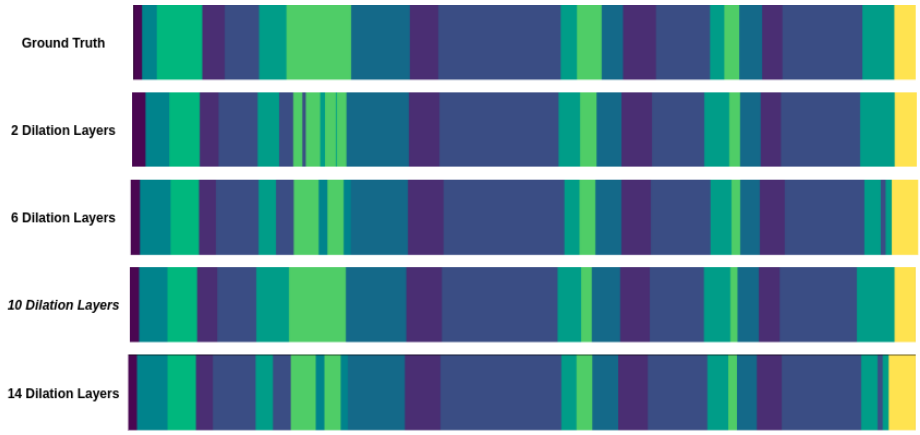


Figure 4.7: Influence of different number of dilation layers. We set the layer number l to 2, 6, 10, 14 both in encoder and decoder dilation block.

4.5 Summary

In this chapter, we propose a symmetric dilated convolution network with self-attention module embedded to jointly segment and classify fine-grained

surgical gestures from the surgical video sequence. Evaluation of JIGSAW dataset indicates that our model can catch the long-term temporal patterns with the large temporal receptive field, which benefits from the symmetric dilation structure. In addition, a self-attention block is applied to build the frame-to-frame relationship to capture the global dependencies, while the temporal max-pooling and upsampling layer further diminish the oversegmentation error. Our approach outperforms the accuracy of state-of-the-art methods both at the frame level and segmental level. Currently, our network is designed with an off-line manner in acausal mode, and we will explore the possibility of improving and applying it for real-time surgical gesture recognition in the future work.

Chapter 5

Surgical Instruction Generation

5.1 Introduction

The revolution of minimally invasive surgery has accompanied by a number of controversies. Besides the benefits of small incisions on the skin, the short recovery period, and cost effective, it has high demand of the surgical skills, requires long training curves, and has high initial complication rate for a novice surgeon. Under this circumstance, providing intra-operative surgical instructions by expert surgeons is imperative when the on-site mentoring is unavailable or insufficient.

Previously, telementoring (Challacombe et al. 2006), which exchanges medical information through video and audio in real time, has been proved as an efficient solution for intra-operative guidance, including pointing out target anatomical structure from the monitor, controlling the camera or the robotic arm, etc. Nonetheless, telementoring is limited by the cost of specific equipment and software, the high demand of transport speed, and legal and ethic issues (Bilgic et al. 2017, Erridge et al. 2019). With the huge development of the related techniques for context awareness assistance, understanding and analyzing the surgical activities in the operation room (OR) opens up the possibility of identifying and providing assistance for surgeons intra-operatively.

Automatic surgical instruction is the process of recognizing and analyzing the surgical activity and generating instruction for surgeons. Most of

existing literatures for surgical workflow analysis focus on surgical phases and fine-grained gestures recognition (Twinanda et al. 2016a, Funke et al. 2019a, Zhang et al. 2020a). However, these methods can be regarded as the classification problem based on pre-defined phases and gestures, thus have no ability of generating the unseen instructions. The most related research topic for us is the medical report generation (Jing et al. 2017, Chen et al. 2020, Bustos et al. 2020), which describes the ***impression***, ***findings***, ***tags***, etc. of a patient in reference to the radiology or pathology. But medical reports always follow similar writing template, while surgical activities data have high heterogeneity even for the same type of surgery on account of different surgical level, medical condition, and patient specific situation.

To our best knowledge, Rojas-Muñoz et al. (2020) is the only prior work for surgical instruction generation. In their work, the authors create the Database for AI Surgical Instruction dataset (DAISI) and use a bidirectional recurrent neural network (RNN) to generate the description for a surgical image. However their work has two limitations. For one, although RNNs are designed for sequence generation with arbitrary length, they suffer from the essential vanishing gradient problem Pascanu et al. (2013). For another, they apply the BLEU score as the only evaluation metric, which is insufficient for natural language evaluation.

In this chapter, inspired by the great performance of transformer model in machine translation Vaswani et al. (2017) and image captioning Cornia et al. (2020) from the open domain, we build our network with an encoder-decoder fully backboned with transformers to predict surgical instructions. Taking an surgical image as the input, we first extract its visual attention features by a fine-tuned ResNet-101 module. Then the encoder attention blocks, decoder attention blocks, and encoder-decoder attention blocks model the dependencies for visual features, textual features, and visual-textural relational features, respectively. On the other hand, sequence generation models are often trained using the cross-entropy (XE) loss and evaluated using non-differential metrics such as BLEU, CIDEr, etc. In order to alleviate the mismatch between training and testing and improve the evaluation performance, we apply the reinforcement learning based self-critical approach Rennie et al.

(2017) to directly optimize the CIDEr score. Experimentally, we extensively explore the performance of different baselines (LSTM-based fully connected and soft-attention models) on DAISI dataset Rojas-Muñoz et al. (2020). The experiments demonstrate that our transformer-backboned architecture outperforms the existing methods as well as our other proposed baselines. The promising instructions generated from the network bring potential value in clinical practice.

5.2 Methodology

5.2.1 LSTM-based captioning models

In order to explain the LSTM-based instruction prediction model, we first introduce the mechanism behind LSTM. Recurrent neural networks (RNN) are designed for modeling sequence data with arbitrary input/output length. However, if a sequence is long enough, the vanilla RNN cannot convey information from the very early beginning. And for the back propagation, it suffers from serious vanishing gradient problem. LSTM (Long Short Term Memory) network (Hochreiter and Schmidhuber 1997) is a type of RNNs, which is capable of carrying long-term temporal dependencies. It depends on cell state and various gates to solve short-memory problem. Typically, a LSTM cell consists of a cell state $c'^{<t>}$ to remember the relevant information throughout the whole processing time, an input gate $i^{<t>}$ to update the cell state, a $f^{<t>}$ forget gate to decide if the information should be kept or discard, and an output gate $o^{<t>}$ to decide the next hidden state. Figure 5.1 and equation 5.1 explain the detailed LSTM mechanism.

$$
\begin{aligned}
c'^{<t>} &= tanh(W_c[a^{<t-1>}, x^t] + b_c) \\
i^{<t>} &= \sigma(W_i[a^{<t-1>}, x^t] + b_i) \\
f^{<t>} &= \sigma(W_f[a^{<t-1>}, x^t] + b_f) \\
o^{<t>} &= \sigma(W_o[a^{<t-1>}, x^t] + b_o) \\
c^{<t>} &= i^{<t>} * c'^{<t>} + f^{<t>} * c^{<t-1>} \\
a^{<t>} &= o^{<t>} * c^{<t>}
\end{aligned}
\tag{5.1}
$$

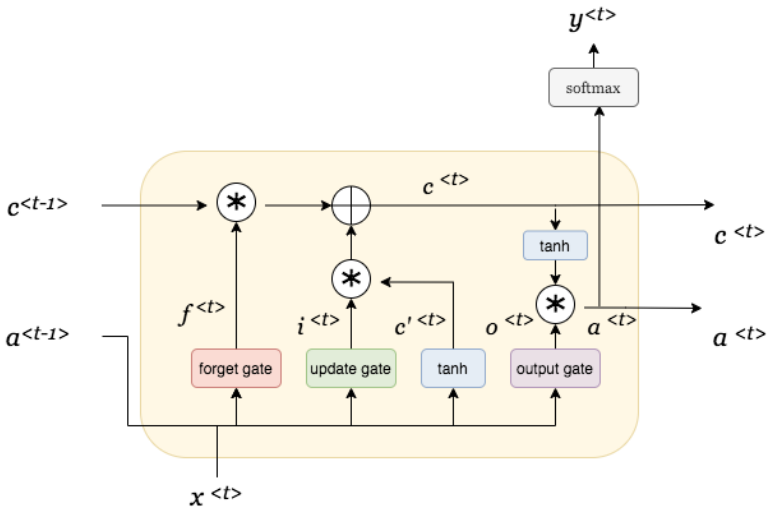


Figure 5.1: LSTM cell

FC model. In a similar way to (Vinyals et al. 2015), we first extract the visual feature for an medical image I from the FC layer of a deep CNN. Before feeding the feature into the LSTM, it is embedded by a linear projection W_I and is considered as the first word $W_I CNN(I)$. As for the text information, each word, which is represented as a one hot vector, is embedded by a linear embedding E with the same dimension as W_I.

$$
\begin{aligned}
x^{<t>} &= E1_{w_{t-1}} \; for \; t \geq 1, \; x^1 = W_I CNN(F) \\
c'^{<t>} &= tanh(W_c[a^{<t-1>}, x^t] + b_c) \\
i^{<t>} &= \sigma(W_i[a^{<t-1>}, x^t] + b_i) \\
f^{<t>} &= \sigma(W_f[a^{<t-1>}, x^t] + b_f) \\
o^{<t>} &= \sigma(W_o[a^{<t-1>}, x^t] + b_o) \\
c^{<t>} &= i^{<t>} * c'^{<t>} + f^{<t>} * c^{<t-1>} \\
a^{<t>} &= o^{<t>} * c^{<t>} \\
s^{<t>} &= W_s a^{<t>}
\end{aligned}
\tag{5.2}
$$

Soft-attention model. Rather than assume all spatial area has the same contribution to the next word prediction, we use the soft-attention (Xu et al. 2015, Rennie et al. 2017) to estimate the specific image region needs to pay

attention at each time step. The cell state $c'^{<t>}$ from Figure 5.2 is changed to: $c'^{<t>} = tanh(W_c[a^{<t-1>}, x^t, z^t] + b_c)$ where z^t is the attention image feature derived by soft-attention as defined in (Xu et al. 2015). Soft-attention (Bahdanau et al. 2014) is a deterministic, differentiable mechanism, which calculates the weights and applies the weighted average of the features across all pixels according to the next word prediction. Given a image feature map with N locations $\{z^1, \dots z^N\}$, then we have

$$\begin{aligned} z^{<t>} &= \sum_{i=1}^{N} \alpha^{<t,i>} z^{<i>} \ where \\ \alpha^{<t,i>} &= \frac{exp(e^{<t,i>})}{\sum_{i=1}^{N} exp(e^{<t,i>})} \ and \\ e^{<t,i>} &= Wtanh(W_e[a^{<t-1>}, z^i] + b_e) \end{aligned} \tag{5.3}$$

In these two methods, $a^{<0>}$ and $c^{<0>}$ are initialized to zero. The output from LSTM is a distribution over the next word $w^{<t>}$: $w_{<t>} \sim softmax(s^{<t>})$. Figure 5.2 is the framework of surgical instruction generation with LSTM.

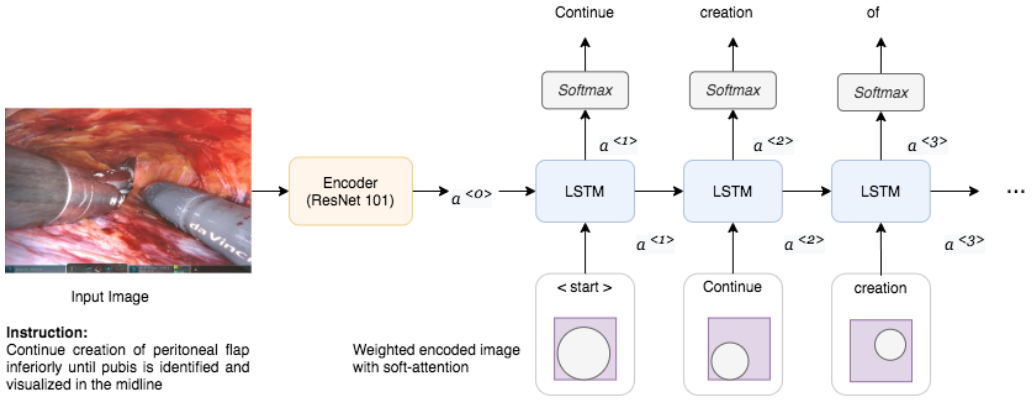


Figure 5.2: The image is encoded by a CNN and input into a LSTM network. For every time step, the last hidden state, the groundtruth word, and the attention from the weighted average across the image work together to predict the next word.

5.2.2 Transformer Captioning Model

In early chapters (Chapter 3 and Chapter 4), we use the self-attention block from transformer model to build framewise relationships for long sequence. In this chapter, we apply the whole transformer encoder-decoder structure for surgical instructions generation, while the input is the extracted spatial features from the surgical image and the output is its corresponding description. The whole architecture can be seen in Figure 5.3.

Encoder. Before the encoder, every feature vector is embedded by a linear embedding to reduce the dimension from 2048 to 512, and it is followed by a dropout layer. The embedded feature is the input of the first encoder layer. The whole encoder is a stack of 6 identical encoder layers, which generates an attention-based representation for the image. Each encoder layer consists of a **multi-head self-attention** layer and a position-wise **fully connected feed-forward network**. The detailed explanation of multi-head self-attention and fully connected feed-forward network is in section 3.2.2, section 3.2.3, section 4.2.2 and the original transformer paper (Vaswani et al. 2017).

Decoder. The input of the decoder is the information retrieved from the last encoder layer and the groundtruth caption. The decoder also consists of six identical layers, with each has two multi-head attention layers (decoder self-attention and encoder-decoder attention layer) and one fully connected feed-forward network. Every decoder self-attention layer is masked to prevent from attending to future locations.

5.2.3 Reinforcement Learning

Reinforcement learning (RL) is an area of machine learning techniques. The target of RL (as shown in Figure 5.4) is to maximize the cumulative reward by training an actor to interact with unknown environment. In this study, we focus on discussing one of the strategies called policy gradient and how it is applied in sequence generation. All the detailed formula derivation can be found in (Rennie et al. 2017). Policy gradient (Sutton et al. 2000) methods target at modelling and optimizing parameterized polices directly.

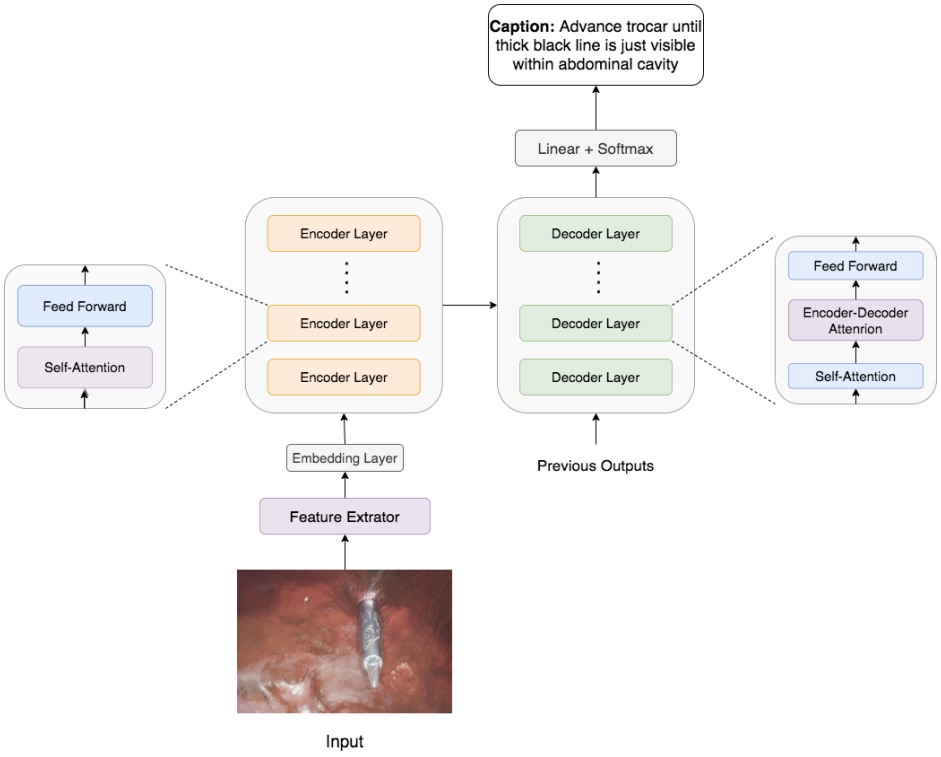


Figure 5.3: Overview of transformer based surgical instruction architecture.

Figure 5.4: An actor interacts with the environment and gets rewards

In sequence generation problem, the 'actor', language model (such as LSTM and transformer), interacts with the environment (image and word features)

to maximize the expected reward (CIDEr score):

$$L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[r(w^s)] \tag{5.4}$$

Here, maximize the expected reward equals to minimize negative expected reward. Where the policy p_θ is defined by the network parameters θ, and $w^s = (w_1^s, \dots w_T^s)$ and w_t^s is the word sampled at time step t. Based on the REINFORCE algorithm (Williams 1992), the corresponding gradient for $L(\theta)$ (a non-differentiable reward function) can be computed as:

$$\nabla_\theta L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[r(w^s) \nabla_\theta \log p_\theta(w^s)] \tag{5.5}$$

In order to eliminate the variance of the gradient, we add a baseline. The baseline can be any function as long as it does not depend on network parameters θ, and this can be represented as:

$$\nabla_\theta L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[(r(w^s) - b) \nabla_\theta \log p_\theta(w^s)] \tag{5.6}$$

According to the chain rule, the final gradient can be expressed as:

$$\nabla_\theta L(\theta) = \sum_{t=1}^{T} \frac{\partial L(\theta)}{\partial S_t} \frac{\partial S_t}{\partial \theta} \tag{5.7}$$

where s_t is a vector with dimension size same as the vocabulary size. It is the score of the word in time t, i.e. the input of the softmax layer. And according to (Ranzato et al. 2015):

$$\frac{\partial L(\theta)}{\partial S_t} \approx (r(w^s) - b)(p_\theta(w_t|h_t) - 1_{w_t^s}) \tag{5.8}$$

where h_t is the output hidden state from the language model and $1_{w_t^s}$ is the one-hot representation for word w_t^s at time t.
In our work, we choose the baseline as (Rennie et al. 2017), which is the reward $r(\hat{w})$ obtained by the current model under the inference algorithm used at test time. As a result, we increase the probability of high reward sample and penalty the low reward sample. The final equation can represent as:

$$\frac{\partial L(\theta)}{\partial S_t} \approx (r(w^s) - r(\hat{w}))(p_\theta(w_t|h_t) - 1_{w_t^s}) \tag{5.9}$$

5.3 Evaluation

5.3.1 Dataset Description

We evaluate our approach on DAISI dataset, which contains 17255 color images from 290 medical procedures, including external fetal monitoring, laparoscopic sleeve gastrectomy, laparoscopic ventral hernia repair, etc. Every procedure is consisted of few images with their corresponding descriptions. We further clean the dataset by deleting noisy and irrelevant images and descriptions such as the description of the author information. Finally, there are 16413 images (along with one caption each) in total, and we assign 13094 images for training, 1646 for validation, and 1673 for testing.

5.3.2 Text Preprocessing

Text preprocessing is an important step to transform the text into a more analyzable and predictable format for the deep learning model. Raw text instructions need to be preprocessed to learn meaningful features and not overfit on irrelevant noise. We follow these steps to clean the text instruction:

1. Converting all words to lower case
2. Expanding abbreviations, including medical abbreviations (e.g. 'a.' to 'artery') and English contractions (e.g. i've to 'i have')
3. Removing numbers, punctuation, and whitespace
4. Tokening the sentence into words

As other image captioning task, we set the threshold of the sentence length to 16, label any word count less than five as 'UNK', and build a vocabulary of size 2212 words. Figure 5.5 shows the distribution of the top 20 words. Except for the most common words in English ('to', 'of', 'the', 'and' etc.), there are some medical specific words in top 20 such as 'muscle', 'remove', 'fascia', etc.

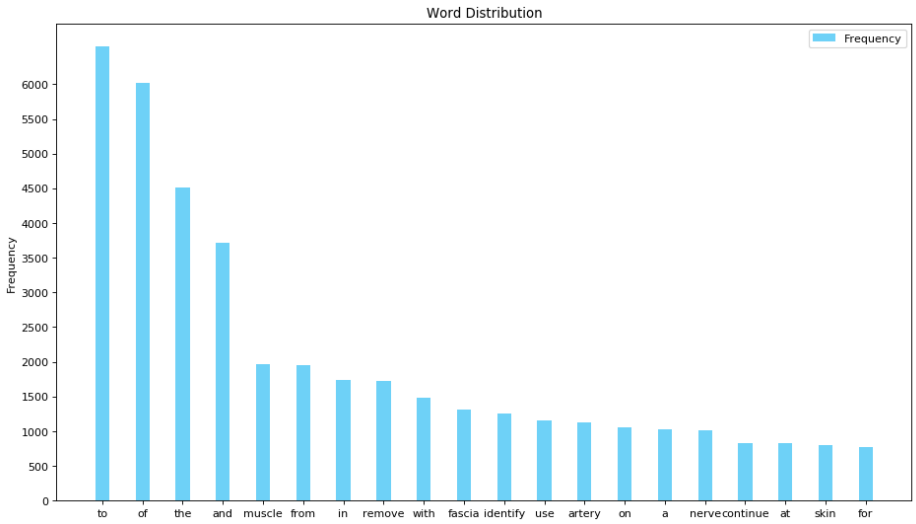


Figure 5.5: The distribution of top 20 words

5.3.3 Feature Extraction

LSTM model. We use Resnet-101 (101 layers) (He et al. 2016) pre-trained on ImageNet classification task to extract image feature. Images are encoded to 2048 dimension vectors with the final convolution layer of resnet follow by an average pooling layer.

Attention model. For the LSTM-based soft-attention model and the transformer model, we also apply the Resnet-101 to encode images. Instead of the average pooling, we apply a spatially adaptive max-pooling layer and end up with $14 \times 14 \times 2048$ dimension output.

5.3.4 Implementation Details

Following the standard procedure in image captioning task, we first train our model with a word-level cross-entropy (XE) loss then optimize the model using reinforcement learning. During the XE training process, the model is trained to predict the next word given previous ground-truth word, while the reinforcement learning process is trained to predict the next token based on the previous prediction.

LSTM-based models. For the LSTM model and LSTM-based soft-attention

model, the image and words embedding dimension and LSTM hidden state size are set to 512. For the XE training, we initialize the learning rate to 10^{-5} and follow the scheduled sampling strategy of (Bengio et al. 2015). During the self-critical evaluation, we use a fixed learning of 10^{-5} using CIDEr score as reward. Both models are optimized using ADAM optimizer (Kingma and Ba 2014) with the batch size of 10.

Transformer model. In regard to the transformer model, we set both encoder and decoder to 6 layers with the dimensionality to 512 for each layer. In addition, the head number for each layer is 8, the feed-forward dimension is 2048, and there is a dropout layer with probability of 0.9 after each attention and feed-forward layer. For the XE training, we initialize the learning rate to 5×10^{-4} and follow the learning rate scheduling strategy with 20000 warm-up steps. During the self-critical evaluation, we use a fixed learning of 5×10^{-5}. Both models are optimized using ADAM optimizer (Kingma and Ba 2014) with the batch size of 5.

5.3.5 Comparison with the State-of-the-Art

We adopt the standard evaluation metrics from image captioning in our experiments: **BLEU** (Papineni et al. 2002), **Rouge-L** (Lin 2004), **METEOR** (Banerjee and Lavie 2005), **CIDEr** (Vedantam et al. 2015), and **SPICE** Anderson et al. (2016). The detailed explanation of these evaluation metrics can be found in 2.2.4.

Since the code in Rojas-Muñoz et al. (2020) is not publicly available, we re-implement their Bi-RNN model. The 4096 dimensional image features are extracted using the last convolutional layer from a pre-trained VGG16 Simonyan and Zisserman (2014b). The Bi-RNN model is trained with 50 epochs by the initial learning rate at 5×10^{-4} and the batch size at 10. Table 5.1 compares our proposed models with the previous work (Rojas-Muñoz et al. 2020). It can be seen that the Bi-directional RNN (Rojas-Muñoz et al. 2020) has relatively lower performance, especially for the 3-gram and 4-gram BLEU score (11.3% and 9.3%) compared with ours (46.4% and 44.9%). In BLEU score evaluation, long $n - gram$ score measures the fluency of the

instruction. It can be concluded that Bi-directional RNN is not capable of generating adequate 'human-like' instructions. Among three other models, LSTM model achieves slightly better performance than LSTM-based soft-attention approach, and the transformer model with reinforcement learning outperforms all other methods in all evaluation metrics. The results indicate that the conventional RNN-based methods have limited ability of catching the dependencies between image features and text informations. While transformer encoder self-attention can encode the dependencies for image pixels, the decoder self-attention is able to model dependencies for textual information, and the encoder-decoder attention builds the relationship between image features and textual information. Figure 5.6 shows some visualization results using the proposed LSTM baseline and transformer framework. The results shows that our method can generate meaningful descriptions for surgical images.

Table 5.1: Comparison with the state-of-the-art (Rojas-Muñoz et al. 2020) for surgical instruction generation task. B1, B2, B3, B4, C, M, R and S stands for 1-4 gram BLEU, CIDEr, METEOR, ROUGE-L and SPICE score respectively.

Surgical Instruction	$B1$	$B2$	$B3$	$B4$	C	M	R	S
DAISI (BiRNN)	21.0	14.4	11.3	9.3	8.32	10.3	22.0	12.1
LSTM	43.7	39.4	37.3	36.2	34.0	24.9	44.6	40.2
LSTM + soft-attn	43.2	38.7	36.3	34.9	32.4	24.3	43.7	38.0
Transformer + rl	**52.8**	**48.7**	**46.4**	**44.9**	**42.7**	**30.7**	**53.1**	**48.4**

5.4 Discussion

5.4.1 The Influence of Reinforcement Learning

To further explore the functionality of each component, we decouple three networks and design an ablative experiment in six settings as follows:

(1) LSTM

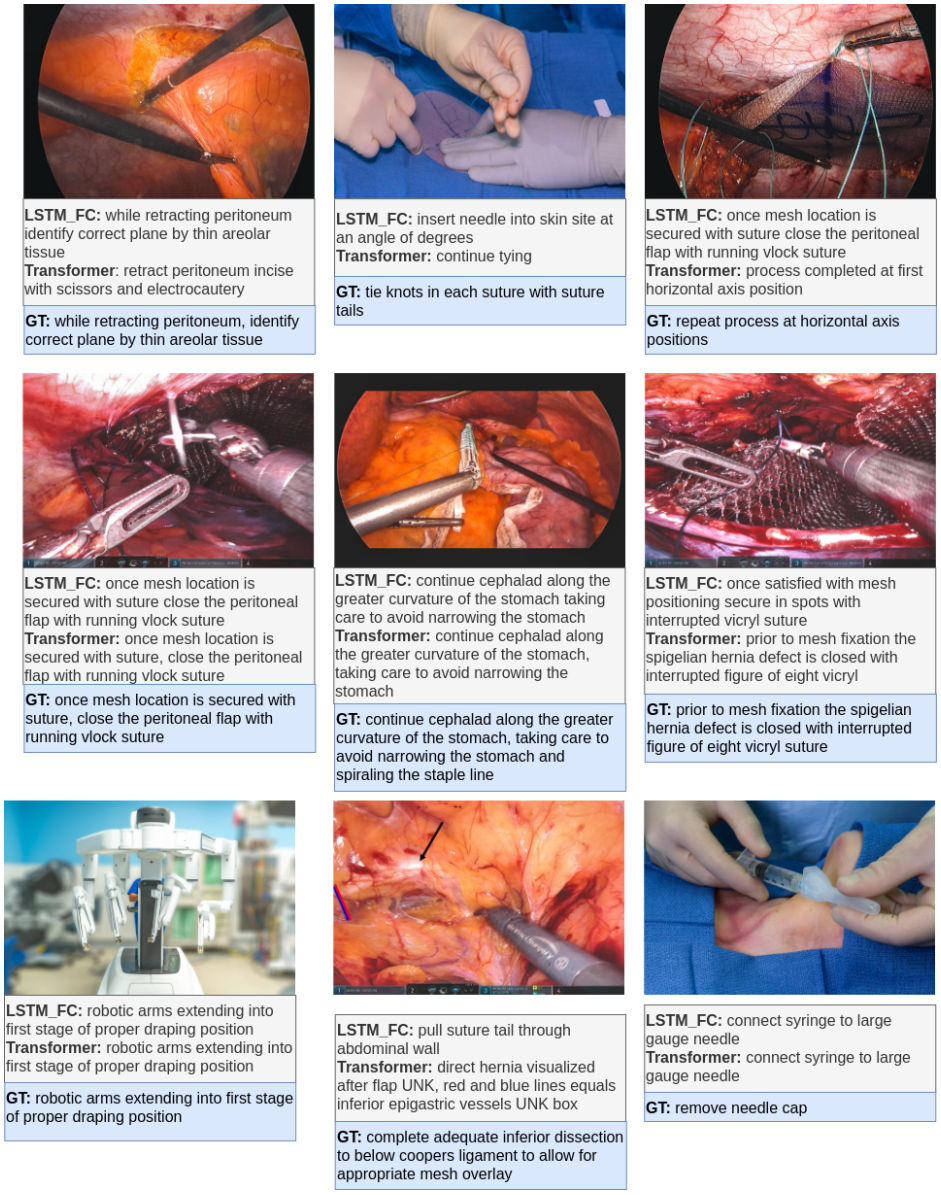


Figure 5.6: Some visualization results from transformer model

(2) LSTM + reinforcement learning

(3) LSTM + soft-attention

(4) LSTM + soft-attention + reinforcement learning

(5) Transformer

(6) Transformer + reinforcement learning

We apply these six different settings to generate instructions from surgical images and evaluate their 1-4 gram BLEU, CIDEr, METEOR, ROUGE-L and SPICE score separately. The experiment results are shown in Table 5.2.

(1) v.s. (3): We add the soft-attention attention module on the top of the LSTM to sequentially attend to different parts of image and aggregate information, but it performs slightly worse (around 1% for each evaluation standard) than the baseline model. This indicates that simple soft-attention mechanism cannot build the correlation between salient pixels and the next prediction word.

(1) v.s. (3) v.s. (5): Without using any recurrent neural units behind LSTM models, transformer model only use the self-attention mechanism to encode the spatial information and decode the text instruction. Transformer model achieves better performance than two LSTM models, which demonstrate its ability to multi-modal contexts.

(1) v.s. (2), (3) v.s. (4), and (5) v.s. (6): During the training procedure, we first train each model with standard XE loss, then we add the reinforcement learning block to optimize the CIDEr score directly. From the results, it can be seen that not only the CIDEr score, but also the performance of other evaluation metrics has been lifted. Specifically, we observe a significant increase in performance when using reinforcement training after the transformer model. The ablative experiment proves the functionality of each component of our method.

5.4.2 Limitations and Challenges

In this subsection, we want to discuss the current challenges and limitations for automatic surgical instruction based on the DAISI dataset. There are mainly four challenges and limitations:

1. **Small dataset size.** Although various single-modal deep learning tasks such as object detection, image segmentation, machine translation, sentiment analysis, etc. has achieved great performance, image

Table 5.2: Ablative study to explore the influence of reinforcement learning. B1, B2, B3, B4, C, M, R and S stands for 1-4 gram BLEU, CIDEr, METEOR, ROUGE-L and SPICE score respectively.

Surgical Instruction	$B1$	$B2$	$B3$	$B4$	C	M	R	S
LSTM	43.7	39.4	37.3	36.2	34.0	24.9	44.6	40.2
LSTM + rl	44.6	40.3	38.3	37.1	35.1	25.4	45.3	41.1
LSTM + attn	43.2	38.7	36.3	34.9	32.4	24.3	43.7	38.0
LSTM + attn + rl	43.4	38.8	36.4	34.8	33.1	24.8	44.1	38.5
Transformer	45.5	41	38.7	37.2	34	25.6	44.3	39.7
Transformer + rl	**52.8**	**48.7**	**46.4**	**44.9**	**42.7**	**30.7**	**53.1**	**48.4**

captioning is a multi-modal process, which use both computer vision and natural language processing techniques to generate text description from an image. Considering the complexity of the image captioning task, deep learning algorithms often require the huge amount of data to tune the parameters and prevent overfitting. For example, the COCO dataset (Lin et al. 2014) has more than 120,000 samples, and same types of objects appear many times. However excluding the noisy and irrelevant images, DAISI dataset only contains 16,413 images. And some of the surgical scenes only appear once. More importantly, understanding the surgical actions depends on the contextual activities and their description, but we can only achieve instruction generation based on single-shot due to the limited size of dataset.

2. **No pre-trained model.** Generally, the first step for image captioning is to extract spatial features pre-trained ImageNet classification task. More advanced algorithms then use Faster R-CNN algorithm (Ren et al. 2016) to detect the object bounding boxes and identify attribute features with Visual-Genome data (Anderson et al. 2018). Next they apply variants of attention mechanism over the extracted bounding boxes to have a better understanding between image representation and text information. We also use the pre-trained ImageNet classification model

to extract the medical image feature, because to our best knowledge, there is no public pre-trained model for medical images. In addition, higher level semantic information (object bounding boxes and attribute features) cannot be extracted from the medical images, for the reason that there are fundamental differences in image content, data size, and task specifications between medical and natural images. This will limit the ability of the model to build the dependencies between the predicted word and the most salient object it should pay attention to.

3. **Complicated descriptions and procedure.** In general area, the caption for an image is a simple sentence, which describe the objects, the attributes of the object (color, number), and their relationships (such as position relationship). As for the medical report generation, a medical report is consisted of few sentences, but they often follow the similar pattern. However, the instruction from DAISI are collected from different sources, including the medical app, textbooks, etc. Surgical instructions from the dataset are usually long and complex. During the pre-processing stage, we have cut every description into a single sentence and set the threshold to 16 words per sentence. But they have no pattern or template, and some instructions even contain the clauses.

4. **One caption per image.** In real situation, an image can be described in different ways. For instance, in order to build human-like model and evaluate the result objectively, Coco captioning task has equipped with 5 different reference translations for each image. Nonetheless, we have only one annotation for each image. It is possible that the evaluation metrics grade an adequate caption a low score only because it does not look similar to the ground truth label.

In summery, understanding surgical action and generating instruction is still at its early stage. Future works include collecting the large training dataset, building the specialized pre-trained model for medical images, regularizing and annotating more reference captions for surgical images.

5.5 Summary

In this chapter, we propose an encoder-decoder architecture fully backboned by transformer to predict surgical instructions from various medical disciplines. The experiment results demonstrate that the transformer architecture is capable of creating the pixel-wise patterns from self-attention encoder, developing text relationships for masked self-attention decoder, and devising the image-text dependencies from encoder-decoder attention. In order to solve the mismatching between the training and testing procedure, we optimize the model with self-critical reinforcement learning, which takes the CIDEr score as the reward after the general cross-entropy training.
Understanding surgical activity and predicting instruction prediction is still at its early stage. Future works include collecting the large training dataset, building the specialized pre-trained model for medical images, regularizing and annotating more reference captions for surgical images.

Chapter 6

A Simulation Platform towards Context-aware Surgical Assistance

Our final aim is to integrate context-aware assistance into whole MIS process by taking advantages of deep learning techniques to automatically learn highly discriminative features from surgical videos. In previous chapters, we have designed an attention network to automatically assess the surgical skills (see Chapter 3), a symmetric dilated model to identify surgical actions (see Chapter 4), and a transformer based approach to generate surgical instructions(see Chapter 5). To achieve our final goal, in this chapter, we design and implement an Unity-based laparoscopic cholecystectomy VR simulator as a starting point.

On the one hand, surgical simulator can integrate pre-operative surgical training, intra-operative activities recognition and guidance generation and post-operative comprehensive surgical skill assessment into a whole ecosystem (see Figure 6.1). On the other hand, after the integration of vision-based skill assessment, surgical gesture recognition and the surgical instruction, more data can be collected without any private and ethics issues. And the data can be applied to train and improve the existing model.

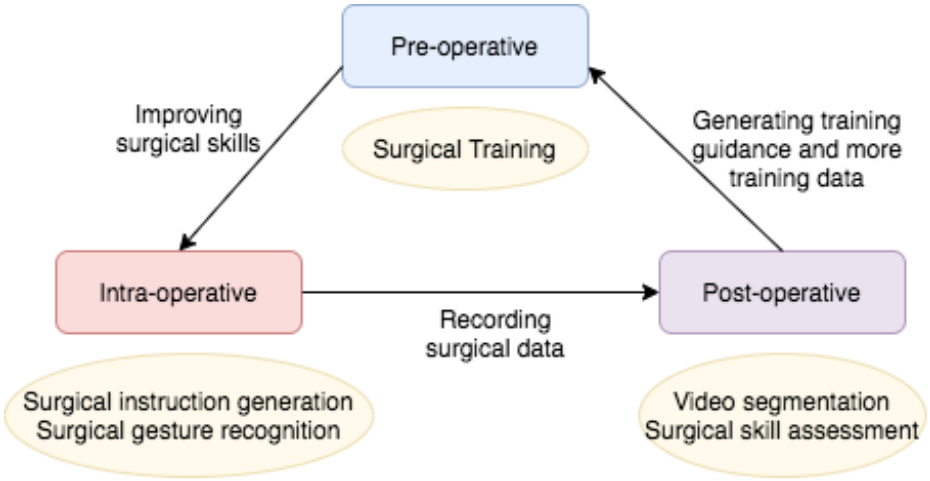


Figure 6.1: VR surgical simulator is an efficient solution to improve the eye-hand coordinate and dexterity skills prior to the surgery. During the surgery, on the one hand, the context-aware system can provide surgical instructions if any inappropriate surgical gesture is detected. On the other hand, surgical video is recorded for post-operative surgical skills analysis and assessment. Furthermore, this process will generate more training data to improve current deep learning model.

6.1 Introduction

Cholecystectomy, a surgical procedure to remove the gallbladder, is one of the most common operating room surgeries. With the explosion of related minimally invasive surgery (MIS) techniques, cholecystectomy can be operated with the assistance of a video camera and several laparoscopic instruments. Benefits of laparoscopic cholecystectomy include few post-operative complications, small incisions on the skin, and a relative shorter recovery period (Suzuki et al. 2000). However, the limited viewing angle and restricted operating space during laparoscopic cholecystectomy often cause undesirable complications (such as bile duct leak or injury). Therefore, it is indispensable for surgeons to acquire pertinent surgical skills prior to a real surgery.

Traditionally, using cadavers or animals for training faces many ethical issues. Training by supervised experienced surgeons is a feasible solution, but it is at the cost of the huge expenditure and a long training period. Fortunately, virtual reality (VR) simulator has gradually become an effective

approach (Alaker et al. 2016) to teach and assess surgical skills outside operation rooms. Medical students and surgeons can practice their professional skills along with the transferable skills such as team-working and emergency reaction using this advanced technique. Many research works have proven the correlation between game-based training and the improvement of surgical skills (Rosser et al. 2007, Knight et al. 2010, Andreatta et al. 2010, Kurenov et al. 2009, Creutzfeldt et al. 2010). Specifically, advantages of the surgical simulator include:

1. offering a secure and efficient environment for surgeons to deeply understand the whole surgery process

2. enabling surgeons to practice two essential skills for MIS, namely eye-hand coordination and the ability to execute 3D actions using a 2D screen as a guide

3. allowing trainees to practice the challenging and significant procedures repeatedly and save the training cost

It is complicated to model and simulate a typical laporoscopic cholecystectomy scene. The simulation of multiple organs and tissues, the interaction between tools and tissues, and feedbacks from both visual and tactile are all necessary components. Developing and integrating all these functional components independently is time and resources consuming. Unity (Goldstone 2009) is a cross-platform modern game engine which has a great support of advanced audio and visual effects. Its integrated development environment, easy profiler, on-shelf tools and modules make Unity stable and highly productive.

In this chapter, we design and develop a laparoscopic cholecystectomy simulator using Unity Game Engine. Our surgical simulator provides trainees an effective platform to practice surgery procedures with realistic visual and haptic feedbacks (shown in Fig. 6.6 and Fig. 6.3). In our design, the organ and soft tissue simulation are based on uflex, a unity plugin originated from NVidia Flex(Korzeniowski 2016), and the haptic feedback is in reference to

Unity 5 Haptic Plugin for Geomagic OpenHaptics 3.3 (Poyade et al. (2014)-). There are mainly two contributions of our work:

1. Developing an interactive laparoscopic cholecystectomy simulator for surgeons to enhance their surgical skills

2. Exploring the use of game engine for novel development of a surgical simulator to achieve efficient and cost-effective solutions

Figure 6.2: Laparoscopic cholecystectomy simulator working environment

6.2 Related Work

Game like medical education and surgical skills training with different objectives have been proven to be valuable where users can improve their related

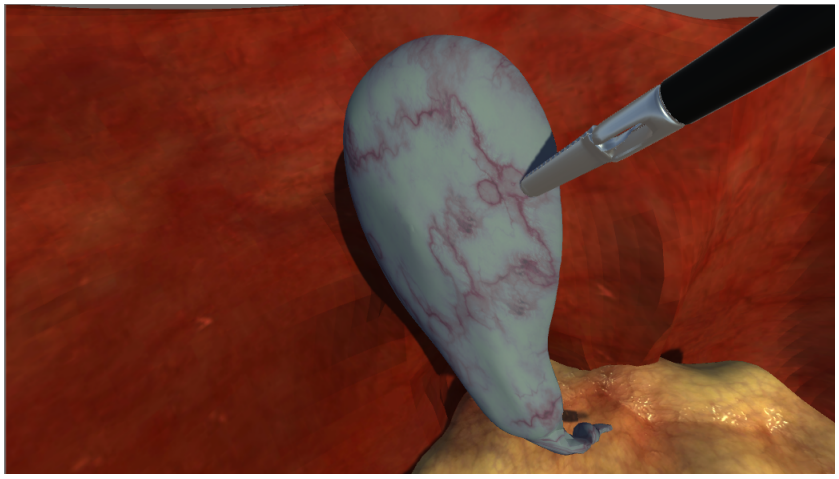


Figure 6.3: Screenshot from cholecystectomy simulator

skills and experience various circumstances that are impractical in the real world due to the safety, cost, or other reasons (Susi et al. 2007). James et al. presented that surgeons who had experienced selected representative video games (Super Monkey Ball 2, Star Wars Racer Revenge, and Silent Scope) performed better in laparoscopic handling and made fewer errors (Rosser et al. 2007). Another interesting experiment (Knight et al. 2010) compared the tagging accuracy between participants who had undertaken the game, Triage Trainer (Blitz Games Studios, Leamington Spa, Uk) training and who had taken traditional card-sort exercise. The Triage Trainer group was more likely to triage precisely in casualties than the card-sort group. Game like training was also used for team training in emergency crisis management (Andreatta et al. 2010, Kurenov et al. 2009), daily clinical tasks for junior doctors, and cardiopulmonary resuscitation (Creutzfeldt et al. 2010), etc. In this research, we focus on the laparoscopic surgical skills trained via the surgical simulator.

Simulators for surgical skills training can be categorized into two types: low fidelity and high fidelity. Evaluation criteria of fidelity are decided by the extent of visual and tactile realism and system interactivity. Synthetic models (Hammoud et al. 2008) and Video Box Trainer belong to low-fidelity simulators. Synthetic models are static bench models such as tissue models

for fascia closure and knot-tying training. Video Box Trainer (Fried et al. 2004) takes advantage of real laparoscopic instruments as well as the camera and video monitors to simulate surgery procedures. Although low fidelity simulators are cost-efficient, easy to make and portable, they usually sacrifice the realism and only offer a single surgical skill training rather than a whole process.

Virtual reality surgery training is a type of high fidelity surgical simulation. This type of technology satisfies the requirements of taking 2D screen as a guide to performing 3D surgery with realistic visual and tactile feedback while also has the ability to monitor and record the training progress. Most of the existing surgical simulators either reinvent all components independently (Qian et al. 2015, Pan et al. 2015) or build from an extendable framework such as SOFA (Kim et al. 2015) and GiPSi (Cavusoglu et al. 2006). Whereas building medical educational or training systems upon game engines is not fully explored. Few research works (Marks et al. 2007b a) have discussed the possibilities and advantages of game engine based surgery simulators. In order to bring more understanding in this direction, we use a robust multipurpose game engine Unity to build our simulation architecture. We also evaluate and compare the pros and cons between the game engine and other simulation methods for surgical training.

Immersive surgery simulation requires accurate physical behaviors, the precision of soft tissue deformation enormously affects the sense of reality of the whole framework (Gallagher et al. 2005, Zhang et al. 2017a). Three types of simulation approaches have been widely applied nowadays.

- Mass spring-based: Mass-spring (Baraff and Witkin 1998, Bouaziz et al. 2014, Liu et al. 2013) system consists of sets of point masses connected by spring dampers. It is a simple and efficient scheme which takes Hooke's Law as the theoretical basis. But it is hard to tune the spring constants to get a desired behavior and usually causes the overshooting problem.

- Finite element-based: Unlike the mass-spring system which discretize an object into finite number of point masses, finite element method

(FEM) (Zienkiewicz and Taylor 2005) is based on continuum mechanics theory. It is capable of handling accurate physical behaviors for different types of elastic and non-elastic material. However, the model complexity associates the technique with difficult initial settings and high computational cost (Sifakis and Barbic 2012).

- Position Based Dynamics (PBD) (Müller et al. 2007): PBD is a method which works on positions directly in each simulation step to resolve constraints. It is fast, stable, and controllable which makes the simulation process highly efficient and functionally suitable for the interactive environment (Bender et al. 2014). Despite the fact that PBD is not physically accurate, it achieves real time surgical simulation which is still visually plausible. We applied the PBD solver with different constraints for our physics simulation as a test prototype.

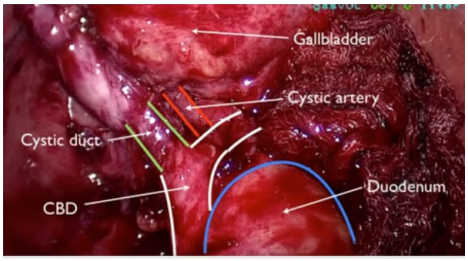
6.3 System Infrastructure

6.3.1 Procedures and Challenges in Laparoscopic Cholecystectomy

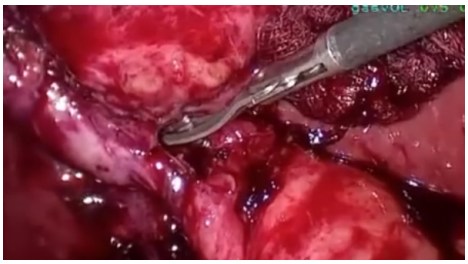
Laparoscopic Cholecystectomy has gradually replaced the open surgery to become the major modality of treating gallstones, gallbladder carcinoma, trauma and porcelain gallbladder. Four primary steps are involved in the surgery as shown in Fig. 6.4: a) Position the patient correctly and insert the trocars and camera through three labeled minimal ports; b) Identify the position of the gallbladder, and then dissect the tissue to find the cystic duct and cystic artery; c) Clip and ligate the cystic duct and cystic artery structure; and d) Tease away adhesions between the gallbladder and the liver, put the gallbladder into a pre-prepared bag and drag it out. In our training system, we focused on the core steps b), c), and d), as a) can be conducted in a separate training session.

During the intra-operative process, some challenges and difficulties require a particular attention, especially for a novice:

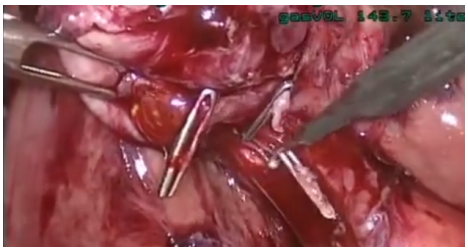
(a) Initial view



(b) Dissect Calot's Triangle



(c) Clipping and cut



(d) Teasing away the gallbladder

Figure 6.4: Surgery process. Image by School of Surgery (Jones 2014)

- Identifying the Calot's triangle properly is the prerequisite to safely remove the gallbladder. Calot's triangle, shown in Fig. 6.5, is the most important anatomy structure in laparoscopic cholecystectomy. It helps the surgeon clearly identify the relative position among the cystic duct, common hepatic duct, and inferior border of the liver.
- Always clipping the cystic duct and cystic artery first and then cutting them to prevent uncontrollable bleeding or bile leaking.
- Gallbladder closely connects with the liver. Correctly separating adhesions between them prevents the harm to the liver.
- In laparoscopic surgery, a surgeon operates 36-39cm long instruments through a tiny entry point, such that the unskilled and careless manipulation is possible to damage intestines or main blood vessels.

6.3.2 Objectives and System Design

This laparoscopic cholecystectomy simulator is designed and developed for medical students and junior doctors to train their surgical skills in a safe,

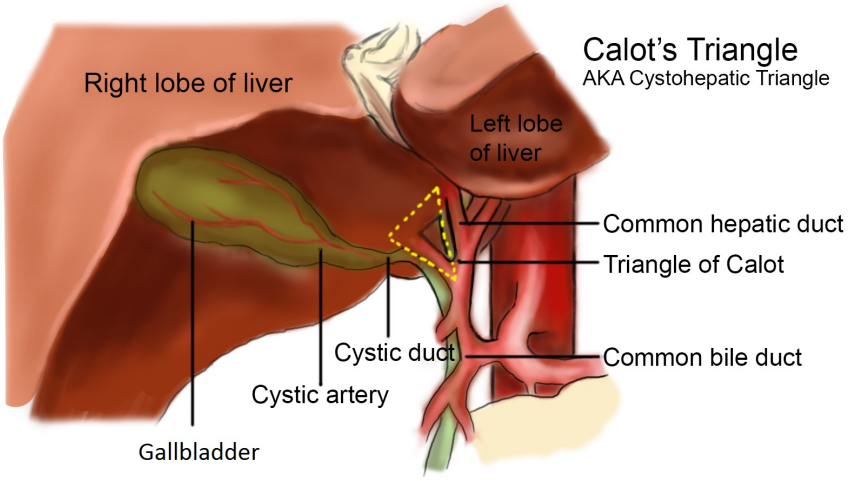


Figure 6.5: Calot's Triangle. (Suzuki et al. 2000)

repeatable and cost-effective manner. Medical students can thoroughly understand the surgery process and practice essential skills before a hand-on operation on a real patient. As for junior doctors lacking practical experiences, they are able to draw lessons from their weakness and practice repeatedly to further enhance operation skills. Laparoscopic experts are involved to monitor the progress of students and juniors to give valuable feedbacks and guidances during the training.

Typically, a surgery training simulator includes a rendering module for displaying 3D models and user interface, a physics simulation module for collision detection, rigid body, soft tissue, and smog/fluids simulation, and an event handling module for input/output. In our development, to exclude unnecessary complexity, we only prototype the anatomical structure for a normal patient without lesions. The patient specific mode will be the future work, which would be of better use to benefit surgical planning. Concretely, we design the game module as below (Fig. 6.6 shows four crucial steps) :

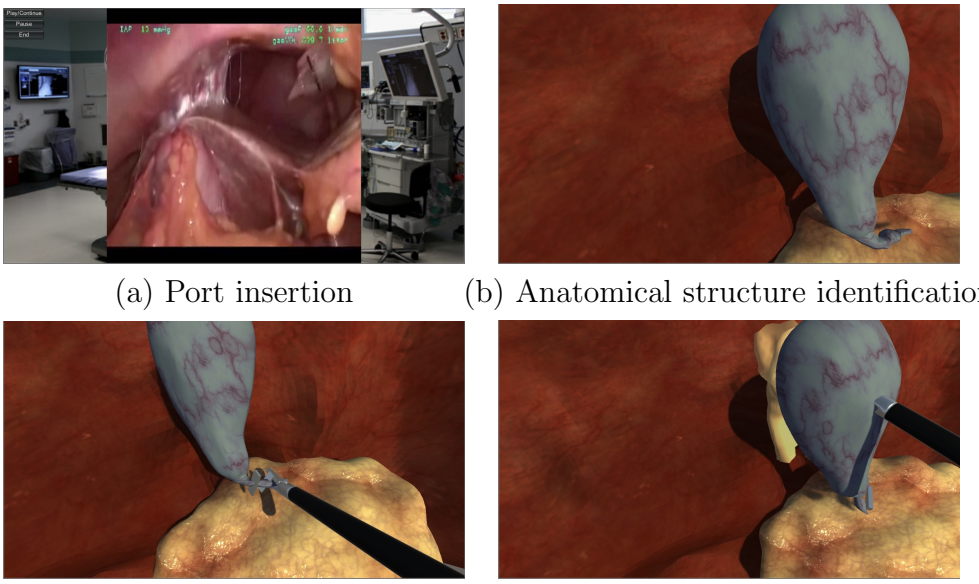
1. A video (Jones 2014) illustrates the process of labeling the patient and inserting trocars through abdomen.

2. Viewpoint selection is important on account of the limited vision field of minimally invasive surgery. Users are able to adjust the camera to get the optimal viewpoint by keyboard.

3. After identifying correlated anatomical structures, trainees need to clip three staples in correct positions on the cystic duct and ligate the gallbladder.

4. With the help of two haptic devices, trainees are required to separate the adhesion between the gallbladder and the liver.

5. The remaining task is rather simple, so we use another video clip to present the procedure of putting the gallbladder into a pre-prepared bag and dragging it out.



(a) Port insertion (b) Anatomical structure identification

(c) Clipping and ligating the cystic duct (d) Teasing away the gallbladder

Figure 6.6: Operation procedures.

6.4 Physical Simulation

NVIDIA Flex is a particle-based simulation library for real time visual effects. In Flex, everything is modeled as a system of particles connected by

different constraints, which is broadly based on PBD and Unified Particle Physics (Macklin et al. 2014). NVIDIA FleX does not support the Unity Game Engine directly. Instead uFlex, a Unity asset integrated low-level Flex native library, is applied in our development. The core idea is to solve a non-linear system sequentially with equality and inequality constraints in different time steps:

$$C_i(\mathbf{p} + \Delta\mathbf{p}) = 0, \quad i = 1, ...n \tag{6.1}$$

$$C_j(\mathbf{p} + \Delta\mathbf{p}) >= 0, \quad j = 1, ...n \tag{6.2}$$

where $\mathbf{p} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_N]^T$ is the vector of particle positions. The unified representation allows modeling different materials and interactions between various types of models, namely rigid body, soft body, cloth, and fluids in a fully efficient and flexible manner. There are mainly two types of the soft tissue associated with laparoscopic cholecystectomy: volumetric soft tissue and surface deformable tissue. We explain their modeling methods as follows.

6.4.1 Volumetric Soft Body Simulation

Liver, gallbladder and fat tissues are regarded as 3D volumetric deformable objects in cholecystectomy, since they keep their intrinsic shape most of the time. When instruments contact or manipulate the organ, only the contacted part deformed. Shape matching constraints of clusters are used to simulate soft bodies. The core idea of shape matching (Müller et al. 2005) can be described as: Attracting initial point sets $\mathbf{x}_i^0$ towards the "goal" positions $\mathbf{p}_i$ by finding the global optimal transformation and rotation matrix against the current state. That is to minimize:

$$\sum_i w_i(\mathbf{R}(\mathbf{x}_i^0 - \mathbf{t}_0) + \mathbf{t} - \mathbf{p}_i)^2 \tag{6.3}$$

where weights $\{\mathbf{w}_i | i = 1, ...n\}$ in this case are mass, $\mathbf{R}$ is the rotation matrix, and $\mathbf{t}$ is the translation vector. The optimal translation vectors turn out to be the displacement of the center of mass for the shape, and the global

rotation matrix is extracted from the global transformation matrix by polar decomposition. Finally, the goal potion can be calculated as:

$$\mathbf{p}_i = \mathbf{R}(\mathbf{x}_i^0 - \mathbf{x}_{cm}^0) + \mathbf{x}_{cm} \tag{6.4}$$

where $\mathbf{x}_{cm}$ is the center of mass.

Throughout the simulation, a soft body consists of several clusters is shown in Fig. 6.7. Each particle belongs to one or more clusters with a weight between 0 and 1, standing for the extent of a particle influenced by the corresponding cluster. The final deformation result is the overlapping and averaging of the constraints which defines in each cluster independently as described by the shape matching model. The more clusters in a soft body, the more elastic it will become. Shape matching constraint does not need the connectivity information, and the result is easy to compute, thus it is efficient for an interactive environment which compromises some physical accuracy.

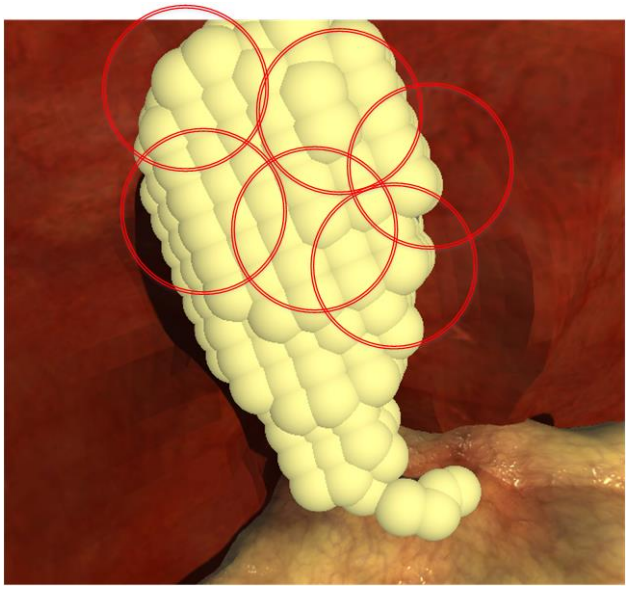


Figure 6.7: Clustering in shape matching

6.4.2 Surface Mesh Simulation

We need to simulate the fascia tissue between the gallbladder and the liver in the dissection procedure. Given the extensible and thin features of the

fascia, it is modeled using cloth simulation. The fascia is displayed as spring networks, in which one triangle handles the stretch and two adjacent triangles handle the bending constraint. For each edge, the stretching constraint function is:

$$C_{stretch}(\mathbf{p}_1, \mathbf{p}_2) = |\mathbf{p}_1 - \mathbf{p}_2| - l_0 \tag{6.5}$$

where l_0 is the rest length of the edge. As for two neighbor triangles ($\mathbf{p}_1$, $\mathbf{p}_3$, $\mathbf{p}_2$) and ($\mathbf{p}_1$, $\mathbf{p}_2$, $\mathbf{p}_4$), the bending constraint is generated as:

$$C_{bend}(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4) = arccos(\mathbf{n}_1 \cdot \mathbf{n}_2) - \varphi_0 \tag{6.6}$$

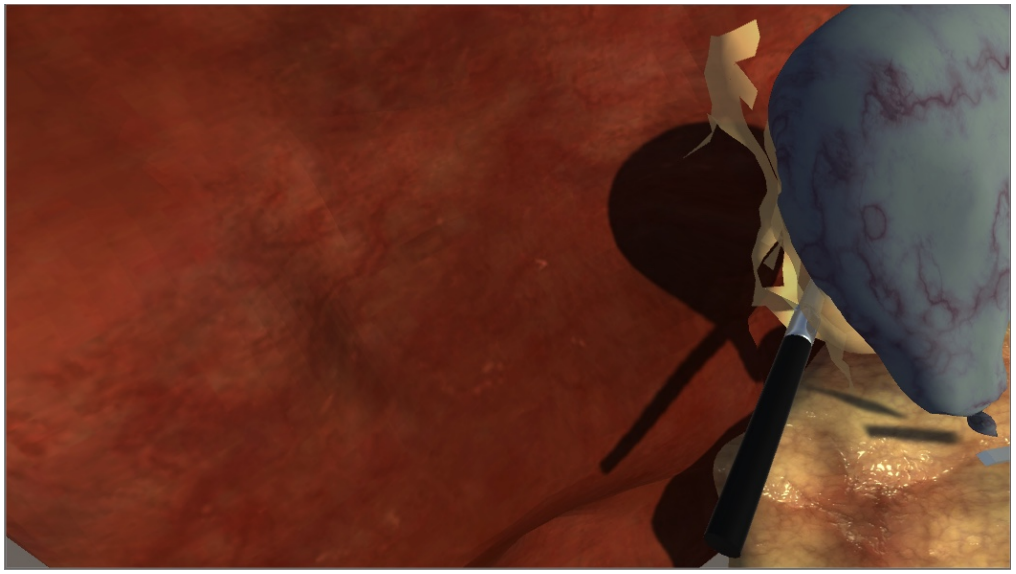
where $\mathbf{n}_1$ and $\mathbf{n}_2$ are the normal vector of two triangles and φ_0 is the dihedral angle between two triangles in the rest pose. The overall goal is to minimize total energy defined as:

$$E = \sum (k_{stretch} C^2_{stretch} + k_{bend} C^2_{bend}) \tag{6.7}$$

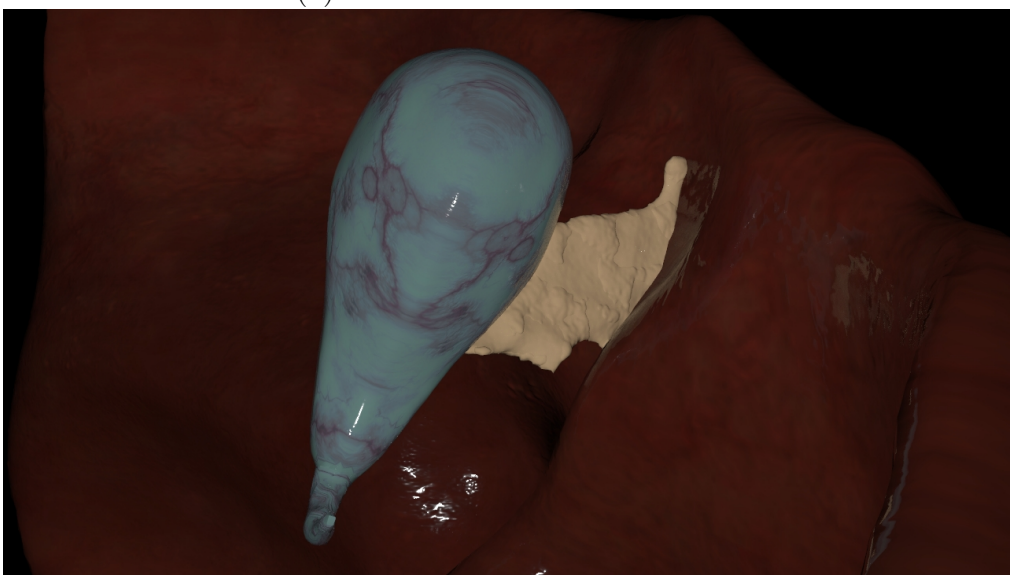
where $k_{stretch}$ and k_{bend} are global stiffness parameters provided by the user.

However, the fascia simulation is not visually plausible in our test, because fascia is not a single layer of smooth structure as the cloth. While splitting the gallbladder from the liver in our simulation, the fascia just looks like a piece of cloth with explicit triangle mesh in the torn partition as shown in Fig. 6.8(a). We design an adhesive constraint to simulate the breakable fascia to enhance the visual effects. Because uFlex can only access to the basic-level library of NVIDIA Flex, it cannot simulate the sticky effects. We tested the algorithm in Houdini, and we will integrate the adhesive constraint into system in the future work. As the result, the breakable fascia is visually satisfying as shown in Fig. 6.8(b).

In cloth simulation, fascia tissue is represented as connective triangle mesh. While in our new algorithm, we simply use lines and the distance constraint to model the fascia. Initially, we randomly generate several points on the surface of the gallbladder and the liver and connect them with lines. In each connection, few points are randomly scattered and marked as breakable group. At the same time, we assign a rest length and a very large stiffness parameter to each line. Finally when we drag the gallbladder, if the distance



(a) Artifact in fascia simulation



(b) Adhesive constraint for the fascia

Figure 6.8: Comparison between original and improved results

between two organs exceeds a given break threshold, the stiffness parameter of random breakable points will be reset to a very small number such that any force can break the connections. In such way, the connection between the gallbladder and the liver no longer looks like a piece of papery cloth with

apparent triangle mesh. Instead, it presents a sticky behavior of the fascia when we drag the gallbladder away.

6.4.3 Haptic Rendering

A system requires at least 500-1000 HZ for smooth haptic rendering (Booth et al. 2003), in which 25-30 Hz for visual rendering in order to provide users realistic feedbacks. Haptic feedback is able to enhance both cognitive ability by offering continuous movement sensory and kinesthetic sense by differentiating different inner structures. Phantom Omni is a feasible and affordable haptic device for virtual-real interaction. It takes the 6 DOF input and generates forces to constrain the 3 DOF output in a high sensory frequency. Haptic rendering in Unity enables the user to control haptic effectors as laparoscopic instruments to interact with organs and tissues.

Figure 6.9 shows the interactive work flow of the haptic device. At the beginning, we predefine the haptic workspace dimensions(Poyade et al. (2014)-). Within the workspace, meshes and transformation matrices of haptic geometries (organs and tissues) are set into haptic frame. During the training process, the position of the haptic proxy (i.e. the graphic representation of the haptic device in the screen, in our case is the position and orientation of the laparoscopic instrument model) keeps updating. If no collisions happen between instruments and the organ, the proxy position and the device position will stay the same. But if the proxy get contact with the haptic objects, the proxy position will be assigned to the device position. According to the position and the collided particle ID information, the simulation algorithm calculates the deformation and gives visual feedbacks to the screen and force feedbacks to the user hand.

6.5 Evaluation and Feedback

A PC with an Intel Xeon 5 CPU, a GeForceGTX 1080 graphics card, and two Phantom Omni devices is used to built this laparoscopic cholecystectomy simulator. The criterion-based quantitative assessment (Jackson et al.

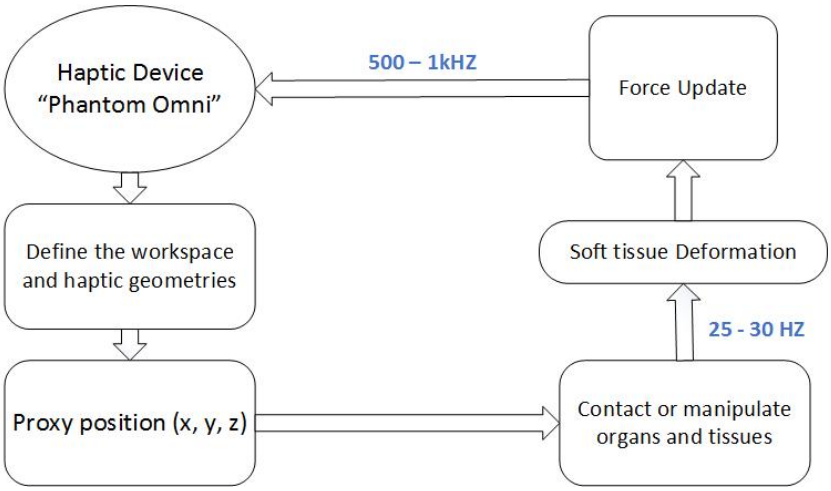


Figure 6.9: Haptic device workflow

2011) is employed to evaluate the overall capacity of the simulator from system performance and usability perspectives. Ten volunteer PhD students and five volunteers from general surgery department (including students and laparoscopic experts) participated in the test and gave their suggestions accordingly.

6.5.1 System Evaluation

There are five criteria chosen to evaluate the system performance: 1) ease of use, 2) interactivity, 3) visual realism, 4) freedom of movement and effectiveness, and 5) system stability. Ten volunteer PhD students are asked to grade each criterion from 1 to 5, representing "Poor" to "Excellent". Figure 6.10 presents the average result of each criterion about the performance of our simulator. The system interactivity, freedom of movement and effectiveness and, ease of use receive the positive feedback with the score around four. All participants mentioned that haptic device with force feedback hugely improve the user experience. However, the visual realism and system stability need to be improved, due to the fact that the related FLEX solver is not physically accurate.

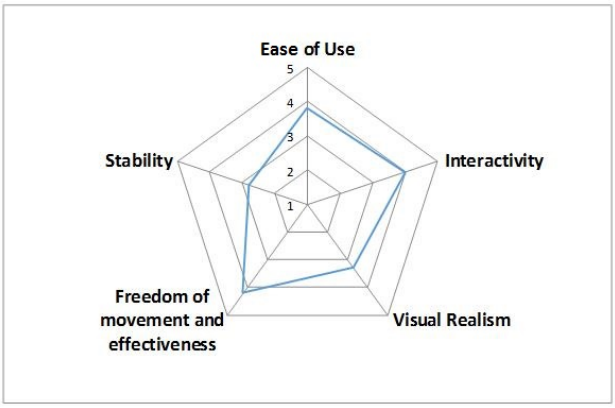


Figure 6.10: System Performance Evaluation *(1-Poor, 2-Fair, 3-Average, 4-Good, 5-Excellent)*

6.5.2 Simulator Usability Evaluation

Five criteria, namely 1) ease of use, 2) anatomy precision, 3) visual realism, 4) integrity of procedure, and 5) utility, are applied to assess the usability of the simulator. Evaluators from general surgery department are asked to grade each criterion from one to five, representing "Poor" to "Excellent". Figure 6.11 presents the average score of each criterion about the usability of our simulator. Due to the small sample numbers, such evaluation provided some guidance and further validation will be carried in the future work. The utility and ease of use reach the highest score around four, thus demonstrating the usefulness of VR surgical simulation. Nevertheless, there is still a space for improvement in anatomy precision, visual realism and integrity of procedure. Surgeons and medical students claim that as there are no professionals involved in our simulator design, some minimal details are neglected. For example, the inner structure of calot's triangle is more complex than the one we built. This defective anatomical representation has the potential to mislead trainers. In the future improvement, we will improve them with the guide from medical experts.

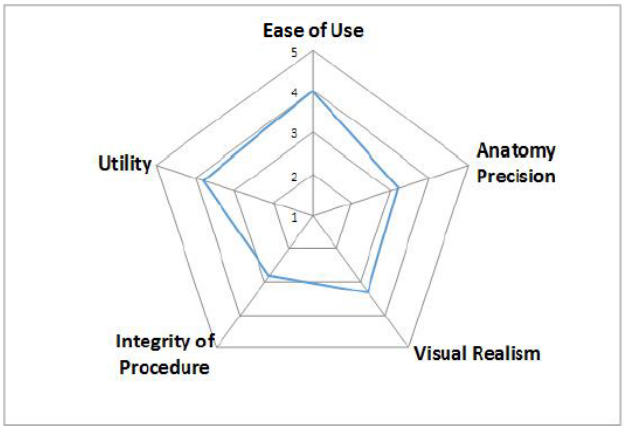


Figure 6.11: Simulator usability evaluation. *(1-Poor, 2-Fair, 3-Average, 4-Good, 5-Excellent)*

6.5.3 Game Engine based Simulator

Most of the existing simulators are either developed independently or based on some open-source framework, whereas we built the whole architecture upon Unity Game Engine. We have discussed our experimental results in a focus group of eight attendances from both computer scientists and laparoscopical surgeons. Comparing with other two types of simulators, we summarized advantages as well as limitations of using Game Engine for the surgical training.

Advantages

- For the self-developed simulators, reinventing essential functional blocks (rendering, physic and event handling components) for a simulator is time-consuming and leads to huge development expenditure. In our case, we spent about four months for the whole development cycle, but it took one year or more for a simulator like (Qian et al. 2015) to achieve the similar function.
- Robust game engine architectures allow developers to concentrate more on the content rather than the implementation, thus could improve the user experience.

- Most of the simulator systems only focus on the technical training of the surgery process. However, other skills such as team collaborative and patient care are also very important for a qualified surgeon. The self-development simulators are not able to network with other simulators for collaborative tasks. None of them are audio supportive (Marks et al. 2007b). Game engine has better support for network and audio.

Limitations

- Game engines are primarily designed for game development. The physical simulation capability of virtual surgery still has a room for improvement. Unlike the ordinary soft body, the structure of human organs and soft tissues are very complex. The physics engine in Unity is not able to provide high fidelity laparoscopic features, e.g. cutting and multi-layer heterogeneous soft tissue simulation.

- It is a trade-off for a game engine to either sacrifice the precise animation for speed or emphasize on pleasant visual without real time interactive ability.

6.5.4 Improvement Suggestions

According to the evaluation results and suggestions from open-ended questionnaires, further improvements need to be achieved:

1. Although the game engine has integrated a system development architecture for audio, rendering, event handling and networking, the visual realism of soft tissue deformation and the precision of collision detection needs to be improved. We will first integrate adhesive constraint simulation for fascia into our simulator.

2. Several participants from medical institution indicate that we need more precise anatomy structures for calot's triangle. The success of laparoscopic cholecystectomy largely relies on whether surgeon could identify the anatomy structure properly.

3. We will also add more interactive modules to improve the integrity and efficiency of the simulator. Currently, only one type of laparoscopic instrument is modelled in the simulator. Several instruments (retractor, electrodes, and anvil graspers) will be integrated, and users will be allowed to change the instruments for different tasks.

6.6 Summary

In this chapter, we have designed and developed a novel game engine based surgical simulator for laparoscopic cholecystectomy. Surgeons are able to train their surgical and decision making skills in this virtual environment before they encountering real patients. The VR surgical training allows trainees to operate the immersive cholecystectomy by using haptic devices with force feedbacks and a computer screen as the guide. The training environment is safe, cost-efficient, and repeatable. The user evaluation results demonstrate that the simulator is easy to use and interactive.

Currently, we only receive feedbacks from few laparoscopic surgeons and medical students. Before it can be validated as a surgery training tool, we need to further improve the evaluation strategy. Controlled experiments will be taken by dividing surgeons into two groups with one trained by the simulator and one without training to determine the benefits and drawbacks of our simulator. Further improving works are desired, including modeling the anatomic structures of Calot's triangle accurately, solving technical challenges related to realistic soft tissue simulation and integrating other training tasks for a better user experience.

Chapter 7

Conclusion and Future Work

In this chapter, we conclude the thesis by summarizing our works. We also discuss the possible future work directions to resolve current limitations.

7.1 Conclusion

This thesis targets at analyzing and understanding surgical activities and building an integrated context-aware system in different surgical activity granularities to assist minimally invasive surgery. Specifically, we focus on vision-based solutions, because surgical video and image data are relatively easy to access. And visual data has high-dimensional semantic features. To achieve this goal, we provide solutions from the coarse to fine level and design a simulation system towards integrated context-aware system.

On the coarsest level, we propose a novel multi-task framework for automatic surgical skill assessment with six detailed evaluation standards by modified objective structured assessments of technical skills (OSATS). The spatial-temporal features are first extracted from a 3D residual network (3D ResNet). Then we present a self-attention based architecture to capture the spatial-temporal features and establish frame-to-frame relationship to automatically find the critical frames for surgical skill determination. We evaluate our approach for three fundamental surgical tasks (suturing, needle passing and knot-tying) and achieve nearly 100% accuracy. As for the detailed surgical technique skills evaluation, six elements are assessed in a multi-task manner

and achieve the performance ranging from 56% to 91%. As the result, our proposed method opens the possibility of automatic surgical assessment and comprehensive performance report generation. The details of this work has been discussed in Chapter 3.

On the fine-grained level, we present a temporal convolutional framework to jointly segment and detect the fine-grained surgical gestures from a RGB surgical video. Rather than using video clips, we take the whole video into consideration and apply the extracted spatial-temporal feature as the input of the the network. The network is composed of a symmetric dilated encoder-decoder to enlarge the receptive field and catch the long term temporal information, an self-attention module in the middle to build the frame-wise adjacent as well as the global relationships, and the pooling and upsampling layer to alleviate the over-segmentation error. Our method has been evaluated on the suturing task from the JIGSAWs dataset over five cross-validation runs. Our results largely outperform the state-of-the-art methods on the frame-wise accuracy up to $\sim$6 points and the F1@50 score $\sim$6 points. The details of this work has been discussed in Chapter 4.

Taking one step further from the video classification and gesture recognition, we aim to recognize surgical actions and generate instructions from still images. Firstly, we apply the pre-trained ResNet-101 from ImageNet classification task to extract the image features. Then we have the long short term memory (LSTM) and LSTM-based attention model as the baseline to predict the text description. Furthermore, the transformer-based encoder-decoder approach is proposed. In order to solve the mismatch between the training ans testing process, we optimize the models with reinforcement learning algorithm, which takes the CIDEr score as the reward after the general cross-entropy training. We validate the task using evaluation metrics from image captioning models for DAISI dataset. Among all the methods, transformer-based method achieves the best performance over all evaluation metrics. It demonstrates that transformer architecture is capable of creating the pixel-wise patterns from self-attention encoder, developing text relationships for masked self-attention decoder, and building the image-text dependencies from encoder-decoder attention. The details of this work has been

discussed in Chapter 5.
In order to integrate the surgical activities analysis in different granularity into a context-aware system. We design and implement an Unity-based laparoscopic cholecystectomy simulator as a starting point. For one, this system is a carrier for building the context-aware system. For another, the simulation system can provide novice surgeon a secure and repeated environment to improve surgical skills pre-operatively. Rather than implement everything from scratch, game engine is a cost-effective, flexible and highly interactive solution, which provide the user plausible physics simulation and realistic haptic feedback. The details of this work has been discussed in Chapter 6.

7.2 Limitations and Future Work

Integrated context-aware system. As we discussed in the thesis, our final goal is to provide surgeons context-aware assistance through whole clinical pathway. On the basis of the laparoscopic cholecystectomy developed in Chapter 6, we will merge the surgical gesture recognition, surgical instruction generation and surgical skill assessment into the system. On the one hand, this integrated system can provide the pre-operative surgical training, intra-operative activities recognition and guidance generation and post-operative comprehensive surgical skill assessment. Surgical report can be created to help surgeons understanding their skill limitations and the way of improvement. On the other hand, we can collect huge amount of data from the whole process without any privacy or ethical issues and the data can be trained to optimize our current algorithms to improve the usability, plausibility and effectiveness of our current algorithms.
Training data and pre-trained medical model. The performance of the deep learning algorithms hugely depend on the amount of training data. Previous research (Twinanda et al. 2016a) has proven the correlation between the network performance and the amount of training data. Currently, for the surgical recognition and skill assessment task, JIGSAWs is the only public dataset, which only have 39, 36, and 28 video for suturing, knot-tying,

and needle passing, respectively. Comparing with the hundreds or thousands training data in other research field, the task with limited data size is harder to train and easier to face overfitting problems. As for the surgical guidance generation tasks, thousands of images are also far from enough. In our future work, more surgical data need to be collected and annotated.

In addition, we extracted the high level features from surgical data using pre-trained model for natural images. The pre-trained model is trained on a large scale benchmark dataset, which contains a wide diversities of object categories (e.g. ImageNet). It helps the network to extract general features which can be reused on the target task. However, natural images have essential differences in image content, amount of data and task specifications with medical images. In (Raghu et al. 2019), the authors surprisingly find that the performance of medical imaging tasks has not significantly been improve by transfer learning. In order to achieve the better performance and get interpretable model, the pre-trained model based on surgical data need to be built specifically.

Other visual data source. In our current surgical skill assessment and surgical gesture recognition tasks, we only use visual information captured by 3D-camera from da Vinci surgical robot. This kind of camera is able to record the whole surgical procedure, which shows the anatomical structure inside human body and tool-tissue interactions. Besides the activities happened inside the patient, the general activities occurring in the operation room (OR) such as the manipulation of the surgeon and the cooperation between the clinical teams, are also important for surgical activities analysis. (Twinanda et al. 2016b) use a multi-view RGBD camera system mounted on the celling of the OR to record the color images and 3D OR-scene structure by the depth camera. They then use the OR-scene videos to recognize the surgical phases of laparoscopic surgery. In our future work, in order to acquire a complete information regarding surgical activities from different perspective, we will simultaneously use visual information from the surgical scene and OR-scene considering their complementary nature for surgical activities analysis.

Semi-supervised or unsupervised learning algorithms. Currently, our approaches work in a fully supervised manner, which means surgical data

need to be densely annotated, and the annotation result is subjective. In order to reduce the cost and human labor for data annotation, it would be interesting to explore the semi-supervised or unsupervised learning algorithms such that only small part of data need to be annotated.

On-line learning. In present settings, the surgical gesture recognition method works off-line in acausal mode. It means all the previous and afterwards information are available for the current prediction. However, during intra-operative process, only the current frame and previous frames are available. We will develop the real-time surgical gesture recognition algorithm in our future work.

References

Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B. B., Zappella, L., Khudanpur, S., Vidal, R. and Hager, G. D., 2017. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 64 (9), 2025–2041.

Alaker, M., Wynn, G. R. and Arulampalam, T., 2016. Virtual reality training in laparoscopic surgery: a systematic review & meta-analysis. *International Journal of Surgery*, 29, 85–94.

van Amsterdam, B., Clarkson, M. J. and Stoyanov, D., 2020. Multi-task recurrent neural network for surgical gesture recognition and progress prediction. *arXiv preprint arXiv:2003.04772*.

Anderson, P., Fernando, B., Johnson, M. and Gould, S., 2016. Spice: Semantic propositional image caption evaluation. *European Conference on Computer Vision*, Springer, 382–398.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S. and Zhang, L., 2018. Bottom-up and top-down attention for image captioning and visual question answering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Andreatta, P. B., Maslowski, E., Petty, S., Shim, W., Marsh, M., Hall, T., Stern, S. and Frankel, J., 2010. Virtual reality triage training provides a viable solution for disaster-preparedness. *Academic emergency medicine*, 17 (8), 870–876.

Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Banerjee, S. and Lavie, A., 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Baraff, D. and Witkin, A., 1998. Large steps in cloth simulation. *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, ACM, 43–54.

Bardram, J. E., Doryab, A., Jensen, R. M., Lange, P. M., Nielsen, K. L. and Petersen, S. T., 2011. Phase recognition during surgical procedures using embedded and body-worn sensors. *2011 IEEE international conference on pervasive computing and communications (PerCom)*, IEEE, 45–53.

Bardram, J. E. and Nørskov, N., 2008. A context-aware patient safety system for the operating room. *Proceedings of the 10th international conference on Ubiquitous computing*, 272–281.

Bay, H., Tuytelaars, T. and Van Gool, L., 2006. Surf: Speeded up robust features. *European conference on computer vision*, Springer, 404–417.

Bender, J., Koschier, D., Charrier, P. and Weber, D., 2014. Position-based simulation of continuous materials. *Computers & Graphics*, 44, 1–10.

Bengio, S., Vinyals, O., Jaitly, N. and Shazeer, N., 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in Neural Information Processing Systems*, 1171–1179.

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A. and Torr, P. H., 2016. Fully-convolutional siamese networks for object tracking. *European conference on computer vision*, Springer, 850–865.

Bettadapura, V., Schindler, G., Plötz, T. and Essa, I., 2013. Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2619–2626.

Bilgic, E., Turkdogan, S., Watanabe, Y., Madani, A., Landry, T., Lavigne, D., Feldman, L. S. and Vassiliou, M. C., 2017. Effectiveness of telementoring in surgery compared with on-site mentoring: a systematic review. *Surgical innovation*, 24 (4), 379–385.

Blum, T., Feußner, H. and Navab, N., 2010. Modeling and segmentation of surgical workflow from laparoscopic video. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 400–407.

Bodner, J., Wykypiel, H., Wetscher, G. and Schmid, T., 2004. First experiences with the da vinci operating robot in thoracic surgery. *European Journal of Cardio-thoracic surgery*, 25 (5), 844–851.

Booth, S., Angelis, F. and Schmidt-Tjarksen, T., 2003. The influence of changing haptic refresh-rate on subjective user experiences-lessons for effective touch-based applications. *Proceedings of eurohaptics*, Citeseer, 374–383.

Bouarfa, L., Jonker, P. P. and Dankelman, J., 2011. Discovery of high-level tasks in the operating room. *Journal of biomedical informatics*, 44 (3), 455–462.

Bouaziz, S., Martin, S., Liu, T., Kavan, L. and Pauly, M., 2014. Projective dynamics: fusing constraint projections for fast simulation. *ACM Transactions on Graphics (TOG)*, 33 (4), 154.

Bustos, A., Pertusa, A., Salinas, J.-M. and de la Iglesia-Vayá, M., 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66, 101797.

Carreira, J. and Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset. *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Cavusoglu, M. C., Goktekin, T. G. and Tendick, F., 2006. Gipsi: a framework for open source/open architecture software development for organ-level surgical simulation. *IEEE Transactions on Information Technology in Biomedicine*, 10 (2), 312–322.

Challacombe, B., Kavoussi, L., Patriciu, A., Stoianovici, D. and Dasgupta, P., 2006. Technology insight: telementoring and telesurgery in urology. *Nature Clinical Practice Urology*, 3 (11), 611–617.

Chen, Z., Song, Y., Chang, T.-H. and Wan, X., 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.

Cheng, J., Dong, L. and Lapata, M., 2016. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*.

Cornia, M., Stefanini, M., Baraldi, L. and Cucchiara, R., 2020. Meshed-memory transformer for image captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10578–10587.

Creutzfeldt, J., Hedman, L., Medin, C., Heinrichs, W. L. and Felländer-Tsai, L., 2010. Exploring virtual worlds for scenario-based repeated team training of cardiopulmonary resuscitation in medical students. *Journal of medical Internet research*, 12 (3).

Dai, B., Fidler, S., Urtasun, R. and Lin, D., 2017. Towards diverse and natural image descriptions via a conditional gan. *Proceedings of the IEEE International Conference on Computer Vision*, 2970–2979.

Datta, V., Mackay, S., Mandalia, M. and Darzi, A., 2001. The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. *Journal of the American College of Surgeons*, 193 (5), 479–485.

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R. and McDonald, C. J., 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23 (2), 304–310.

Deng, Z., Jiang, Z., Lan, R., Huang, W. and Luo, X., 2020. Image captioning using densenet network and adaptive attention. *Signal Processing: Image Communication*, 115836.

Ding, L. and Xu, C., 2017. Tricornet: A hybrid temporal convolutional and recurrent network for video action segmentation. *arXiv preprint arXiv:1705.07818*.

Ding, L. and Xu, C., 2018. Weakly-supervised action segmentation with iterative soft boundary assignment. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6508–6516.

DiPietro, R., Lea, C., Malpani, A., Ahmidi, N., Vedula, S. S., Lee, G. I., Lee, M. R. and Hager, G. D., 2016. Recognizing surgical activities with recurrent neural networks. *International conference on medical image computing and computer-assisted intervention*, Springer, 551–558.

Dollár, P., Rabaud, V., Cottrell, G. and Belongie, S., 2005. Behavior recognition via sparse spatio-temporal features. *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, IEEE, 65–72.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K. and Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2625–2634.

Doughty, H., Mayol-Cuevas, W. and Damen, D., 2019. The pros and cons: Rank-aware temporal attention for skill determination in long videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7862–7871.

Erridge, S., Yeung, D. K., Patel, H. R. and Purkayastha, S., 2019. Telementoring of surgeons: a systematic review. *Surgical innovation*, 26 (1), 95–111.

Fard, M. J., Ameri, S., Chinnam, R. B., Pandya, A. K., Klein, M. D. and Ellis, R. D., 2016. Machine learning approach for skill evaluation in robotic-assisted surgery. *arXiv preprint arXiv:1611.05136*.

Farha, Y. A. and Gall, J., 2019. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3575–3584.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L. and Muller, P.-A., 2018. Evaluating surgical skills from kinematic data using convolutional neural networks. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 214–221.

Feichtenhofer, C., Fan, H., Malik, J. and He, K., 2019. Slowfast networks for video recognition. *Proceedings of the IEEE international conference on computer vision*, 6202–6211.

Franke, S., Meixensberger, J. and Neumuth, T., 2013. Intervention time prediction from surgical low-level tasks. *Journal of biomedical informatics*, 46 (1), 152–159.

Fried, G. M., Feldman, L. S., Vassiliou, M. C., Fraser, S. A., Stanbridge, D., Ghitulescu, G. and Andrew, C. G., 2004. Proving the value of simulation in laparoscopic surgery. *Annals of surgery*, 240 (3), 518.

Funke, I., Bodenstedt, S., Oehme, F., von Bechtolsheim, F., Weitz, J. and Speidel, S., 2019a. Using 3d convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 467–475.

Funke, I., Mees, S. T., Weitz, J. and Speidel, S., 2019b. Video-based surgical skill assessment using 3d convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14 (7), 1217–1225.

Gallagher, A. G., Ritter, E. M., Champion, H., Higgins, G., Fried, M. P., Moses, G., Smith, C. D. and Satava, R. M., 2005. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Annals of surgery*, 241 (2), 364–372.

Gao, J., Wang, S., Wang, S., Ma, S. and Gao, W., 2019. Self-critical n-step training for image captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6300–6308.

Gao, Y., Vedula, S. S., Reiley, C. E., Ahmidi, N., Varadarajan, B., Lin, H. C., Tao, L., Zappella, L., Béjar, B., Yuh, D. D. et al., 2014. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. *MICCAI Workshop: M2CAI*, volume 3, 3.

Goldstone, W., 2009. *Unity game development essentials*. Packt Publishing Ltd.

Hammoud, M. M., Nuthalapaty, F. S., Goepfert, A. R., Casey, P. M., Emmons, S., Espey, E. L., Kaczmarczyk, J. M., Katz, N. T., Neutens, J. J. and Peskin, E. G., 2008. To the point: medical education review of the role of simulators in surgical training. *American Journal of Obstetrics & Gynecology*, 199 (4), 338–343.

Haouchine, N., Cotin, S., Peterlik, I., Dequidt, J., Lopez, M. S., Kerrien, E. and Berger, M.-O., 2014. Impact of soft tissue heterogeneity on augmented reality for liver surgery. *IEEE transactions on visualization and computer graphics*, 21 (5), 584–597.

Hara, K., Kataoka, H. and Satoh, Y., 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 6546–6555.

He, K., Girshick, R. and Dollár, P., 2019a. Rethinking imagenet pre-training. *Proceedings of the IEEE international conference on computer vision*, 4918–4927.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, X., Yang, Y., Shi, B. and Bai, X., 2019b. Vd-san: Visual-densely semantic attention network for image caption generation. *Neurocomputing*, 328, 48–55.

Herdade, S., Kappeler, A., Boakye, K. and Soares, J., 2019. Image captioning: Transforming objects into words. *Advances in Neural Information Processing Systems*, 11137–11147.

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural computation*, 9 (8), 1735–1780.

Hu, H., Gu, J., Zhang, Z., Dai, J. and Wei, Y., 2018. Relation networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3588–3597.

Hu, W., Xie, N., Li, L., Zeng, X. and Maybank, S., 2011. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41 (6), 797–819.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K. Q., 2017. Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Jackson, M., Crouch, S. and Baxter, R., 2011. Software evaluation: criteria-based assessment. *Software Sustainability Institute*.

Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.-W. and Heng, P.-A., 2018. Sv-rcnet: Workflow recognition from surgical videos using recurrent convolutional network. *IEEE transactions on medical imaging*, 37 (5), 1114–1126.

Jing, B., Xie, P. and Xing, E., 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.

Jones, K., 2014. Laparoscopic cholecystectomy explained. URL "https://www.youtube.com/watch?v=a7rIFlvZM0I&t=139s".

Judkins, T. N., Oleynikov, D. and Stergiou, N., 2009. Objective evaluation of expert and novice performance during robotic surgical training tasks. *Surgical endoscopy*, 23 (3), 590.

Karaman, S., Seidenari, L. and Del Bimbo, A., 2014. Fast saliency based pooling of fisher encoded dense trajectories. *ECCV THUMOS Workshop*, volume 1, 5.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1725–1732.

Kataoka, H., Wakamiya, T., Hara, K. and Satoh, Y., 2020. Would mega-scale datasets further enhance spatiotemporal 3d cnns? *arXiv preprint arXiv:2004.04968*.

Kerr, B. and PATRICK, J., 1999. The training of the surgeon: Dr. halsted's greatest legacy. *small*, 96, 62.

Kim, Y., Kim, L., Lee, D., Shin, S., Cho, H., Roy, F. and Park, S., 2015. Deformable mesh simulation for virtual laparoscopic cholecystectomy training. *The Visual Computer*, 31 (4), 485–495.

Kingma, D. P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knight, J. F., Carley, S., Tregunna, B., Jarvis, S., Smithies, R., de Freitas, S., Dunwell, I. and Mackway-Jones, K., 2010. Serious gaming technology in major incident triage training: a pragmatic controlled trial. *Resuscitation*, 81 (9), 1175–1179.

Korzeniowski, P., 2016. uflex integrates nvidia flex with unity3d.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A. et al., 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123 (1), 32–73.

Krizhevsky, A., Sutskever, I. and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097–1105.

Kuehne, H., Gall, J. and Serre, T., 2016. An end-to-end generative framework for video segmentation and recognition. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 1–8.

Kurenov, S. N., Cance, W. W., Noel, B. and Mozingo, D. W., 2009. Game-based mass casualty burn training. *Studies in health technology and informatics*, 142, 142–144.

Lalys, F. and Jannin, P., 2014. Surgical process modelling: a review. *International journal of computer assisted radiology and surgery*, 9 (3), 495–511.

Laptev, I., 2005. On space-time interest points. *International journal of computer vision*, 64 (2-3), 107–123.

Lea, C., Choi, J. H., Reiter, A. and Hager, G. D., 2016. Surgical phase recognition: from instrumented ors to hospitals around the world. *Medical image computing and computer-assisted intervention M2CAIMICCAI workshop*, 45–54.

Lea, C., Flynn, M. D., Vidal, R., Reiter, A. and Hager, G. D., 2017. Temporal convolutional networks for action segmentation and detection. *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 156–165.

Lea, C., Hager, G. D. and Vidal, R., 2015. An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks. *2015 IEEE winter conference on applications of computer vision*, IEEE, 1123–1129.

Lei, P. and Todorovic, S., 2018. Temporal deformable residual networks for action segmentation in videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6742–6751.

Li, Y., Liang, X., Hu, Z. and Xing, E. P., 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 1530–1540.

Lin, C.-Y., 2004. Rouge: A package for automatic evaluation of summaries. *Text summarization branches out*, 74–81.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L., 2014. Microsoft coco: Common objects in context. *European conference on computer vision*, Springer, 740–755.

Lipton, Z. C., Berkowitz, J. and Elkan, C., 2015. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B. and Sánchez, C. I., 2017. A survey on deep learning in medical image analysis. *Medical image analysis*, 42, 60–88.

Liu, D. and Jiang, T., 2018. Deep reinforcement learning for surgical gesture segmentation and classification. *International conference on medical image computing and computer-assisted intervention*, Springer, 247–255.

Liu, S., Zhu, Z., Ye, N., Guadarrama, S. and Murphy, K., 2017. Improved image captioning via policy gradient optimization of spider. *Proceedings of the IEEE international conference on computer vision*, 873–881.

Liu, T., Bargteil, A. W., O'Brien, J. F. and Kavan, L., 2013. Fast simulation of mass-spring systems. *ACM Transactions on Graphics (TOG)*, 32 (6), 214.

Lu, J., Xiong, C., Parikh, D. and Socher, R., 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 375–383.

Lu, J., Yang, J., Batra, D. and Parikh, D., 2018. Neural baby talk. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7219–7228.

Macklin, M., Müller, M., Chentanez, N. and Kim, T.-Y., 2014. Unified particle physics for real-time applications. *ACM Transactions on Graphics (TOG)*, 33 (4), 153.

Maier-Hein, L., Vedula, S. S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S. et al., 2017. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1 (9), 691–696.

Marks, S., Windsor, J. and Wünsche, B., 2007a. Collaborative soft object manipulation for game engine-based virtual reality surgery simulators.

Marks, S., Windsor, J. and Wünsche, B., 2007b. Evaluation of game engines for simulated surgical training. *Proceedings of the 5th international conference on Computer graphics and interactive techniques in Australia and Southeast Asia*, ACM, 273–280.

Martin, J., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C. and Brown, M., 1997. Objective structured assessment of technical skill (osats) for surgical residents. *British journal of surgery*, 84 (2), 273–278.

Mavroudi, E., Bhaskara, D., Sefati, S., Ali, H. and Vidal, R., 2018. End-to-end fine-grained action segmentation and recognition using conditional random field models and discriminative sparse coding. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 1558–1567.

Müller, M., Heidelberger, B., Hennix, M. and Ratcliff, J., 2007. Position based dynamics. *Journal of Visual Communication and Image Representation*, 18 (2), 109–118.

Müller, M., Heidelberger, B., Teschner, M. and Gross, M., 2005. Meshless deformations based on shape matching. *ACM transactions on graphics (TOG)*, ACM, volume 24, 471–478.

Nakawala, H., Ferrigno, G. and De Momi, E., 2017. Toward a knowledge-driven context-aware system for surgical assistance. *Journal of Medical Robotics Research*, 2 (03), 1740007.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K., 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Padoy, N., Blum, T., Ahmadi, S.-A., Feussner, H., Berger, M.-O. and Navab, N., 2012. Statistical modeling and recognition of surgical workflow. *Medical image analysis*, 16 (3), 632–641.

Pan, J. J., Chang, J., Yang, X., Liang, H., Zhang, J. J., Qureshi, T., Howell, R. and Hickish, T., 2015. Virtual reality training and assessment in laparoscopic rectum surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 11 (2), 194–209.

Pan, Y., Yao, T., Li, Y. and Mei, T., 2020. X-linear attention networks for image captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10971–10980.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Parlak, S. and Marsic, I., 2013. Detecting object motion using passive rfid: A trauma resuscitation case study. *IEEE Transactions on Instrumentation and Measurement*, 62 (9), 2430–2437.

Parlak, S., Marsic, I. and Burd, R. S., 2011. Activity recognition for emergency care using rfid. *BODYNETS*, 40–46.

Parmar, P. and Tran Morris, B., 2017. Learning to score olympic events. *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 20–28.

Pascanu, R., Mikolov, T. and Bengio, Y., 2013. On the difficulty of training recurrent neural networks. *International conference on machine learning*, 1310–1318.

Pavlopoulos, J., Kougia, V. and Androutsopoulos, I., 2019. A survey on biomedical image captioning. *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, 26–36.

Peters, J. H., Fried, G. M., Swanstrom, L. L., Soper, N. J., Sillin, L. F., Schirmer, B., Hoffman, K., Committee, S. F. et al., 2004. Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. *Surgery*, 135 (1), 21–27.

Poyade, M., Kargas, M. and Portela, V., (2014)-. Haptic plug-in for unity.

Qian, K., Bai, J., Yang, X., Pan, J. and Zhang, J., 2015. Virtual reality based laparoscopic surgery simulation. *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology*, ACM, 69–78.

Raab, S., 1998. Computer-aided surgery apparatus. US Patent 5,748,767.

Raghu, M., Zhang, C., Kleinberg, J. and Bengio, S., 2019. Transfusion: Understanding transfer learning for medical imaging. *arXiv preprint arXiv:1902.07208*.

Ranzato, M., Chopra, S., Auli, M. and Zaremba, W., 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Reiley, C. E. and Hager, G. D., 2009. Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. *International conference on medical image computing and computer-assisted intervention*, Springer, 435–442.

Ren, S., He, K., Girshick, R. and Sun, J., 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39 (6), 1137–1149.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J. and Goel, V., 2017. Self-critical sequence training for image captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7008–7024.

Richard, A., Kuehne, H. and Gall, J., 2017. Weakly supervised action learning with rnn based fine-to-coarse modeling. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 754–763.

Rohrbach, M., Amin, S., Andriluka, M. and Schiele, B., 2012. A database for fine grained activity detection of cooking activities. *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 1194–1201.

Rojas-Muñoz, E., Couperus, K. and Wachs, J., 2020. Daisi: Database for ai surgical instruction. *arXiv preprint arXiv:2004.02809*.

Ronneberger, O., Fischer, P. and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Rosser, J. C., Lynch, P. J., Cuddihy, L., Gentile, D. A., Klonsky, J. and Merrell, R., 2007. The impact of video games on training surgeons in the 21st century. *Archives of surgery*, 142 (2), 181–186.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al., 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115 (3), 211–252.

Schlegl, T., Waldstein, S. M., Vogl, W.-D., Schmidt-Erfurth, U. and Langs, G., 2015. Predicting semantic descriptions from medical images with convolutional neural networks. *International Conference on Information Processing in Medical Imaging*, Springer, 437–448.

Sefati, S., Cowan, N. J. and Vidal, R., 2015. Learning shared, discriminative dictionaries for surgical gesture segmentation and classification. *MICCAI Workshop: M2CAI*, volume 4.

Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J. and Summers, R. M., 2016. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2497–2506.

Sifakis, E. and Barbic, J., 2012. Fem simulation of 3d deformable solids: a practitioner's guide to theory, discretization and model reduction. *ACM SIGGRAPH 2012 Courses*, ACM, 20.

Simonyan, K. and Zisserman, A., 2014a. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 568–576.

Simonyan, K. and Zisserman, A., 2014b. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, B., Marks, T. K., Jones, M., Tuzel, O. and Shao, M., 2016. A multi-stream bi-directional recurrent neural network for fine-grained action detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1961–1970.

Staniūtė, R. and Šešok, D., 2019. A systematic literature review on image captioning. *Applied Sciences*, 9 (10), 2024.

Stauder, R., Okur, A., Peter, L., Schneider, A., Kranzfelder, M., Feussner, H. and Navab, N., 2014. Random forests for phase detection in surgical workflow analysis. *International Conference on Information Processing in Computer-Assisted Interventions*, Springer, 148–157.

Susi, T., Johannesson, M. and Backlund, P., 2007. Serious games: An overview.

Sutton, R. S., McAllester, D. A., Singh, S. P. and Mansour, Y., 2000. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 1057–1063.

Suzuki, M., Akaishi, S., Rikiyama, T., Naitoh, T., Rahman, M. and Matsuno, S., 2000. Laparoscopic cholecystectomy, calot's triangle, and variations in cystic arterial supply. *Surgical endoscopy*, 14 (2), 141–144.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Tao, L., Elhamifar, E., Khudanpur, S., Hager, G. D. and Vidal, R., 2012. Sparse hidden markov models for surgical gesture classification and skill evaluation. *International conference on information processing in computer-assisted interventions*, Springer, 167–177.

Tao, L., Zappella, L., Hager, G. D. and Vidal, R., 2013. Surgical gesture segmentation and recognition. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 339–346.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M. and Padoy, N., 2016a. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36 (1), 86–97.

Twinanda, A. P., Winata, P., Gangi, A., Mathelin, M. and Padoy, N., 2016b. Multi-stream deep architecture for surgical phase recognition on multi-view rgbd videos. *Proc. M2CAI Workshop MICCAI*, 1–8.

Varadarajan, B., Reiley, C., Lin, H., Khudanpur, S. and Hager, G., 2009. Data-derived models for segmentation with application to surgical assessment and training. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 426–434.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 5998–6008.

Vedantam, R., Lawrence Zitnick, C. and Parikh, D., 2015. Cider: Consensus-based image description evaluation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Vinyals, O., Toshev, A., Bengio, S. and Erhan, D., 2015. Show and tell: A neural image caption generator. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Wang, H., Kläser, A., Schmid, C. and Liu, C.-L., 2011. Action recognition by dense trajectories. *CVPR 2011*, IEEE, 3169–3176.

Wang, H. and Schmid, C., 2013. Action recognition with improved trajectories. *Proceedings of the IEEE international conference on computer vision*, 3551–3558.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X. and Van Gool, L., 2016. Temporal segment networks: Towards good practices for deep action recognition. *European conference on computer vision*, Springer, 20–36.

Wang, T., Huang, J., Zhang, H. and Sun, Q., 2020. Visual commonsense r-cnn. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10760–10770.

Wang, X., Girshick, R., Gupta, A. and He, K., 2018. Non-local neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wang, Y. and Mori, G., 2009. Max-margin hidden conditional random fields for human action recognition. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 872–879.

Wang, Z. and Fey, A. M., 2018. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *International journal of computer assisted radiology and surgery*, 13 (12), 1959–1970.

Williams, R. J., 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8 (3-4), 229–256.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. *International conference on machine learning*, 2048–2057.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H. and Courville, A., 2015. Describing videos by exploiting temporal structure. *Proceedings of the IEEE international conference on computer vision*, 4507–4515.

Zhang, J., Chang, J., Yang, X. and Zhang, J. J., 2017a. Virtual reality surgery simulation: A survey on patient specific solution. *International Workshop on Next Generation Computer Animation Techniques*, Springer, 220–233.

Zhang, J., Nie, Y., Lyu, Y., Li, H., Chang, J., Yang, X. and Zhang, J. J., 2020a. Symmetric dilated convolution for surgical gesture recognition. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 409–418.

Zhang, L., Sung, F., Liu, F., Xiang, T., Gong, S., Yang, Y. and Hospedales, T. M., 2017b. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*.

Zhang, Q., Chen, L., Tian, Q. and Li, B., 2013. Video-based analysis of motion skills in simulation-based surgical training. *Multimedia Content and Mobile Devices*, International Society for Optics and Photonics, volume 8667, 86670A.

Zhang, Q. and Li, B., 2011. Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model. *Proceedings of the 2011 international ACM workshop on Medical multimedia analysis and retrieval*, ACM, 19–24.

Zhang, S., Guo, S., Huang, W., Scott, M. R. and Wang, L., 2020b. V4d: 4d convolutional neural networks for video-level representation learning. *arXiv preprint arXiv:2002.07442*.

Zhao, D., Chang, Z. and Guo, S., 2019. A multimodal fusion approach for image captioning. *Neurocomputing*, 329, 476–485.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A., 2016. Learning deep features for discriminative localization. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zia, A. and Essa, I., 2018. Automated surgical skill assessment in rmis training. *International journal of computer assisted radiology and surgery*, 13 (5), 731–739.

Zienkiewicz, O. C. and Taylor, R. L., 2005. *The finite element method for solid and structural mechanics*. Butterworth-heinemann.