Instance Segmentation of Laparoscopic Instruments and Organs

By

AQSA RIAZ

Fall2015-RIME00000118298

School of Mechanical and Manufacturing Engineering
National University of Sciences and Technology
H-12 Islamabad, Pakistan

Instance Segmentation of Laparoscopic Instruments and Organs

By

AQSA RIAZ

Fall2015-RIME00000118298

Supervisor

Dr. Hassan Sajid

Co-supervisor

Dr. Yasar Ayaz

A thesis submitted in partial fulfillment of the requirement for the degree of Masters of Science In

Robotics and Intelligent Machine Engineering

Department of Robotics and Intelligent Machine Engineering

School of Mechanical and Manufacturing Engineering (SMME)

National University of Sciences and Technology(NUST)
H-12 Islamabad, Pakistan

July 2018


بسم الله الرحمن الرحيم

Dedication

This thesis dedicated to my parents and family who have always supported me throughout my life and inspired me to become the person I am today.

ACKNOWLEDGMENT

Firstly, and foremost praise be to ALLAH the Great and Almighty surrounded me under his auspices during my MS study, in the department of School of Mechanical and Manufacturing Engineering, National University of Sciences and Technology (NUST).

I would like to express my utmost gratitude to my supervisor Dr. Hassan Sajid for his advice, invaluable guidance and comments during experimental work and thesis writing as he has been a regular source of encouragement and support. His extensive knowledge in the field, oral presentation skills, and everyday striving for high standards have been indelibly imprinted on me. I thank Dr. Yaser Ayaz for his never ending willingness to help in a myriad of situations. I thank him for his interest in my topic, especially when I needed new approaches.

Most of all, I wish to extend my love to my family, who have always supported me throughout my life and inspired me to become the person I am today. A must thank to my mother Sajida Riaz for always encouraging me not to do a job halfway and for always being there for me. I owe thank to my father Riaz Ahmad for always helping me to see the bright side of things, for his sage advice. A deep thanks to my siblings for being my best friends and for their help.

Aqsa Riaz.

DECLARATION

I, Aqsa Riaz declare that this thesis titled “Instance segmentation for laparoscopic instruments and organs” and the work presented in it are my own and has been generated by me as a result of my own original research.

I confirm that:

- This work was done wholly or mainly while in candidature for a Master of Science degree at NUST
- Where any part of this thesis has previously been submitted for a degree or any other qualification at NUST or any other institution, this has been clearly stated
- Where I have consulted the published work of others, this is always clearly attributed
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work
- I have acknowledged all main sources of help
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself

Aqsa Riaz,

Fall2015-RIME00000118298

COPYRIGHT NOTICE

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr. / Ms. **Aqsa Riaz** (Reg No. **Fall2015-RIME00000118298**) of **SMME** (School/College/Institute) has been vetted by undersigned, found completed in all respects as per NUST Statues / Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil Degree. It is further certified that necessary amendments as pointed out by GCE members of the scholar have also been incorporated in the said thesis.

Signature: _____

Name of Supervisor: _____

Date: _____

Signature of (HoD):_____

Date: _____

Signature of (Principal):_____

Date: _____

PLAGIARISM CERTIFICATE (TURNITIN REPORT)

This thesis has been checked for plagiarism. Turnitin report endorsed by Supervisor is attached.

Signature of Student

Registration Number

Fall-RIME00000118298

Signature of Supervisor

CONTENTS

TABLE OF FIGURES

LIST OF TABLES

ABBREVIATIONS

MIS	Minimally Invasive Surgery
AI	Artificial Intelligence
NN	Neural Network
CNN	Convolutional Neural Network
2D	Two Dimensional
3D	Three Dimensional
SVM	Support Vector Machine
HMM	Hidden Markov Model
Fig.	Figure
LSTM	Long Short-Term Memory
VGGNet	Visual Geometry Group Network
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
FCN	Fully Convolutional Network
R-CNN	Region- Convolutional Neural Network
RNN	Recurrent Neural Network
COCO	Common Objects in Context
RGB	Red Green Blue
CRF	Conditional Random Field
ML	Machine Learning
ReLU	Rectified Linear Unit
conv	Convolutional
BatchNorm	Batch Normalization
BCELoss	Binary Cross Entropy Loss
BCEWithLogitLoss	Binary Cross Entropy with Logits Loss

IoU	Intersection Over Union
TP	True Positive
FP	False Positive
FN	False Negative
API	Application Program Interface
AP	Average Precision
P	Precision
TH	Threshold
Sec	Second

ABSTRACT

Robot assisted laparoscopic surgeries gained more light in the past few years especially after Da Vinci has been introduced in the medical field. Advancement in AI techniques and robotic procedures led scientist to the advancement in medical surgeries by automate the surgical procedures. In the presented work, we propose a novel segmentation algorithm for identification, label and classification of the tissues, organs and surgical tools in the endoscopic video feed of the human torso region. This thesis serves as the first step towards autonomous minimal invasive surgery. It has two main contributions: first, we contribute an annotated dataset called M2CAISeg created from actual endoscopic video feed of surgical procedures, and secondly, we propose a state of the art deep learning algorithm for instance segmentation. The trained model will be cross-validated followed by comprehensive evaluation on the test set of the proposed dataset.

Chapter 01

INTRODUCTION & MOTIVATION

1.1. Introduction

Human visual and neural system has no difficulty in understanding the picture they are seeing. They easily perceive the information from the environment and act accordingly. Understanding a scene means quickly extracting the gist of a scene which involves cognitive abilities. While looking at picture (Figure 1), Humans quickly classify that it's a scene of some football match. The people in this match are players. They also can categorize the gender and predict the emotions to some extent.

Figure 1- *Sample football match Scene*

In expert systems, the process of analyzing, categorizing, perceiving, enlarging the dynamic images in real time from 2D dynamic scene just like human does, is called scene understanding. The parameters from environment has been taken from network of sensors. In this regard, for characterizing the scene observed from sensor, scene understanding includes both extracting and adding the semantic from sensor data. Scene may have multiple objects of different kind interacting with each other or with their environment. Scene itself can be of different kinds and size it may be consist of an instant movement (eat something), or several hours (driving a car). Sensors used in

network to observe the data may include cameras of different kind (infrared, omni directional), sonar sensors, microphones, radars, smoke detector etc. Scene understanding lies in the area of cognitive vision which means it deals with computer vision, software engineering and cognition, the three major areas.

It is used to target the major problems like

- Detection
- Localization
- Recognition
- Classification
- Understanding

Systems build on scene understanding not only extracts visual feature information by detecting corners or edges but it should be robust, and adaptable network which can adapt, learn from environment variables, weigh alternative solutions and produce more accurate results with new strategies and interpretation.

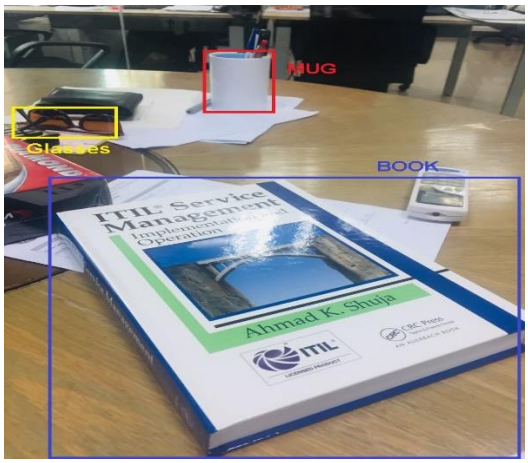


Figure 2- *Scene Understanding*

The capacity to display the robust performance in the situation that is not foreseen when the system is build, is the main characteristic of scene understanding system. The system should expect the events, its possible operations and predict environment future configurations with novel variations that can help to generalize the new context and

communicate to the other system just like humans. Robotics in which system can adapt and modify their environment and multi-media document analysis in which we can get some limited contextual information, are related but somehow different domains.

Scene understanding has wide range of applications due to its general formulation. There are many successful applications build on scene understanding such as swimming pool monitoring, inferring semantic of remote sensing data, traffic monitoring, intrusion detection, robotic surgeries, urban scene understating for automotive applications.

1.2. Motivation

Current surgical procedures, including robotic surgeries requires high level of scene segmentation, skills and domain knowledge. Robotic Surgery is one of the field in which scene understanding is widely used. It's a sub domain of Minimally Invasive Surgery (MIS) in which surgeries are done through tiny holes with the help of mini size surgical instruments. These Robot assisted surgeries are well-known for its precision, control and flexibility and way head of conventional procedures. Clinical robotic surgery systems consist of mechanical arms having surgical instruments and camera arm. Da Vinci is the advanced surgical robot, widely used throughout the advance US hospitals. Doctors or surgeon operate these robots through control panel inside or outside the operating room.

The advancements in AI making an impact on all other field just like advancement in autonomous or self-driving car. These advancement can be made in medicine and health care applications, a field which has lot of research potential. Right now robots are used by surgeons but like autonomous car, we can utilize the AI techniques to automate surgical procedures through recent advancements in AI and Robotics. The proposed work lay down the foundation in this direction.

1.3. Goal

Goal of this proposed work is to make a robust algorithm which perform instance segmentation (Classification + Semantic Segmentation) on medical images of human torso region. This work is a first step to autonomous surgical robot. This algorithm is

designed for laparoscopic cholecystectomy, minimal invasive surgery for removing gallbladder from human torso region.

Chapter 02

DATASET

2.1. Annotation Methodology

We have studies the laparoscopic surgical procedures in details. A laparoscopic surgery includes small incisions with mini instruments specially designed for these procedures called laparoscopic instruments, instead of large open incision. Procedure takes place by inserting small video camera through small incision and surgeon operates through big screen on which camera live feed is coming. Laparoscopic surgeries are of many types:

- Laparoscopic Cholecystectomy
- Laparoscopic Hysterectomy
- Laparoscopic Appendectomy

Through deep studies of these procedure we have identified that data for dataset is largely and easily available for cholecystectomy. So our focus is on laparoscopic cholecystectomy. To understand the procedure, good knowledge of human torso region, procedure, and laparoscopic instruments is necessary. Procedure of removing gallbladder is called cholecystectomy. For digestion, liver of human body produces bile and juices and to help digestion bile is sent through bile duct in to the small intestine by gallbladder. But some time it gets necessary to remove gallbladder when gall stones also called cysts are diagnosed. These cysts blocks the bile duct and result in serious damage to human digestive system. Gallbladder can be removed through two type of procedures:

- Open gallbladder surgery
- Laparoscopic cholecystectomy

First type of procedure includes large incision and removal of gallbladder. It takes lot of time to heal after the treatment besides it's quite painful as well. Second one is simple,

easier and faster recoverable method which includes several small incisions of about few mm in abdomen region. Recovery time is of one week.

2.1.1. Laparoscopic surgical instruments

There are many instruments with different sizes and quality have been used in surgery. It mainly depends upon the surgeon's choice. Mostly used laparoscopic instruments include: Trocar, grasper, Bipolar, Hook, Scissors, and Clippers etc.

Graspers [Figure: 3] are used to grasp and manipulate between the tissues during surgeries with precision . There are different kind of graspers use in surgeries by different surgeons. Grasper for cholecystectomy are different from the graspers use in appendectomy. It is also used to pull out the gallbladder through specimen from abdomen.

Figure 3*-Grasper (Surgical Tool)*

Bipolar [Figure: 4] provide the function of coagulate and cut between vessels and tissues without changing the instrument. It is of different sizes, adjustable to small hands. Its handle is adjustable (rotated through 360) and came with extra accessories.

Figure 4*- Bipolar (Surgical Tool)*

To dissect, mobilize and cut the tissues or vessels, surgeons use different type of hooks [Figure: 5] and scissors [Figure: 6].They may be of straight or curved type. Scissors are used to cut different type of tissues mostly containing staples while to cut joint, tough fibrous tissues hook is used.

Figure 5*-Hook (Surgical Tool)*

Figure 6*-Scissor (Surgical Tool)*

Irrigator suction is used to suck the blood and other biles to clear the operative area for precision. Blood vessels need to be tied up when to be cut and for that clippers are used. Clippers can be disposable having 20 clips or reusable having one clip at a time. Specimen bags are made of strong material, use to contain the cysts and other masses from operative area. They are impervious to cancer cell.

2.1.2. Dataset preparation

Dataset preparation is a joint work with Salman Maqbool. We have used M2CAI 2016 Surgical Tool Detection Dataset [3]. This dataset contains tool presence annotation for 7 tools after every 25 frames. It has total of 15 videos lasted for almost 2 to 3 hrs. For training we separate 10 videos and for testing we take 5 videos. We have annotate pixel level annotation of the different organs and surgical instruments. Total 350 images have been annotated. Dataset classes are identified and unique RGB values have been assigned to each of the class [table 1].

Organs	**Assigned RGB Values**	**Instruments**	**Assigned RGB Values**
Liver	0,170,85	Grasper	170,85,0
Gallbladder	255,170,85	Bipolar	255,85,0
Fat	0,255,85	Hook	85,170,0
Upper wall	170,255,85	Scissors	85,255,0
Intestine	255,0,255	Clipper	170,255,0
Artery	255,0,170	Irrigator	170,0,85
Black	0,0,0	Specimen Bag	255,0,85

***Table 1**-RGB values for Identified Dataset Classes*

We have annotate difficult regions and regional boundaries manually using Photoshop and MS Paint. We develop a tool for the annotation who assign colors to image regions using MegaPixels in Matlab. The tool works on Nearest Neighbor search in RGB space and smooth boundary values. It apply 5 x 5 median filter to remove the resultant salt and pepper noise. Few samples of dataset are mentioned in the Figure7.

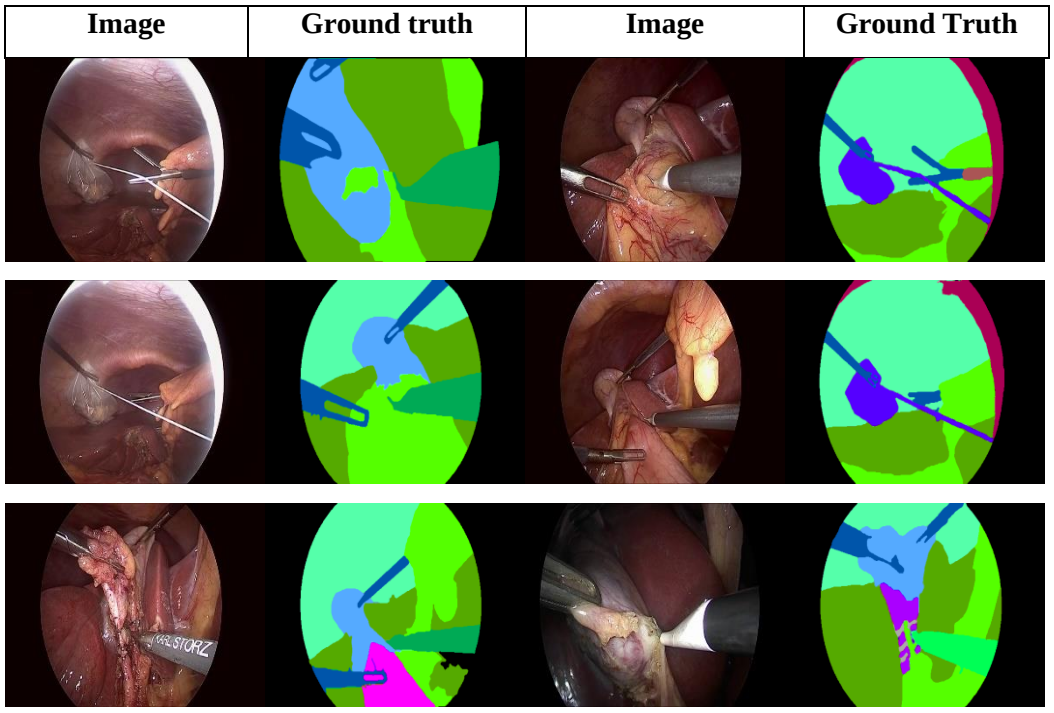


***Figure 7**-Dataset Images with their Ground Truth*

2.1.3. Difficulties while preparing dataset

- Time taken: Approximately 30 minutes per image over 2 months.
- Redundant work so hard to keep focused.
- Lack of outsourcing opportunities.
- Variations in data - Sometimes difficult to classify a region in a single category.
- Lack of access to specialized medical personnel for guidance.
- General difficulty of the pixel level annotation.

89Chapter 03

LITERATURE REVIEW

3.1. Laparoscopic Image and Video Analysis

Minimum Invasive Surgery (MIS) is now becoming successfully leading method of treatment use in medical procedures. This low risk technique is budget friendly and cost effective. It is preferred by most of the patients due to its fast recoveries and less healing time. To enhance the patient safety, and assist the surgeons to navigate tools by reproducing the training programs, automatic analysis of videos of minimum invasive surgeries potentially play the vital role. Surgical videos are main reliable resource for the information. To achieve the purpose the successful results of 3D reconstruction of the surgical field area and tool tracking have been achieved. [20] (Patricia Sánchez-González *et al*., 2011)

To reduce the damage risks while intervention, the most important task is to the keep the tools centered in the field of view of the image. Minimum invasive surgery techniques lack the orientation characteristics and depth perception and to overcome the problems 3D localization of tools play the pivot role. Which itself is the critically separate issue. To guide the surgeon about the height of sensitivity of delicate areas, the 3D tool presence knowledge is important. The loss of depth information of map linked with endoscopic camera's image is the limitation to the laparoscopic surgery. [21] (Patricia Sánchez-González *et al*., 2011)

3.1.1. Surgical phase identification

Segmentation of videos of robot assisted surgeries and laparoscopic videos in terms of context has been shown the performance and efficiency of work flow. It may use for education about the surgeries and even time critical consultation. Modern science wants to exclude the manual video analysis. And many policies implemented in hospitals and

substructure running in the health care centers prevent the recordings of the procedure and save the large amount of data. To overcome the situation a system had presented that generates laparoscopic and robot assisted surgeries based video segmentation while using the minimum resources. It generates the segmentation videos automatically on surgical procedures according to their surgical phases with minimum training data. Without needing of analysis of non- rigid and variable environment, this system takes variability of video stream by using the combination of SVM and HMM with an augmented feature space. [4] (Mikhail Volkov *et al.*, 2011)

Surgical work flow gradually becoming more important in the tasks like scheduling the surgeons, updating and optimization of scheduling of operating room, prediction of upcoming events, if something goes wrong it alerts and suggests the modifications and monitoring the surgical process, etc [6]. This paper was the result of c challenge called M2CAI workflow Challenge and ranked in top three performing solutions. The task in hand was to analyze and detect the phase from the videos of MIS operation surgeries. The paper presented a very profound residual convolutional network consists of 50 layers called *EndoRCN* is extracting feature from visual information of the videos. These features are highly discriminative and later encoded by using LSTM network to complete the video analysis tasks. [7] (Yueming Jin *et al.*, 2016)

3.1.2. Detection of Tool presence

Many medical applications such as database video's automatically indexing, optimization of real time operating room's scheduling etc are using surgical workflow recognition. Previously this task has been done with using tool usage signals and visual features. Data collected for both are mostly manual handcrafted or by using manual annotation process. The novel approach has been introduced in this paper in which EndoNet (a convolutional neural network architecture) has been used for phase detection. The network learns the feature automatically from surgical videos. This network first trained through fine tune process then this trained network use for both tasks: tool presence detection or phase recognition. After that network extracts the visual features and to obtain the final estimated phase, these feature pass through the

Hierarchical HMM and Support Vector Machine (SVM). [3] (Andru P. Twinanda *et al*, 2016)

CNNs are currently use in almost all type problems related to the ML and CV. One of the major and complicated problem is feature extraction which requires expert knowledge of the field. The methodology to solve this differs from one task to another. The solution presented in this paper is the result of combination of two deep convolutional neural network architectures: VGGNet [9] and GoogleNet [10]. To achieve the better accuracy they customized the parameters of both networks. The solution presented in this paper ranked 1st in the M2CAI Tool presence detection challenge. [8] (Ashwin Raju *et al.*, 2016)

3.1.3. Surgical Tool Segmentation

Segmentation is much more complicated and tough task than image classification and object detection. In image classification we classify the object in to different object classes present in the image and in object detection we identify the object by using bounding boxes with in the image while in segmentation image is converted in to segments which help to analyze the image. It classify each pixel to its object class which means a label for each pixel. FCN is the architecture modified from AlexNet [2], VGGNet [9] GoogleNet [10] by transforming fully connected layers to fully convolutional layers. This architecture takes non fixed size input and produce many feature maps with dense representation and small sizes and to create the output of the same size segmented image, used unsampling technique. It extract abstract information from upper layers to resolve the problems of "what" and for the solution of "where" problems it extracts the lower layer's local information. [11] (Jonathan Long *et al.*, 2015)

In this research a novel approach with two versions (real-time and non-real-time) have been presented which is based on fully convolutional neural networks that resulted in high speed of optical flow along with deep and accurate segmentations of highly deformable parts. Low amount of images have been used to fine tune the FCN. There

was no need to use hand craft features. Method has been tested on benchmark datasets which include ex vivo and in in vivo cases. [12] (Luis C.Garica-Peraza-Herrera, *et al.*, 2016)

For flow identification and phase recognition, segmentation of tools use in surgeries are important. It helps to track and estimate the tool angle and pose in the surgical scene. In domain of context aware surgical systems, flow identification remains the unresolved task. The proposed method hybrid in this work is utilizing both CNN and RCNN to get the highest accuracy of surgical tools segmentation. EndoVis a MICCAI 2016 Endoscopic Vision Challenge Robotic instruments dataset is used for training and testing the model and achieved better results compared to the state of the art models on the same dataset. [13] (Mohamed Attia *et al.*, 2017)

In many computer assisted surgeries the most important part is real-time tool segmentation from endoscopic video. It also has a crucial importance in robotic surgical data science. In this paper two novel approaches based on deep learning architecture has been proposed for segmentation of non-rigid surgical tools automatically. Both architectures extract the multi-scale feature extraction without losing the quality of accurate segmentation at all resolutions. Multi scale constraints are encoded inside the network architecture, first method implemented it by cascading aggregation of prediction and other one enforce it by means of holistically nested architecture. Both architecture have less number of parameters as they proposed for real-time semantic labeling. Both architectures are validated on ex vivo dataset having multiple tools presence. The methods show the significant results in terms of accuracy as compared to the state of the art convolution networks. [14] (Luis C.Garica-Peraza-Herrera *et al.*, 2017)

In this work, the technique has developed which incorporate the real-time surgical tools segmentation with color images. The method is develop in the context of robot assist laparoscopic surgeries. It detects and tracks the gray regions and investigates the metallic instrument images inside the abdominal cavity. First of all the difficulties such as changing lighting conditions, non-uniform background environment due to breathing,

and visual appearances of specular reflections, need to be tackled. Then for laparoscopy a technique is developed which extract the discriminant color feature with significant capabilities of dealing variations and specularities and by doing this it achieved an automatic color segmentation.[15] (C.Doignon *et al.,* 2005)

The study in this paper also tackled the problem of specular reflections, non-uniform backgrounds and camera position, image blurriness and resolution by proposing a technique that have advantage of interdependency between the surgical instruments localization and segmentation. By reformulating 2D surgical tool's pose estimation as heatmap regression which resulted in enabling the simultaneous, robust and real time regression for both tasks using deep learning. Experiments shows the significant improvement in performance than regressing the tool presence directly. [18] (Iro Laina *et al.,* 2017)

3.2. Instance Segmentation

The domain that dealt with the problems of classification and semantic segmentation and solved them simultaneously is called Instance. In this regard a model called MaskLab has been presented in this paper. Mask lab produces three outputs, first is box detection using Faster-RCNN, from pixel wise segmentation it produces its second output which is semantic segmentation logits and last output is direction prediction logits. It predict the each pixel's direction towards its instance center. Then utilizing semantic segmentation and direction prediction it perform background and foreground segmentation. Then apply predicted cox label on semantic segmentation logit and according to the predicted box it crops the region. For each channel, it assemble the regional logits by performing direction pooling on direction prediction logits. Then it concatenated the both cropped features and applied 1 by 1 convolution on it to get final instance segmentation.[22] (Liang-Cheih Chen *et al.*, 2017)

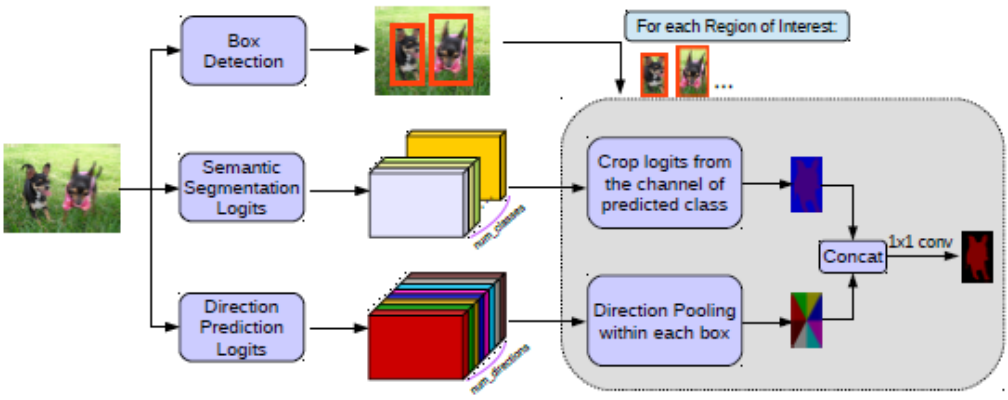


***Figure 8**-MaskLab complete methodology*

The approach used in this paper for object instance segmentation is quite flexible and conceptually simple. The model which is called Mask R-CNN is extended from Faster R-CNN by adding a branch that predict masks for objects. It has the advantage of detecting objects in the image by using Faster-RCNN while predicting an object mask simultaneously. Mask R-CNN is easy to train. It is train and tested on MS-COCO and gave significant results for all three COCO suite of challenges. [23] (Kaiming He *et al*,. 2017)

In this paper the approach is consist of two main modules: semantic segmentation subnetwork and instance segmentation module. This model produces the segmentation map in which each pixel has assigned an object class and identity label of instance. The output of semantic segmentation module is passed to the instance subnetwork. Along with the cues form object detector, it uses the initial category level segmentation within CRF tto predict the instances. This approach doesn't require any post processing. [24] (Anurag Arnab & Philip Torr , 2017)

Chapter 04

METHODOLOGY

4.1. Problem Statement

The first thing for any robot is to see and understand the scene, and then plan its actions. Surgical procedures in which robotic systems are indulge, demand high level abstraction n in data and robust representations to overcome the limitations. And for that high level scene understanding is necessary. Our work focuses on the general broad category of scene understanding. In particular, on pixel level scene understanding.

Goal: Segmentation + Classification at Pixel

4.2. Proposed Network

The network architecture is consisted of an encoder and decoder followed by pixel wise classification layer. The encoder is based on the typical architecture of a convolutional network. It is consisted of the repetitive application of a 4x4 kernel size convolutional layer with stride 1 and no padding, then normalize the output of convolutional layer, we use Batch normalization layer. After that a rectified linear unit layer (ReLU) has been used. To reduce the size of representation or down sample we use stride 2 in each layer of encoder except the first one. We double the no. of feature channels at each convolutional layer step. In decoder every step is consisted of a convolutional transpose

layer which up sample the input, batch normalization layer and dropout layer used for regularization and to enable generalization of the training followed by ReLu layer. Drop out layer used in first 3 decoder layers. To produce class probabilities for each pixel independently, one of the output produced by final decoder layer sent to a multi-class soft-max classifier. 2nd output fed to the classifier consists of convolution layer and sigmoid for the tool classification. Architecture is designed for the less data, performs less number of iteration. It uses pre-trained values.

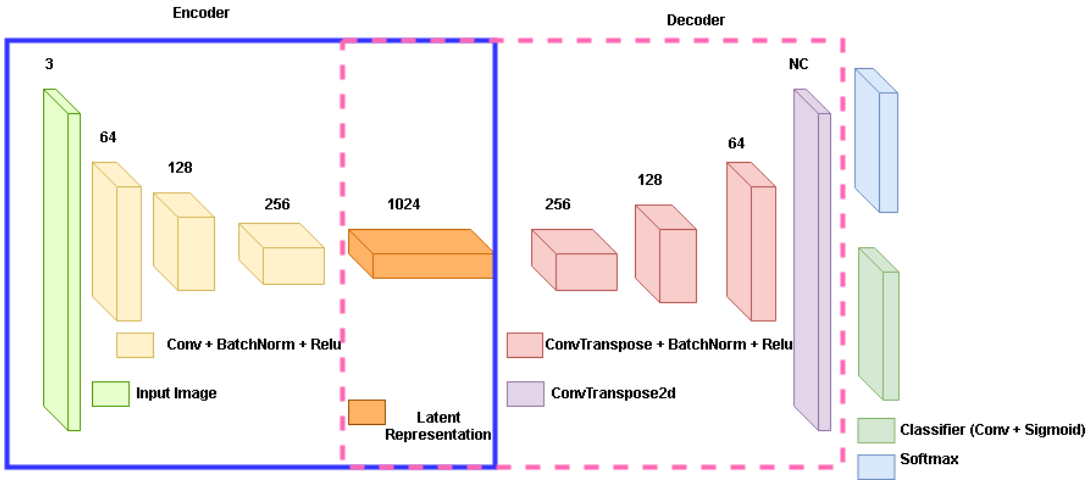


***Figure 9**-Proposed Architecture*

4.2.1. Input Preprocessing

Before feeding the input image to the network, few operations have been applied on the image. The size for all input images should be 256 x 256. Online TenCrop data augmentation is used during training including 224 x 224 side and center crops and their horizontal flips. Input image has been normalized by using the per-channel mean [0.295, 0.204, 0.197] with standard deviation [0.221, 0.188, 0.182] of complete M2CAI 2016 surgical tool detection frames. The values were calculated over the 581935 images.

4.2.2. Convolutional layers

Convolutional layers consists of filters with parameters and wrights that need to be learned. These filters height and weights are smaller than the input's height and weight

the dot product of these learnable filters is computed with the input images. These filters slide across the input image and calculate dot product of input and filter. Parameters used in proposed networks are:

- Kernel size : 4 x 4
- Stride : 2
- Padding:1 x1

For the last layer of the encoder stride is 1 with 0 padding.

4.2.3. Batch Normalization Layers

During training pre-processing on input helps normalization of input data to prevent the saturation of non-linear classifier. Problem of internal covariate shift arises in intermediate layers when change in activation distribution happens constantly during training. To avoid or tackle the covariate problem, we have used batch normalization method. It calculates the variance and mean of inputs of layers.[19] After reviewing the previous calculated batch statistic it normalize the next layer inputs and to obtain the layer output it applies scaling and shifting. It also give the slight regularization effect, since we are using mini-batches and adds slight noise to the activations. We don't use batch normalization in the first encoder layer.

4.2.4. Rectified Linear Unit (ReLU) Layers

We have used ReLU, a non-linear activation function which enable the learning of non-linear functions. This fast converging activation function is most commonly uses in covnets. It is very easy to use and practically works well. Computation of ReLU is quite simple. For all the positive values it gives linear identity and for negative values it give zero.

4.2.5. Convolutional Transpose Layer

Transposed convolutional layers are known for the decompression of compressed representation into different domains. In our proposed network we are using encoder and decoder, encoder is doing the compression job. This layer is the mirror to the conv layer just to make sure that the size of output is same as input side. It also enable the

network to be used by images of any size. To up-sample the input we fractionally pad the input and then perform the convolution on it.

Parameters used in this section are:

- Kernel size : 4 x 4
- Stride : 2
- Padding:1 x1

For the first decoder layer, we use:

- Stride:1
- Padding :0

4.2.6. Dropout Layers

To counter the over-fitting problem we have used dropout layer which randomly dropout the nodes and improve the generalization and regularization of convolutional neural network. We have use 3 dropout layers in our first three decoder layers. Network uses 2D variant of dropout, where filter maps made zero with a probability of 0.5 during training. Over an ensemble of models, it acts like averaging predictions at test time.

4.2.7. SoftMax and Classifier

Softmax layer is used at the very end along with the classifier to convert the network predictions in probabilities. Probability is calculated for each pixel for it belonging to each class. Classifier consist of convolutional layer with sigmoid function to classify the tools in the images.

4.3. Training Details

Dataset:

- Split into Training and Test set
- Training Set: 245 images
- Test Set: 62 images

Optimizer:

- Adam with the default parameters:
    - beta1 = 0.9
    - beta2 = 0.999
    - epsilon = 10e-8
    - Weight Decay = 0.0005
    - Step Learning Rate Decay:
    - Initial Learning Rate: 0.0001
    - Step Decay: Decrease by half after every 10 epochs

- Loss Function: BCEWithLogitsloss
- Epoch: 1500
- Batch Size: 2

Unsupervised pre-training over the whole M2CAI training set (581935 images) for image reconstruction for 1 epoch.

- Learning Rate: 0.01
- Batch Size: 64
- Mean Squared Error Loss

4.4. BCEWithLogitsLoss

This loss combines a Sigmoid and the BCELoss in one single class. The difference between the network output and ground truth of images has been computed. It is computed depth-wise, for every pixel, and then averaged across all pixels. Adam Optimizer is then used to minimize the loss.

Chapter 5

RESULTS

5.1. Evaluation Criteria

For evaluation criteria we used the Jaccard Index / Intersection over Union (IoU)

$$IoU = TP / (TP + FP + FN)$$

It is evaluated for each class, and the overall mean on the dataset is reported. Pixel-wise criteria is used while evaluating the True Positives (TP), False Positives (FP), and False Negatives (FN).

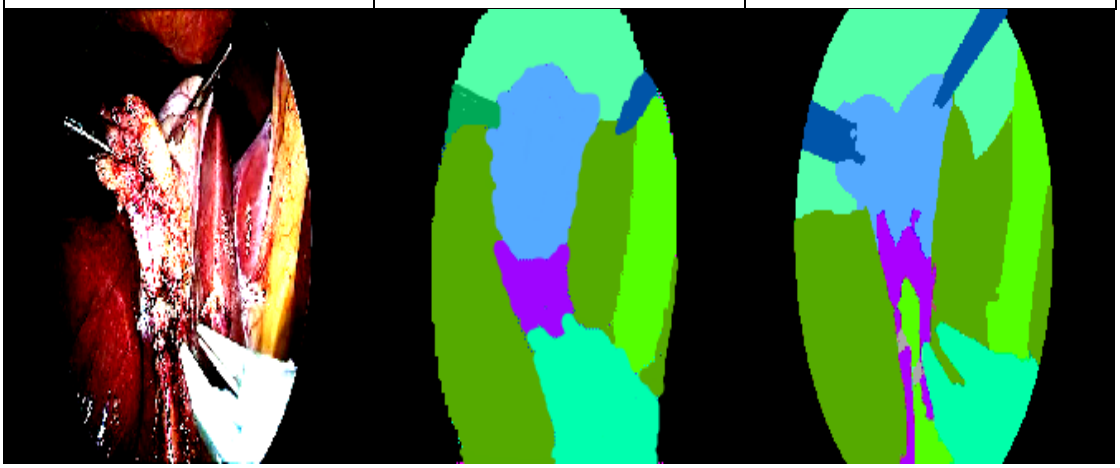
We presented the class-wise Precision, mean, Recall, and the F1 scores:

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1\ Score = (2 * Precision * Recall) / (Precision + Recall)$$

Following are the results of instance segmentation for laparoscopic tools and organs of torso region

Original Image	Generated Image	Ground Truth
		

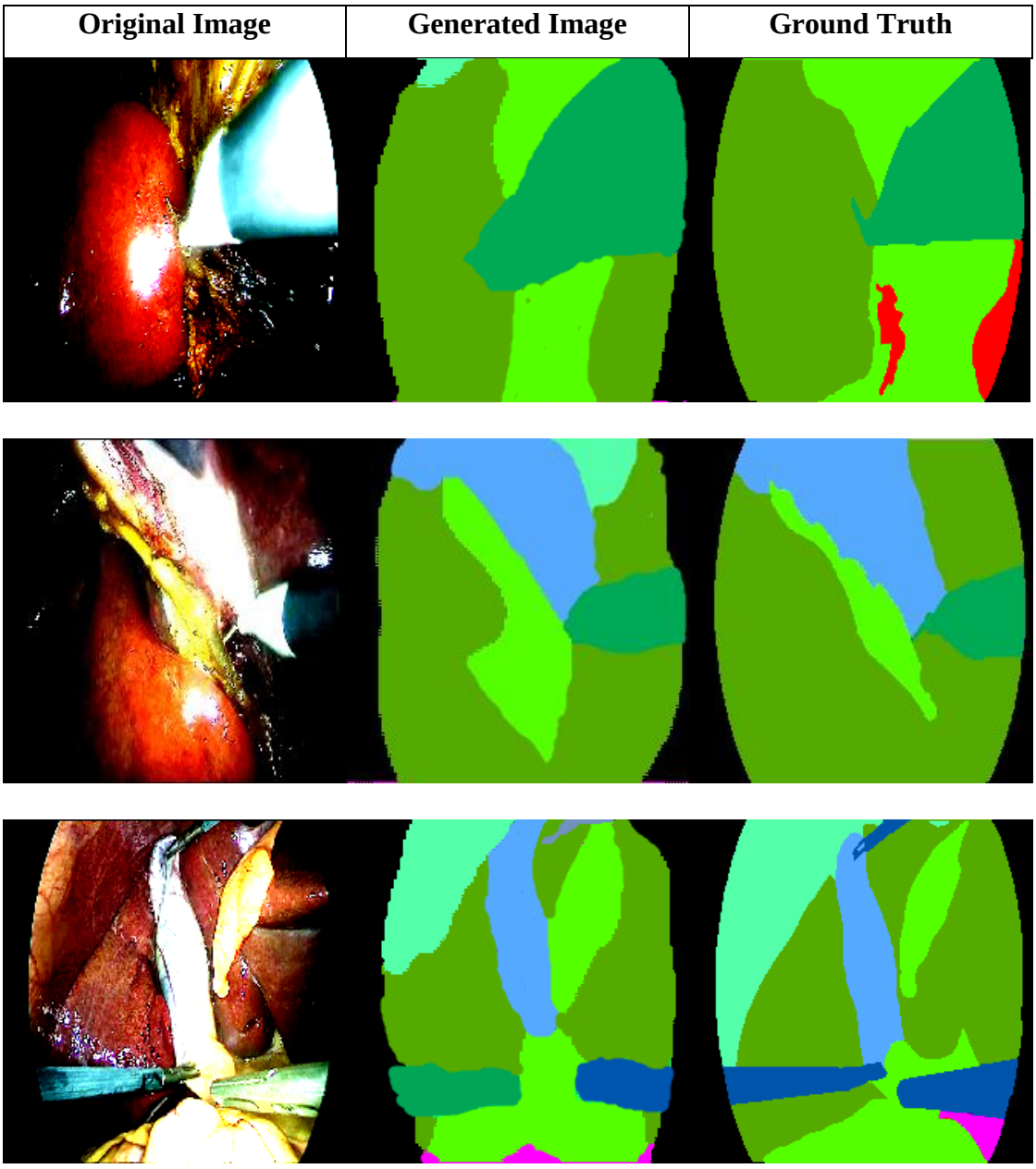


***Figure 10**-Image representation of Instance Segmentation for laparoscopic tools and torso region organs*

The architecture is created to achieve the results on small size dataset. The mean precision, recall and F1 score of the organs and tools are 0.548, 0.4841 and 0.4951 respectively.

Class / Measure	Precision	Recall	F1 Score
Grasper	0.6241	0.6434	0.6333
Hook	0.6220	0.5907	0.6059
Scissor	0.0170	0.1439	0.0048
Clipper	0.5052	0.5218	0.5134
Irrigator	0.4281	0.1904	0.2635
Specimen bag	0.5021	0.3271	0.3960
Liver	0.6841	0.6934	0.6887
Gallbladder	0.5942	0.5151	0.5518
Fat	0.5425	0.6564	0.5940
Upper Wall	0.6233	0.4959	0.5483
Artery	0.4521	0.1615	0.2379
Intestine	0.4912	0.3933	0.4368
Black	0.9625	0.9610	0.9617
Mean	**0.5408**	**0.4841**	**0.4951**

Table 2*-Results of Instance segmentation on small dataset*

Comparison of our proposed architecture with MaskR-CNN.

Organs and Laparoscopic Tools	F1 Measurement (Proposed Network)	F1 Measurement (MaskR-CNN)
Grasper	0.6333	0.2934
Hook	0.6059	0.2019
Clipper	0.5134	0.1921
Irrigator	0.2635	0.1069
Specimen bag	0.3960	0.1489
Liver	0.6887	0.3745
Gallbladder	0.5518	0.2785

Fat	0.5940	0.2812
Upper Wall	0.5483	0.2616
Artery	0.2379	0.0512
Intestine	0.4368	0.1569
Black	0.9617	0.6303
Over all Mean	**0.4951**	**0.2331**

Table 3- *Approximate F1measurement between our proposed networks with MaskR-CNN*

Qualitative analysis of both architecture is represented in form of bar graphs.

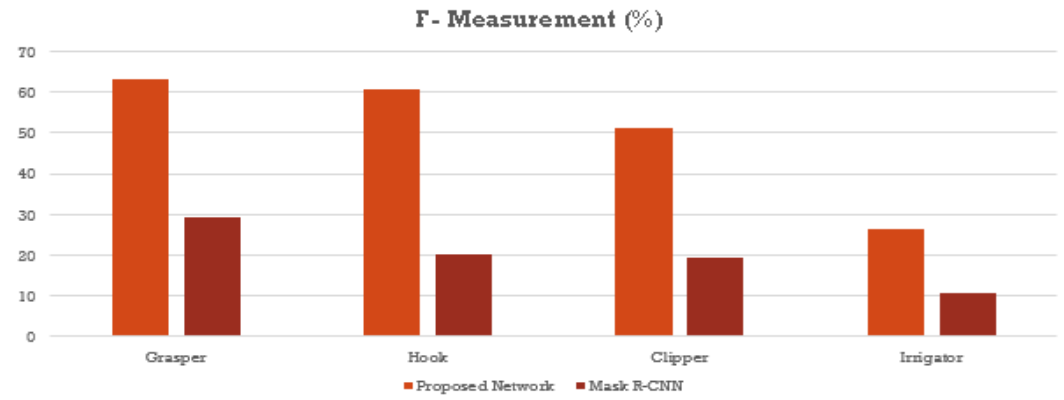


Figure 11*-Qualitative analysis for Grasper, Hook, Clipper and Irrigator of proposed network and Mask R-CNN*

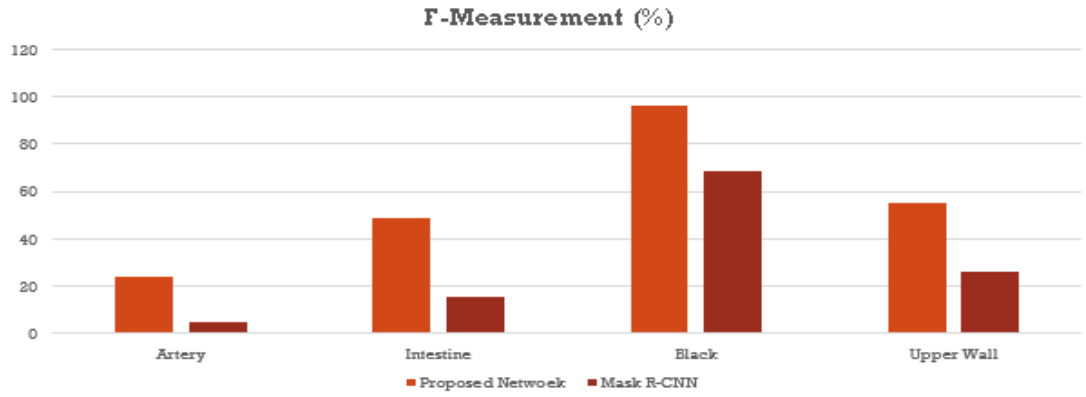


Figure 12*-Qualitative analysis for Artery, Intestine, Black and Upper Wall of proposed network and Mask R-CNN*

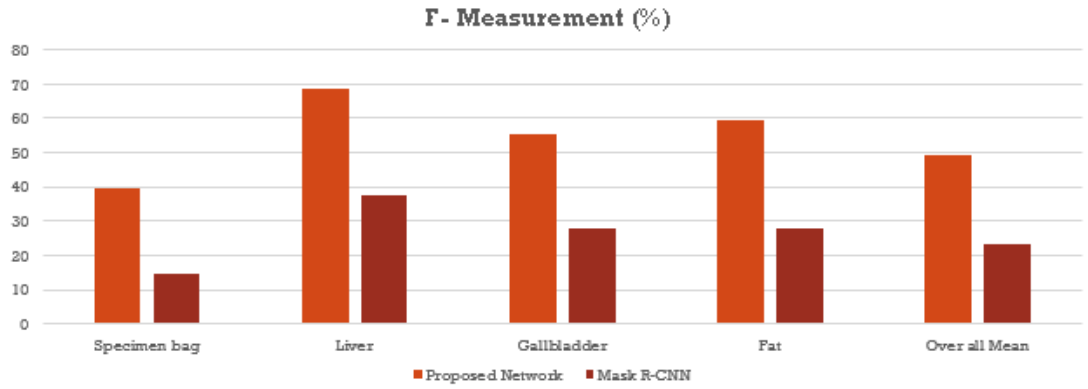


Figure 13*-Qualitative analysis for Specimen bag, Liver, Gallbladder and Fat and overall of proposed network and Mask R-CNN*

5.2. Failure Cases

There are some failure cases as well on which our proposed architecture was fail to produce results.

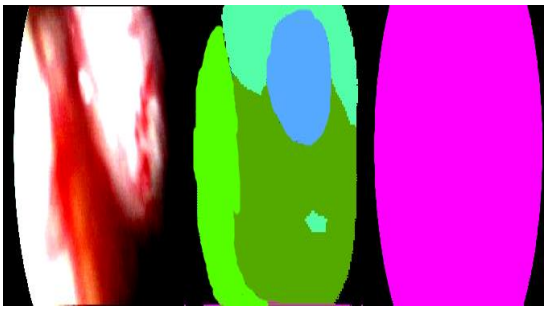


Figure 14- *Failure Case 1*

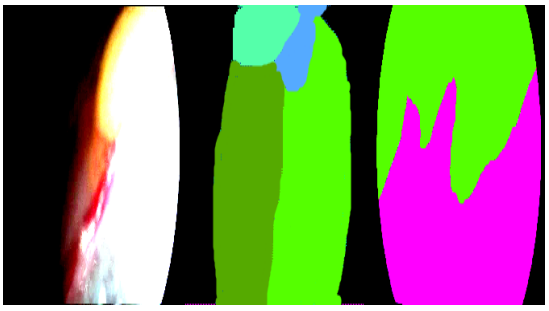


Figure 15*-Failure Case 2*

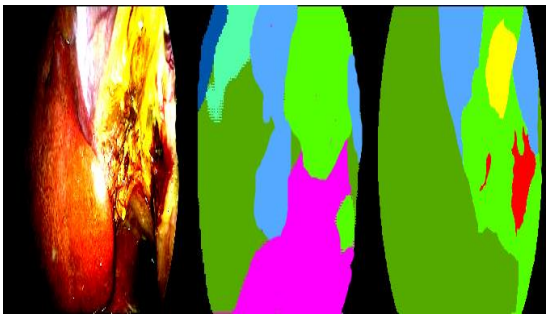


***Figure 16**-Failure Case 3*

CONCLUSION AND FUTURE ASPECTS

Segmentation and classification of medical images have been the active research area in bio medical for past few years. Despite the lot of work has done in this field still there are more challenges to cover. One of the problems is having a dataset with ground truth. From actual endoscopic video feed of surgical procedures we have created an annotated dataset. Most of the work done on medical images are either on semantic segmentation or classification of the object present in the image. Our proposed technique tackle the both problems simultaneously which is called instance segmentation.

There is still lot of capacity in which many improvements can be done. We have only focus on cholecystectomy in which we incorporate only major tools. Instance segmentation for other kind of laparoscopic surgeries is also under discussion. Furthermore in these major tools, there are varieties in term of shapes, size and properties which need to be catered. Right now the dataset is created from video feed but there is an opportunity of creating dataset from live video feed in future.

REFERENCES

[1] Brostow, G. J., Fauqueur, J., & Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, *30*(2), 88-97.

[2] Song, S., Lichtenberg, S. P., & Xiao, J. (2015, June). SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR* (Vol. 5, p. 6).

[3] Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., & Padoy, N. (2017). Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, *36*(1), 86-97.

[4] Volkov, M., Hashimoto, D. A., Rosman, G., Meireles, O. R., & Rus, D. (2017, May). Machine learning and coresets for automated real-time video segmentation of laparoscopic and robot-assisted surgery. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*(pp. 754-759). IEEE.

[6] Stauder, R., Ostler, D., Kranzfelder, M., Koller, S., Feußner, H., & Navab, N. (2016). The TUM LapChole dataset for the M2CAI 2016 workflow challenge. *arXiv preprint arXiv:1610.09278*.

[7] Jin, Y., Dou, Q., Chen, H., Yu, L., & Heng, P. A. (2016). EndoRCN: Recurrent Convolutional Networks for Recognition of Surgical Workflow in Cholecystectomy Procedure Video. *IEEE Trans. on Medical Imaging*.

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

[9] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*

[10] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015, June). Going deeper with convolutions. Cvpr.

[11] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

[12] García-Peraza-Herrera, L. C., Li, W., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., ... & Ourselin, S. (2016, October). Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. In *International Workshop on Computer-Assisted and Robotic Endoscopy* (pp. 84-95). Springer, Cham.

[13] Attia, M., Hossny, M., Nahavandi, S., & Asadi, H. Surgical Tool Segmentation Using A Hybrid Deep CNN-RNN Auto Encoder-Decoder

[14] Garcia-Peraza-Herrera, L. C., Li, W., Fidon, L., Gruijthuijsen, C., Devreker, A., Attilakos, G., ... & Ourselin, S. (2017). Toolnet: Holistically-nested real-time segmentation of robotic surgical tools. *arXiv preprint arXiv:1706.08126*.

[15] Doignon, C., Graebling, P., & De Mathelin, M. (2005). Real-time segmentation of surgical instruments inside the abdominal cavity using a joint hue saturation color feature. *Real-Time Imaging*, *11*(5-6), 429-442.

[16] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. CoRR abs/1512.03385 (2015).

[17] Pakhomov, D., Premachandran, V., Allan, M., Azizian, M., & Navab, N. (2017). Deep Residual Learning for Instrument Segmentation in Robotic Surgery. *arXiv preprint arXiv:1703.08580*.

[18] Laina, I., Rieke, N., Rupprecht, C., Vizcaíno, J. P., Eslami, A., Tombari, F., & Navab, N. (2017, September). Concurrent segmentation and localization for tracking of surgical instruments. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*(pp. 664-672). Springer, Cham.

[19] Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456).

[20] P. Sánchez-González et al. Laparoscopic video analysis for training and image guided surgery, *Minimally Invasive Therapy and Allied Technologies*, vol 20, pp. 311-320 2011(ISSN: 1364-5706).

[21] Sánchez-González P, Oropesa I, Díaz S et al. How can video analysis help laparoscopic surgeons? , Graz, Austria: Proc's 2011 SCATh Joint Workshop on New Technologies for Computer & Robot Assisted Surgery; 2011

[22] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In CVPR, 2018

[23] K. He, G. Gkioxari, P. Doll´ar, and R. Girshick. Mask R-CNN. *arXiv:1703.06870*, 2017

[24] A. Arnab and P. H. Torr. Pixelwise instance segmentation with a dynamically instantiated network. In CVPR, 2017