Semantic Segmentation of Surgery Tools Using U-Net and U-Net16

Soumya Sharma

Department of Electrical Eng. IIT BHU Varanasi, India

Email:soumya.sharma.eee22@itbhu.ac.in

Karan Singh

Department of Electrical Eng. IIT BHU Varanasi, India

Email:karan.singh.eee22@itbhu.ac.in

Abstract— **This report presents semantic segmentation models designed for the identification and labeling of instruments in robotic-assisted surgeries. The objective is to accurately classify each pixel within the surgical scene according to its corresponding instrument class. Semantic segmentation of surgical instruments serves as a fundamental task in the development of robotic -assisted surgery systems, facilitating instrument tracking, pose and surgical phase prediction ,etc.. By developing robust methods for semantic segmentation of surgical instruments, this project aims to drive progress across multiple domains of study**

Keywords—Semantic Segmentation, U-Net, VGG16, Surgery Tools

I. Introduction

Robot-assisted minimally invasive surgery (MIS) has revolutionised the field of medicine by bringing the advantages of laparoscopic surgery—such as reduced trauma and faster time of recovery—to a wider range of procedures and patients. Surgeons now have improved precision and control over anatomy using advanced instruments and high-definition 3D vision systems. To further enhance surgeon capabilities, there's a need to integrate various data sources, comprising pre- and intra-operative surgical imaging, with the endoscopic view. This requires selectively presenting relevant information to surgeons without overwhelming them with unnecessary data. Key to this is identifying objects in the endoscope's view and understanding their significance. Pixel-wise segmentation of endoscopic images, using deep convolutional neural networks (CNNs), is crucial for this task. The aim of our work is to implement semantic image segmentation of individual frames obtained from a robotic arm camera feed in robot-assisted surgery, providing surgeons with a clearer and more intelligent view of the surgical environment to support informed decision-making during procedures.

II. Model Description

A. U-Net

U-NET, a specialized architecture for Biomedical Image Segmentation, was developed by Olaf Ronneberger and his team at the University of Freiburg in Germany in 2015. Since its inception, it has gained widespread recognition as one of the most efficient approaches for semantic segmentation tasks. Renowned for its capacity to learn from limited training data, U-NET stands out as a fully convolutional neural network.

U-NET — Network Architecture

U-NET is characterized by a U-shaped architecture, featuring four encoder blocks and four decoder blocks linked by a bridge. Within the encoder network, termed the contracting path, spatial dimensions undergo halving, while the number of filters (feature channels) doubles at each encoder block. Conversely, the decoder network doubles the spatial dimensions while halving the number of feature channels.

1.Encoder Network

Within the encoder network, its primary function is to act as a feature extractor, the image is passed through a series of encoder blocks such that the number of feature maps gets doubled in the successive blocks as we move from the top to the bottom of the U-Net. Each encoder block that we are using comprises of two 3X3 conv blocks which is followed by Rectified Linear Activation Unit (ReLU) which introduces non-linearity. The output of the ReLU block serves as a skip connection for the upcoming decoder block.Following the ReLU blocks we have a 2X2 max-pooling layers This is done in order to reduce the computational burden on our model. The spatial parameters of the output of the previous layer is halved.

2.Skip Connections

The skip connections serve two primary purposes: enhancing the decoder's ability to produce more refined semantic features and facilitating the flow of gradients to previous layers, thereby preventing degradation. Thus, skip connections act as shortcut connections, allowing gradients to flow more effectively during backpropagation. This streamlined gradient flow aids the network in acquiring improved representations, ultimately contributing to enhanced learning capabilities.

3.Bridge

The bridge plays a pivotal role as the essential connection between the encoder and decoder networks, guaranteeing the seamless transmission of information. It consists of two 3x3 convolutions, each of which is succeeded by a ReLU activation function.

4.Decoder Network

The decoder network consists of transpose convolution layers that are used for upscaling the image which were downscaled buy the encoder network. The number of transpose conv layers is identical to the number of conv layers present in the encoder path and they are also concatenating with the skip connection from the corresponding encoder conv block.

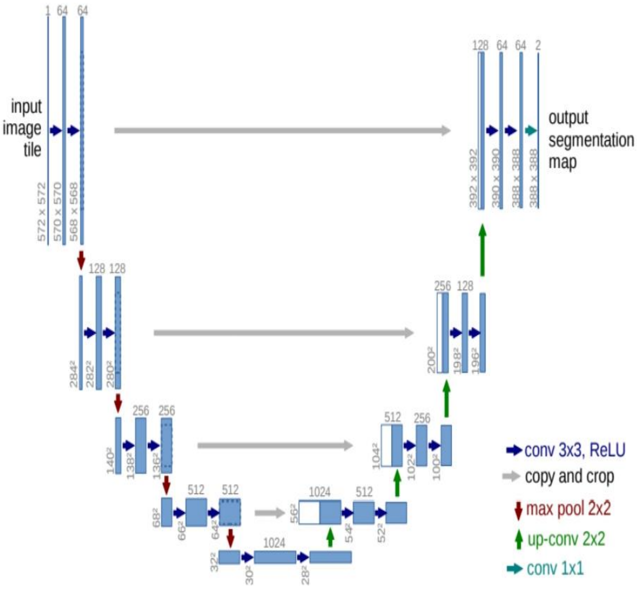


Figure 1 : U-Net Model Architecture

B. *U-Net 16*

Our fundamental network architecture draws inspiration from the well-established U-Net structure, renowned for its efficacy in medical image semantic segmentation tasks and numerous advantages. Comprising an encoder and a decoder, the network follows a two-stage process: feature extraction by the encoder and segmentation mask prediction by the decoder. Skip connections between the encoder and decoder play a pivotal role by amalgamating low-level feature maps with their higher-level counterparts, thereby enhancing the precision of segmentation outcomes. To enhance segmentation performance, we introduce modifications to the basic U-Net architecture by integrating VGG-16 as the backbone network, resulting in a structure termed U-Net16. Specifically, we eliminate the last max-pooling layer, along with all subsequent fully connected and softmax layers from the VGG net. Consequently, the remaining convolutional layers form the encoder portion of the network. For constructing the decoder, we employ transposed convolution layers. At each layer of the decoder, the output is concatenated with the output of the corresponding convolutional layer in the encoder. Subsequently, we implement five upsampling operations using convolution layers to ensure the desired channel number for the output feature maps.

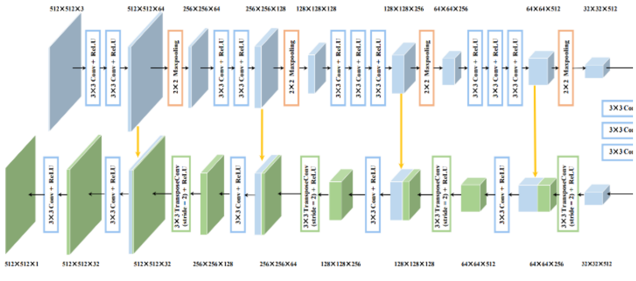


Fig. 2: The network architecture of U-Net-16, using a VGG-16 as backbone.

Figure 2: Model Architecture of U-Net 16

III. Use of Segmentation-models library

Segmentation models library encompasses a comprehensive suite of tools, algorithms, and pre-trained models specifically tailored for semantic segmentation tasks. This section delves into the significance of the library and highlights its key components.

The Segmentation Models library is a powerful Python tool designed for image segmentation tasks using neural networks. It leverages the capabilities of Keras and TensorFlow frameworks to provide a high-level API for building sophisticated segmentation models with minimal code complexity. This note highlights the key features and usage of this library.

A. *Key Features*

- High-Level API: With just a few lines of code, users can create complex segmentation models, including popular architectures like U-Net, FPN, Linknet, and PSPNet.
- Model Architectures: The library offers 4 distinct model architectures optimized for both binary and multi-class image segmentation, each supporting up to 25 different backbone networks.
- Pre-trained Weights: All available backbones come with pre-trained weights, enabling faster convergence and improved performance when training models
- Losses and Metrics: The library includes a variety of segmentation-specific loss functions (e.g., Jaccard, Dice, Focal) and evaluation metrics (e.g., IoU, F-score) to facilitate model training and evaluation.

B. *Models and Backbones*

The library supports various segmentation models:

- FPN (Feature Pyramid Network)
- Linknet
- PSPNet (Pyramid Scene Parsing Network)
- U-Net

C. *Available backbone networks encompass a range of options:*

- MobileNet
- ResNet
- VGG (VGG16, VGG19)
- EfficientNet
- Inception
- SE-ResNet
- SENet154
- ResNeXt

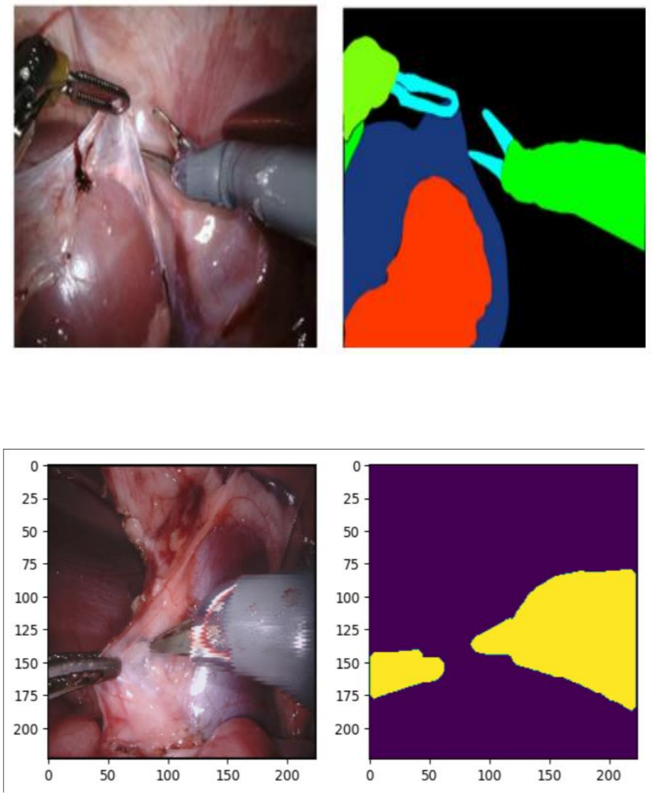


Figure 3 : Sample from the dataset

IV.Dataset Used

In this project, Dataset used is from Robotic Scene Segmentation Sub-Challenge of Endovis.

1) Data Collection: The dataset comprised 19 videos of surgeries, which were divided into 15 training sets and 4 test sets. Each sequence stemmed from a singular porcine training procedure recorded on da Vinci X or Xi systems, utilizing specialized hardware. The frames underwent manual filtering until each sequence attained a total of 300 frames.

2) Data Annotation: The data was annotated by a team of trained professionals using specialized software, which generated boundaries around each semantic class. Quality control measures were implemented throughout the annotation process. Annotating anatomical data requires following a complex protocol to address ambiguities and ensure consistent labeling while adhering to higher-level objectives for system development. Additionally, a challenge is presented by the difficulty or impossibility of identifying structures in many camera views. Instead, an extended sequence of images capturing the structure from various angles and distances needs to be reviewed by annotators. Many annotation tools operate on individual images without seamlessly integrating the viewing of video sequences into the workflow, thereby complicating the process.

V.Metrics

The following metrics were used to evaluate the performance of our models.

1. IOU (intersection over union):

In computer vision, IOU stands for Intersection over Union, and it’s a common evaluation metric used to assess the accuracy of object detection and image segmentation algorithms. IOU measures the overlap between the predicted region and the ground truth region, helping to quantify how well an algorithm’s output aligns with the actual object or region in the image. IOU is also known as the Jaccard index and is calculated using the following formula:

$$IOU = \frac{Area\ of\ Inter\ sec\ tion}{Area\ of\ Union} \quad (1)$$

“Area of Intersection” refers to the number of overlapping elements (pixels or regions) between the predicted segmentation and the ground truth segmentation. It represents the area where both segmentations agree. “Area of Union” represents the total number of elements (pixels or regions) encompassed by both the predicted segmentation and the ground truth segmentation. The IOU metric provides a value between 0 and 1, where: A score of 0 indicates no overlap or agreement between the predicted and ground truth regions, meaning there is no common area between them. A score of 1 indicates perfect overlap, where the predicted region is identical to the ground truth.

2. F1 score:

The F1 score, a key evaluation metric in machine learning, combines precision and recall to provide a comprehensive assessment of a model's performance. By accounting for both false positives and false negatives, it offers a balanced measure. Conversely, the accuracy metric calculates the proportion of correct predictions made by a model across the entire dataset. However, accuracy may be misleading in cases of imbalanced datasets, where classes have varying sample sizes, as it can skew towards the majority class, yielding an inaccurate evaluation of the model's true capability. Precision and recall serve as crucial evaluation metrics in machine learning, particularly for classification tasks.

Precision measures the accuracy of positive predictions made by the model, while

recall quantifies the model's ability to correctly identify actual positive cases.

The F1 score is a single metric that combines both precision and recall into a harmonic mean. Maximizing the F1 score implies optimizing for both high precision and high recall simultaneously. This balanced approach makes the F1 score a popular choice among researchers for evaluating their models, often used in conjunction with accuracy.

A confusion matrix serves as a valuable tool for evaluating the performance of a classification model. It presents the model's predictions in a tabular format, depicting both correct and incorrect predictions for each class. The rows of the matrix represent the actual classes, while the columns represent the predicted classes. For a binary classification scenario involving classes labeled as "positive" and "negative," the confusion matrix comprises four key components:

a) True Positives (TP): The count of samples accurately predicted as "positive."

b) False Positives (FP): The count of samples incorrectly predicted as "positive."

c) True Negatives (TN): The count of samples accurately predicted as "negative."

d) False Negatives (FN): The count of samples incorrectly predicted as "negative."

$$\Pr e\,cision = \frac{TP}{TP + FP} \qquad (2)$$

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

$$F1Score = \frac{2 \cdot \Pr e\,cision \cdot Recall}{Recall + \Pr e\,cision} \quad (4)$$

3. Dice Coefficient

The Dice coefficient serves as a prevalent similarity metric utilized across various domains such as image segmentation and natural language processing, among others. It quantifies the similarity between two sets, denoted as A and B. This coefficient varies between 0 and 1, with 1 signifying complete identity between the sets and 0 indicating no overlap between them.It is defined as:

$$DiceCoefficient = \frac{2 \cdot |A \cap B|}{|A| + |B|} \quad (5)$$

Where |A| represents the number of elements in set A, and |B| represents the number of elements in set B. |A ∩ B| represents the number of elements that are present in both sets.

VI.Losses

a) Jaccard-distance loss : The Jaccard distance, also known as the Jaccard similarity coefficient, is a measure of similarity between two sets. It’s defined as the size of the intersection divided by the size of the union of the sets. The Jaccard distance between two sets A and B can be calculated using the following formula:

$$J = 1 - \frac{|A \cap B|}{|A| + |B|} \qquad (6)$$

Where |A ∩ B| represents the size (cardinality) of the intersection of sets A and B. |A ∩ B| represents the size (cardinality) of the union of sets A and B. The Jaccard distance ranges from 0 to 1, with 0 indicating that the sets are identical and 1 indicating that the sets have no elements in common. Now, in the context of machine learning and specifically for loss functions, the Jaccard distance is often used as a similarity metric for tasks like image segmentation or object detection. In these tasks, the Jaccard distance is used to compare the overlap between predicted and ground truth regions. The Jaccard distance can be transformed into a loss function known as the Jaccard loss or Jaccard distance loss. This loss function penalizes predictions that have low overlap with the ground truth regions. It’s often used in tasks where the classes are imbalanced or where accurately predicting the spatial extent of objects is important. The Jaccard loss is usually defined as 1 minus the Jaccard similarity coefficient, so it’s effectively minimizing the dissimilarity between predicted and ground truth regions.

This is expressed as:

$$JaccardLoss = 1 - \frac{|A \cap B|}{|A| + |B|} \qquad (7)$$

In the context of neural network training, the Jaccard loss is commonly used as a loss function along with other metrics such as Intersection over Union (IoU) to optimize models for tasks like semantic segmentation or object localization.

b)Categorical Cross entropy :
Categorical Cross-Entropy (also known as Softmax Cross-Entropy) is a commonly used loss function in classification tasks, particularly when dealing with multiple classes. It measures the dissimilarity between the true distribution of the data and the predicted probability distribution. Here’s how it’s calculated: Suppose you have N classes, and for each input sample, the model predicts a probability distribution over these classes. Let’s denote the predicted probabilities as yˆi where i=1,2,...,N, and the true labels as yi , where yi is 1 if the sample belongs to class i, and 0 otherwise. The categorical cross-entropy loss L for a single sample is calculated as:

$$L = -\sum_{i=0}^{n} y_i \cdot \log(y_{\tilde{i}}) \qquad (8)$$

This formula calculates the cross-entropy loss for each class and sums them up. Essentially, it measures how well the predicted probabilities match the true distribution of the data. The true distribution is represented by the one-hot encoded vector y, where only one entry is 1 (indicating the true class) and the rest are 0. Here’s a breakdown of the components of the formula:

yi : if 1 then belongs to positive class else negative
yˆi : The predicted probability of class i by the model .

The utilization of the negative sign in cross-entropy loss stems from its equivalence to maximizing the likelihood of the model's predictions aligning with the true distribution. Put simply, categorical cross-entropy loss imposes greater penalties on the model for substantial errors in predicting the true class probabilities. Minimization of this loss fosters a model behavior that assigns high probabilities to correct classes and low probabilities to incorrect ones.

VII.Results and Conclusions

U-Net 16 Model Performance

Metric	***Train data***	***Validation data***
Loss	0.0125	0.0168
Dice Coefficient	0.9623	0.9434
Accuracy	0.9876	0.9833

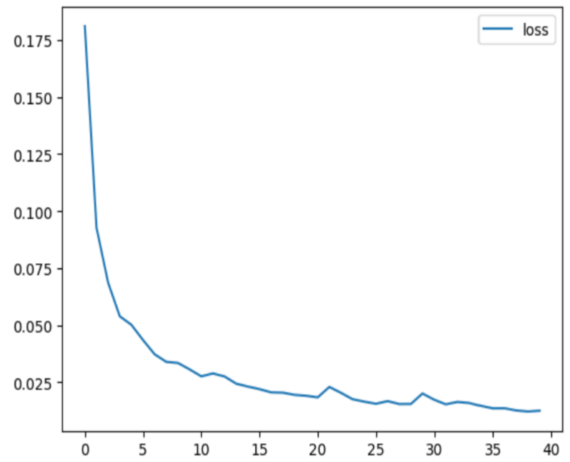


Figure 4 : Loss vs Training Epochs for U-Net

U-Net 16 Model Performance

Metric	***Train data***	***Validation data***
Loss	0.4920	0.4914
IOU	0.2046	0.2068
F1-Score	0.2262	0.2272

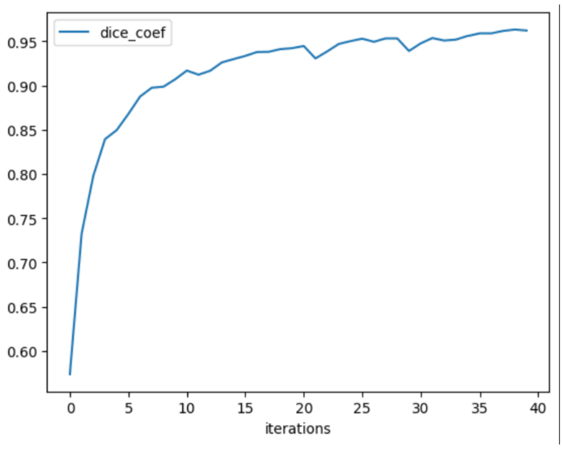
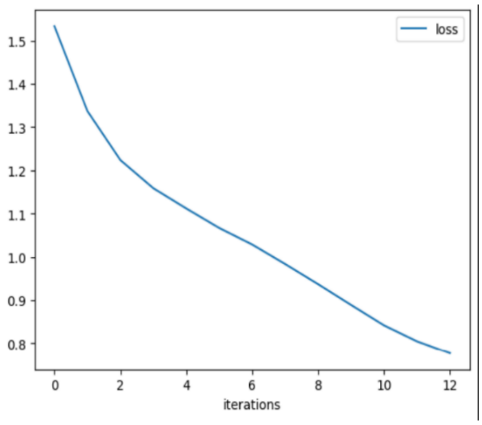


Figure 5 : Dice Coefficient vs Training Epochs for U-Net

Figure 6: Loss vs Training Epochs for U-Net 16



Prediction of U-Net-16 on test images

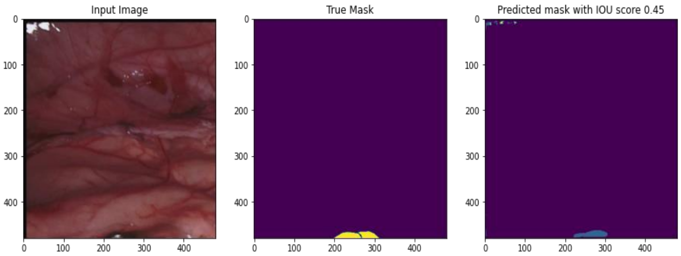


Figure 7 : Test Image 1 for U-Net16

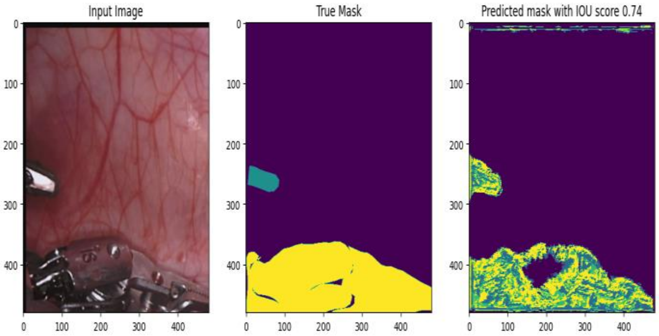


Figure 8 : Test Image 2 for U-Net 16

Prediction of U-Net on test images

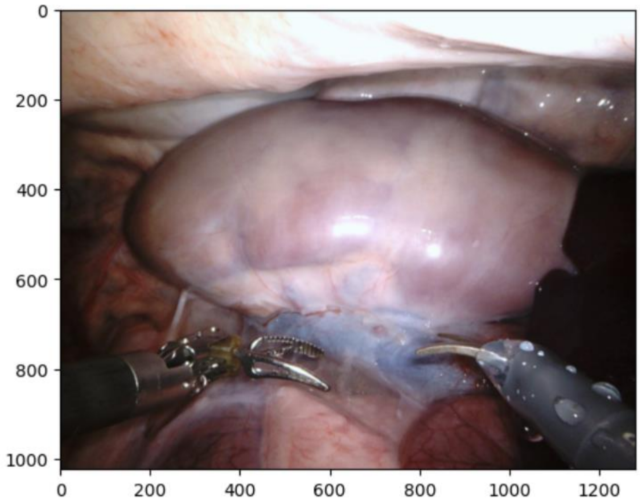
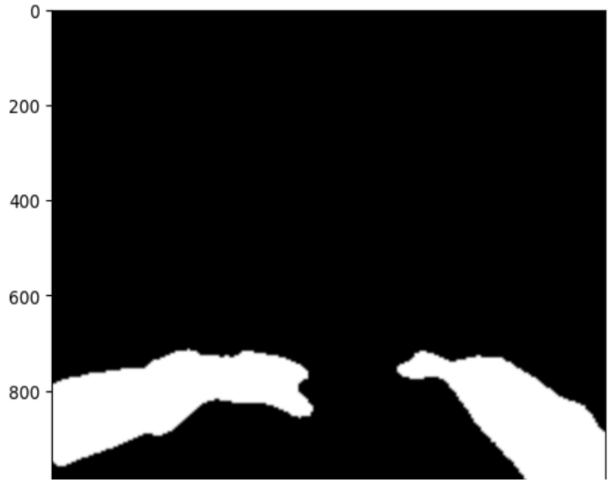


Figure 8:Sample test image for U-Net



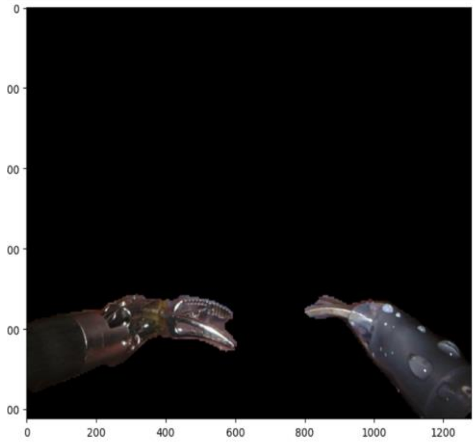


Figure 9 : Output predictions of U-Net

As seen from the above graphs for loss against the number of epochs the model is able to converge and able to identify the surgical tools which are present in the surgical frame with adequate accuracy. Among the two models that were used for performing the segmentation task at hand it has been found that the U-Net model is able to perform better .This observation may be attributed to the fact that when we use VGG-16 as a backbone it comes with pretrained weights that might be derived by training the model on image data different from the one used in our work. With limited resources the training of VGG-16 backbone was not viable option for our work hence we can say that better performance may be obtained if we train the entire U-Net16 on the given dataset

VIII. Refrences:

1) *https://www.researchgate.net/publication/352761398_2018_Robotic_Scene_Segmentation_Challenge*
2) *https://arxiv.org/abs/1505.04597(U-Net: Convolutional Networks for Biomedical Image Segmentation)*
3) *https://www.researchgate.net/figure/The-network-architecture-of-U-Net-16-using-a-VGG-16-as-backbone_fig2_347257459*