Laparoscopic Surgical Image Segmentation - CS231N Final Project Report

Brian Sutjiadi, MD
Stanford University
bsutjiad@stanford.edu

Abstract

In this project, I apply supervised learning methods to perform anatomical semantic segmentation of images taken from laparoscopic surgery. My primary aim will be to develop a model that can perform single class segmentation and identify the gallbladder in images taken from gallbladder surgery. In particular, I implement a U-Net model, perform hyperparameter tuning, and evaluate the performance of the model against pre-trained/fine- tuned models in order to assess whether transfer learning is an effect approach at addressing the data shortage issue that plagues biomedical and surgical domains.

1. Introduction

Surgery has traditionally been practiced with the "open" technique. In open surgery, the tissues and organs requiring surgical therapy are directly visualized and manipulated. However, technological advancements have enabled the development of minimally invasive surgery. Minimally invasive surgery is done by inflating a body cavity, such as the abdomen, with CO_2 to create working room, making small 5-12 mm incisions, and placing ports into the body cavity of interest. A camera is then inserted through one of these ports to visualize tissues requiring operative therapy. The operation is then entirely carried out under indirect visualization, where the surgeon uses long instruments inserted through the ports, and visualizes/supervises the movement of those instruments through a live-streamed video feed. In many surgical specialties, minimally invasive techniques have replaced traditional open techniques as the standard of care. Consequently, there is an abundance of image/video data that is generated throughout the course of everyday surgical practice.

Being able to perform safe and effective surgery requires that surgeons are able to recognize the visualized anatomy, and map the things that they see to their own mental model of what the anatomy should be like. In other words, surgeons must perform semantic segmentation of the things they see in order to perform safe surgery. Surgical mistakes/errors (eg. cutting the wrong structure) are often the result of misclassification during this process of semantic segmentation. Thus, semantic segmentation is of critical importance to the task of performing surgery.

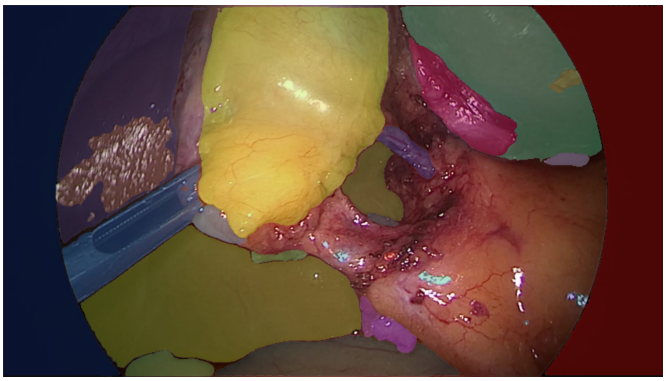
Anatomical segmentation can be particularly challenging in cases of high disease severity. It is well known that higher severity cases are associated with higher rates of complications, a fact that is at least partially due to difficulty in performing anatomical segmentation in complicated cases. Having an automated system to aid in anatomical semantic segmentation could prove to be incredibly useful to surgeons managing patients with severe disease, and could help to reduce complication rates and improve surgical safety.

Minimally invasive gallbladder surgery (laparoscopic cholecystectomy) is one of the most commonly performed procedures in the United States. As such, data from laparoscopic cholecystectomies is the most available for use in the development of computer vision models. While the gallbladder operation involves fairly complex and intricate anatomy, the gallbladder itself is usually visually distinct and it is usually the first thing that medical students and surgical trainees learn to distinguish. As such, it is a ripe target for segmentation model development.

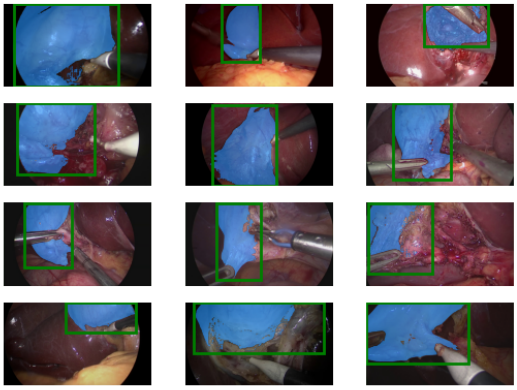
2. Related Work

Muscagni *et al*. were the first to publish on the application of supervised learning techniques to the problem of semantic segmentation of laparoscopic surgical images [6]. They approached this problem by using a pre-trained DeepLabv3+ model, and fine tuning it with expert labeled surgical images. The DeepLabv3+ model applies an Xception encoder to an atrous convolution along with a decoder network to perform semantic segmentation [1, 2].

Surgical images and videos are relatively scarce sources of data for training machine learning models, due to data privacy requirements and other restrictions due to its nature as healthcare data. Furthermore, annotated surgical image data is even more scarce, as expert surgeons and clinicians are the only ones with the training to interpret this data; and unlike other medical domains such as radiology and pathol-



(a) SAM2 Automatic Mask Generation



(b) SAM2 Mask with Box-Prompt

Figure 1: SAM2 Semantic Segmentation Output. Figure 1a shows the output of SAM2 automatic segmentation. The light yellow in the upper middle portion of the image represents the gallbladder. This qualitatively appears to be a reasonable segmentation, although it can be improved. This current segmentation misses a sliver of the gallbladder immediately to the left of the yellow segmentation mask. Figure 1b shows reasonable appearing segmentation masks with a given bounding box prompt.

ogy, useful annotations are not routinely generated as part of the surgeons workflow. This paucity of data has made model development in the surgical image/video domain particularly challenging, and is likely a main contributor toward stifled innovation and progress in this domain.

In a prior class (CS 131), I attempted to approach this problem with an unsupervised, edge detection-based segmentation method. Unsupervised methods are desirable because they don't require the curation of expert labeled data for model development, and because they are often efficient to run and don't require an expensive training process. Specifically, I used Farid edge detection, applied a graph-search method to link pixels within contiguous "islands" delimited by edge- pixels, "within-island" color averaging, and k-means clustering to create segmentation masks. However, as will be shown in section 6, this approach did not work well, likely related to the fact that edges are too non-specific of a feature in isolation to be useful for semantic segmentation.

Transfer learning is a well-described approach to overcoming challenges in developing robust models with limited annotated data. However, successful transfer learning depends on the existence of similarities between domains [14]. I hypothesize that images from laparoscopic surgery are far enough out of distribution from the training data that was used to train the DeepLabv3+ model, and consequently, that the transfer learning process doesn't contribute significantly to model performance. In this project, I implement a U-Net model and train it using the same data that Muscagni *et al*. used, and evaluate the performance of this model trained from scratch.

3. Data

The Endoscapes dataset [7] is a publicly available dataset that consists of selected frames taken laparoscopic cholecystectomy videos. A total of 201 videos were obtained from surgeries performed at a single hospital in France from 2015 to 2019. These videos were then temporally segmented according to the phase of surgery, and frames from the dissection phase of the operation were sampled and annotated by surgeons. Specifically, the frames were annotated with bounding boxes around anatomical structures (gallbladder, cystic artery, cystic duct, hepatocystic triangle, cystic plate) and surgical tools. Segmentation masks were also manually annotated. These images were then divided by the dataset authors into an 80%-20% train-test split. The training set was then further split 75%-15% into training and validation splits.

Unfortunately, upon reviewing the available annotations, the provided segmentation masks were not usable. The annotations are provided in COCO format, which specifies that segmentation masks can either be encoded as polygon coordinates or as run-length encoding. However, it appears as if the segmentation masks are in some kind of binary format that is not described in any of the associated dataset publications or documentation [7].

3.1. Dataset Pre-processing

This project focuses on single-class case of identification of the gallbladder within surgical images. However, the dataset also includes images that do not contain a gallbladder. Because I was not sure how negative examples would affect model training and evaluation, I made the decision to remove all images that did not contain a gallbladder from the dataset.

Due to the aforementioned issues with utilizing the provided segmentation masks, I needed to generate my own segmentation masks to utilize as ground truth labels for model training and evaluation. Given the time constraints of the course and significant effort that would have been involved, I decided not to manually annotate the images in the dataset. Instead, I attempted to generate segmentation masks by utilizing the Segment Anything Model 2 (SAM2) [9]. Initially, I was concerned that the SAM2 model would perform poorly because surgical images may be out of distribution for these off-the-shelf models. However, after applying SAM2 to a test image (Fig. 1a), the resulting segmentation appeared to be reasonable.

Having reassurance that the SAM2 segmentations would be usable, I proceeded to generate masks for the remainder of the dataset. Fortunately, the model was able to accept prompts in the form of bounding boxes, in which case it would identify one type of object located within the given box prompt. Using this feature, I used the dataset's bounding box annotations as prompts to the model. The resulting segmentation masks largely appeared very reasonable (Fig. 1b). However, I noticed that there were a few images where the mask was incorrect (eg. the model created a mask corresponding to an instrument or the blank letterbox instead of identifying the pertinent anatomy). As a result, I manually reviewed all of the generated masks and discarded the clearly incorrect segmentation masks as well as a few subjectively low-quality masks.

The original data set consisted of a total of 1933 images, divided into 1212 (62.7%) training images, 409 (21.2%) validation images, and 312 (16.1%) test images. After removal of the images without a gallbladder, there were 1174 (64.6%) training images, 357 (19.7%) validation images, and 285 (15.7%) test images. Finally, after removing images which had erroneous segmentation masks, the dataset consisted of 1098 training images (65.0%), 335 validation images (19.8%), and 256 test images (15.2%).

4. Methods

As mentioned in section 2, the goal of this project is to assess whether transfer learning is applicable to a surgical context, given the possibility that surgical images are out-of-distribution when compared to the standard images used to train computer vision models. In order to answer this question, I implement a modified U-Net model [10], train it only on labeled data, evaluate its performance, and compare this model's performance to the performance of the published model that utilizes transfer learning.

The U-Net model has become ubiquitous in solving semantic segmentation problems in biomedical contexts because it effectively and efficiently learns from relatively small datasets [10]. Because of how effectively it learns from small datasets, the architecture has found broad application to problems in radiology and pathology [13, 5]. However, it has not yet been applied to surgical images.

I started my evaluation by implementing the U-Net model. I made a few modifications to the published U-Net architecture. First, I included padding with each of the convolutional layers so that the image dimensions remained the same throughout the convolutions. Next, I added batch normalization between the convolutional layers to make the training process less sensitive to initialization [8]. I also deviated from the original U-Net paper with respect to training. Instead of utilizing stochastic gradient descent (SGD), I utilized the Adam optimizer due to prior experiences in the assignments which suggested that Adam may converge faster than SGD. Similar to the original description, I utilized binary cross-entropy loss as the loss function.

The U-Net paper also described data augmentation as a way to improve performance with limited data. Consequently, I also performed data augmentation by adding random transformations of the original data (rotations and flipping) to the training dataset. I selected these augmentations as I felt that they would capture the majority of the frame-to-frame variation of the gallbladder appearance that would naturally occur during the course of surgery. I then normalized the input dataset prior to training. For validation and testing, normalization without augmentation was applied to the images.

For quantitative model evaluation, the main metric I used was Intersection Over Union (IOU) to compare the similarity between the predicted and ground-truth segmentation maps. The IOU serves as a more realistic representation of segmentation performance as opposed to pixel-level prediction accuracy, as some of the input images may have class imbalances which can lead to situations where the objective metric is very good while the qualitative alignment of the predicted segmentation is poor.

Having established the code base needed for model training and evaluation, I then turned my attention to hyperparameter tuning. The main hyperparameters I sought to optimize were the learning rate, batch size, and number of epochs. I assessed the effects of various hyperparameter changes by observing the changes to training and validation set losses. Upon discovering the set of hyperparameters that produced the lowest validation set loss, I then saved the model and applied it to the test set.

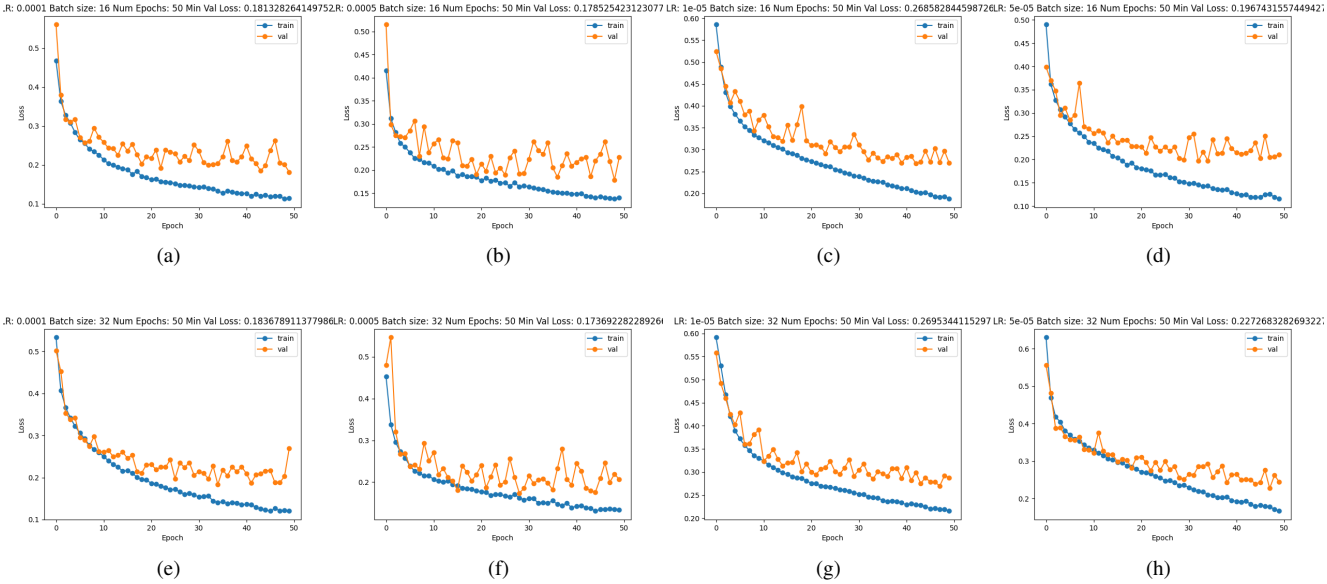


Figure 2: Evaluating Relationship between Batch Size and Learning Rate. Hyperparameters along with the minimum validation set loss for each trial are listed above the graph. Overall, batch size of 16 vs 32 does not appear to have a significant effect on the minimum validation loss.

5. Experiments

Due to literature suggesting a relationship between learning rate and training batch size in terms of model performance [11, 3, 4], I began hyperparameter tuning by training the model with selected combinations of learning rates and batch sizes. After investigating this relationship, I then ran trials to determine the best learning rate. Finally, I evaluated the effect of various batch sizes while holding the learning rate constant.

5.1. Evaluating the Interplay between Batch Size and Learning Rate

I started hyperparameter tuning by running training trials with selected combinations of learning rates and batch sizes Fig. 2. I found that batch sizes of 16 and 32 did not make a significant difference in terms of the minimum validation loss achieved throughout the trial. However, I noticed that with the learning rates $< 5\text{e}{-4}$ (Figs. 2b-2d and 2f-2h), the training loss appears to be continuing to decrease at the 50th epoch, which suggested that more training epochs could potentially lead to greater performance. In all of the trials, it appears as if overfitting starts to become apparent around the 30th epoch. Additionally, I noticed that trials with larger batch sizes may be overfitting less than trials with smaller batch sizes, although if this effect is truly present, it is small.

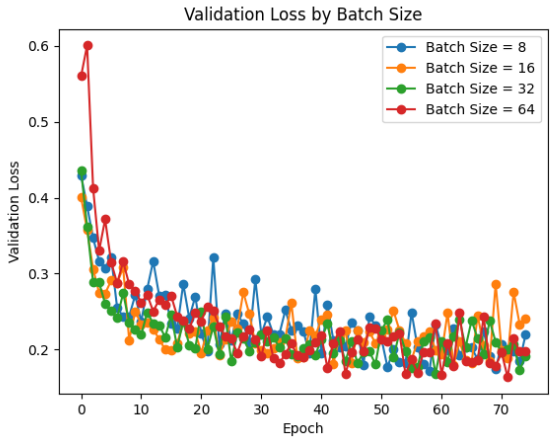


Figure 3: Effect of Batch Size.

5.2. Searching for the Optimal Learning Rate

Using the insights from the experiment in section 5.1, I increased the number of epochs per trial from 50 to 75. Also from previous experiments (not shown), I knew that the optimal learning rate was somewhere in the range between $1\text{e}{-3}$ and $1\text{e}{-5}$. Thus, I held the batch size constant at 16, and searched through learning rates within this range. Specifically, I tried learning rates of $3\text{e}{-4}, 2\text{e}{-4}, 1\text{e}{-4}, 9\text{e}{-5}, 8\text{e}{-5}, 7\text{e}{-5}$ in one experiment,

(a) Input Image (b) U-Net Segmentation Mask (c) Bounding box-based Segmentation Mask

(d) Input Image (e) U-Net Segmentation Mask (f) Bounding box-based Segmentation Mask

Figure 4: Qualitative Segmentation Evaluation. Figures 4a-4c represent examples of qualitatively good segmentations, whereas Figures 4d-4f represent examples of qualitatively poor segmentations

and $2.5\text{e}{-4}, 2.25\text{e}{-4}, 2\text{e}{-4}, 1.75\text{e}{-5}$ in the next experiment. The learning rate with the lowest validation set loss was $2.25\text{e}{-4}$.

5.3. Assessing the Effect of Batch Size

While the experiments in section 5.1 suggested that batch size has a negligible effect on model performance, I hypothesized that I was coming to this conclusion because I did not test a broad enough range of batch sizes. Consequently, I decided to revisit this question by training models with batch sizes of 8, 16, 32, and 64. I held the learning rate fixed at $2.25\text{e}{-4}$, and I trained the model for 75 epochs. As can be seen in Figure 3, I still was not able to observe an effect of batch size on model performance.

6. Conclusion and Discussion

In this project, I implemented a modified U-Net model, trained it to perform semantic segmentation on laparoscopic surgical images, and compared its performance with an unsupervised edge detection-based method as well as a pretrained/fine-tuned DeepLabv3+ model. The performance of each of the approaches, as assessed by the average IOU across the test set, is shown in Table 1.

From these results, it is apparent that the unsupervised edge detection-based approach performs poorly. While analysis of the unsupervised method is outside the scope of this project, it is insightful at suggesting that the supervised learning architectures (such as the U-Net architecture explored in this project) are capturing more general features and emergent properties rather than just being simple edge detectors. Also, the failure of this approach also demonstrates that effective semantic segmentation relies on understanding these higher level features.

On the other hand, while the U-Net architecture is capable of delivering impressive classification performance in selected examples (Fig. 4b), training a model from scratch is still unable to match the performance of the pretrained/fine-tuned model as presented by Mascagni *et al.* [6]. There are several possible reasons this may be the case.

With this project specifically, as described in section 3, I was unable to utilize the true ground truth labels provided by the dataset authors. Consequently, I was required to generate my own ground truth segmentation maps. While the segmentation masks generated by SAM2 were mostly reasonable (Fig. 1), this was not the case for all input images (Fig. 4c). Additionally, in some cases (eg. Fig. 4a), the U-Net segmentation mask appears qualitatively more accurate than the generated ground truth label. The key here is that perhaps the average IOU obtained by the U-Net model as seen in Tab. 1 is an underestimate due to faulty ground truth labels.

The U-Net architecture crucially relies on data augmentation to make efficient use of small datasets. In my implementation, I only implemented fairly minimal data augmentation strategies and was unable to revisit this aspect of the implementation due to time constraints. However,

Method	Average IOU
Edge Detection	0.358
U-Net	0.661
DeepLabv3+	0.885

Table 1: Average IOU.

in reviewing images in the dataset, it seems apparent that many of the depictions of gallbladders are fairly similar to one another, and differ mostly as a result of the ways that the surgeon's instruments are manipulating the organ and by parts of the organ that have been altered during the course of the surgery. While simple shape and color transformations/augmentations likely won't be able to create new data mimicking dissected tissue, I believe stretching and rotational transformations could reasonably mimic different appearances of the organ resulting from surgeon retraction. Thus, including these transformations as part of the augmentation strategy could potentially further increase the performance of the U-Net classifier.

Another possible reason why the modified U-Net implementation performs worse is that the DeepLabv3+ model utilizes transfer learning and model fine tuning instead of training from scratch. While I initially hypothesized that transfer learning would not work well for surgical images, given the difference in distribution that I thought was present between surgical images and standard computer vision dataset images, it seems that transfer learning may actually prove to be quite useful in solving problems related to surgery and biomedicine.

With respect to other take-aways and learning points from this project, I learned that hardware limitations are a significant barrier to taking on even a simple project such as this one. When attempting to investigate the effect of batch sizes on model performance, as suggested by Hoffer *et al*. and Keskar *et al*. [3, 4], I ran into GPU out-of-memory (OOM) errors I could not work around. The largest batch size I could run trials on was 64, and as seen in Figure 3, there was no obvious difference in the model performance at these batch sizes. The batch sizes that Hoffer *et al*. and Keskar *et al*. work with in their papers are on the order of 500-2000. Granted, they utilized the CIFAR and MNIST datasets, which are tiny images compared to the images I work with in this project, but the point still stands that hardware is a significant barrier to creating robust and high-performing models. Perhaps another reason why the U-Net model has poorer performance than the DeepLabv3+ model is that I simply did not have the computing resources to train a higher performing model.

References

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.

[2] F. Chollet. Xception: Deep learning with depthwise separable convolutions, 2017.

[3] E. Hoffer, I. Hubara, and D. Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks, 2018.

[4] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *CoRR*, abs/1609.04836, 2016.

[5] J. Li, K. Sarma, K. C. Ho, A. Gertych, B. S. Knudsen, and C. W. Arnold. A multi-scale u-net for semantic segmentation of histological images from radical prostatectomies. *AMIA Annu Symp Proc*, pages 1140–1148, 2018.

[6] Mascagni, Vardazaryan, Alapatt, Urade, Emre, Fiorillo, Pessaux, Mutter, Marescaux, Costamagna, Dallemagne, and Padoy. Automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Annals of Surgery*, 275(5):955–961, 2022.

[7] P. Mascagni, D. Alapatt, A. Murali, A. Vardazaryan, A. Garcia, N. Okamoto, G. Costamagna, D. Mutter, J. Marescaux, B. Dallemagne, and N. Padoy. Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. *Scientific Data*, 12(1), Feb. 2025.

[8] A. Persson. Machine learning collection. Available online at: `https://github.com/aladdinpersson/Machine-Learning-Collection`, last accessed on 06.04.2025.

[9] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[10] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[11] S. L. Smith, P. Kindermans, and Q. V. Le. Don't decay the learning rate, increase the batch size. *CoRR*, abs/1711.00489, 2017.

[12] H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath. Semantic segmentation using vision transformers: A survey, 2023.

[13] J. Walsh, A. Othmani, M. Jain, and S. Dev. Using u-net network for efficient brain tumor segmentation in mri images. *Healthcare Analytics*, 2:100098, 2022.

[14] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *CoRR*, abs/1911.02685, 2019.