Semantic Segmentation of Endoscopic Surgical Scenes using Encoder-Decoder Architectures

Douglas Summerlin
Clark School of Engineering, University of Maryland, College Park, MD, USA
Contact: dsummerl@umd.edu, phone +1-240-205-3707

Abstract— **Accurate localization and tracking of surgical instruments remains a key challenge for enabling robotic systems to perform precise, tissue level interventions. Due to the advent of endoscopic robotic surgical systems, interoperative camera footage has been explored as an avenue for training deep-learning based semantic segmentation models to identify surgical instruments within the surgical scene. This study investigates the semantic segmentation performance of three encoder-decoder based deep learning architectures: Basic U-Net, SwinUNETR, and a custom implementation called VGG16-AttnUNet. The results generated from the trained models seemingly indicate that the introduction of attention gates, pre-trained convolutional encoders, and vision transformers make negligible improvements in segmentation accuracy while suffering significantly in inference time. Additionally, the results generated from these models interestingly seem to outperform results reported by the recent literature by as much as 1.80%, 3.64%, and 7.11% in binary, part, and instrument type segmentation respectively, indicating potential issues with study design that may warrant investigation and revalidation of the techniques employed in this work.**

Keywords—Robotic Surgery, Deep Learning, Computer Vision, Minimally Invasive Surgery, Semantic Segmentation

I. Introduction

Robotic Minimally Assisted Surgery (RMIS) has seen increased exploration and usage in recent years due to the ability of these systems to perform more precise and difficult procedures while minimizing recovery time and tissue damage for the patient. Robotic surgical systems often rely on endoscopes for visualization of the surgical scene beyond the incision site, which presents an opportunity to use computer vision techniques and artificial intelligence to obtain useful feedback for onboard autonomous systems and clinical operators. For instance, understanding the orientation of instruments equipment within the surgical scene remains a fundamental problem of RMIS, as localization errors can present a significant hazard when interacting with high-risk anatomical structures such as vasculature or nervous system tissue. In addition to navigation assistance, the implementation of such algorithms in RMIS vision systems can be used to mask augmented reality overlays to prevent obstruction of renders from the surgical instruments if the segmentation model can make predictions with low temporal latency, as demonstrated in Fig. 1 below.

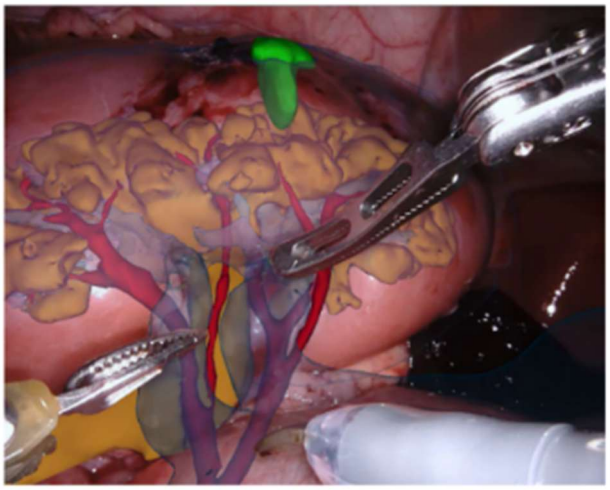


Fig. 1 From the 2017 MICCAI challenge of masking an instrument so that an augmented reality overlay does not occlude the surgeon's view [1].

Effective localization of surgical instruments requires the identification of instruments and their constituent parts the background tissue in the image using a camera or another remote sensing apparatus. With the advent of deep learning, automated semantic segmentation algorithms have been developed as a method to address these challenges while potentially providing real-time performance speed.

Since its inception in 2015, the U-Net network [9] and its many variants have demonstrated proficiency in performing semantic image segmentation due to the ability of encoder-decoder frameworks to capture low-level and high-level features. U-Net combines a contracting convolutional path for preserving global image context and a decoder path that up-samples the derived feature maps to the original image resolution. Additionally, skip connections are implemented on each layer to link encoder phase feature maps to the corresponding decoder path feature maps to preserve fine details. In a systematic review published by Fernandes et al. [3] in 2023 on deep learning applications in laparoscopic tool detection, the authors found that the 3 best performing network types in performing segmentation tasks were either U-Net or autoencoder, both of which similarly utilize an encoder-decoder framework. This review reported scores in the range of 72.3–86.0% for accuracy and 85.63–90.2% for DICE similarity coefficient of the networks attempting binary instrument segmentation, as visualized in Fig. 2 below.

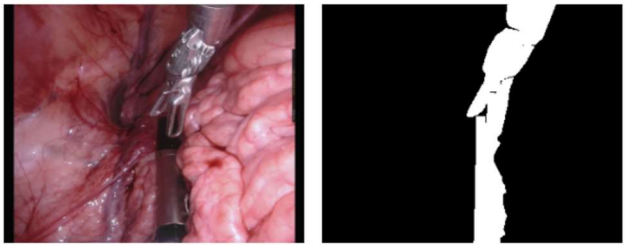


Fig. 2 Exemplar case of binary instrument segmentation of surgical instruments.

Many researchers have altered the core architectural framework of the original basic U-Net in attempts to improve its image segmentation performance. Hasan and Linte [5] introduced U-NetPlus, a modified U-Net architecture with a pre-trained encoder backbone using VGG-11 and VGG-16 with batch normalization to accelerate convergence time and mitigate optimization challenges. U-NetPlus achieved notable segmentation performance on surgical video frames, including a 90.20 DICE score for binary segmentation. They also reported a 76.26% DICE score for instrument part segmentation, which assesses the ability of the network to segment the different mechanical subsystems of the surgical tools in the scene. Sun et al. [10] proposed a novel lightweight encoder-decoder framework that utilizes MobileNetV3 as the encoder backbone with Ghost feature maps [4] to reduce the computational cost associated with traditional convolutional operations. These modifications improved real-time segmentation prediction ability in surgical contexts, with the authors reporting a detection time of 27ms of their model, representing a 2.5x inference speed improvement over the basic U-Net framework.

Xia et al. [11] proposed that a nested U-Net algorithm designed with a U-Net structure at every layer of the network structure to effectively fuse global and local contextual information captured at different scales. They reported improvements ranging from 0.46-1.06% and 3.58-7.25% in the performance of binary and instrument type segmentation, respectively, when compared to reports of previous works. Hayat et al. [8] developed the SEGRSNet which introduces super-resolution techniques and a combined spatial and channel attention block, both of which were introduced to enhance feature maps and sharpen image details of low-resolution endoscopic images. They report DICE performance gains of 0.4% in binary and 7.34% in part segmentation when compared to the work of [11] but suffered from poor performance in the instrument type segmentation prediction modality reporting only 23.79% DICE score.

In recent years, the advent of transformer architectures and their subsequent widespread adoption have led to increased interest in attention-based deep learning methods in computer vision tasks as well. Hatamizadeh et al. [7] proposed the SwinUNETR (Sliding Window U-net Transformer) in 2022 to perform 3D segmentation of brain tumors, applying the potential of Vision Transformers to model long-range dependencies and contextual relationships to medical imaging tasks.

Inspired by these recent works, this study will attempt to explore the impact integrating attention mechanisms into U-net encoder-decoder architectures to assess how attention influences segmentation performance in surgical applications. To objectively assess the improvements of attention-based U-nets against the original basic U-Net framework, a pair of attention-based networks will be developed and trained alongside a Basic U-Net baseline model. One of the attention U-Net models will be a custom implementation that utilizes a pre-trained encoder to incorporate features derived from large datasets and attention gates to selectively focus on areas of interest within the surgical scene. These U-Net models will be comprehensively compared against state-of-the-art methods to assess the extent to which the addition of attention mechanisms can meaningfully improve performance with respect to segmentation accuracy and computational efficiency in the context of Robotic Minimally Assisted Surgery.

II. Methods

A. Dataset Selection

The MICCAI 2017 EndoVis Robotic Instrument Segmentation Challenge [1] dataset will be utilized for training, validation and testing of each of the developed U-Net models. The dataset contains 8 training video sequences, each with 225 frames at 1280×1024 resolution. Additionally, the dataset provides 9 test sequences, including 8 continuations of the footage provided in the training set with 75 additional frames each, as well as one completely novel set with 300 frames maintaining the same 1280×1024 image resolution. All images in the dataset were captured at 1Hz from surgical footage obtained from an Intuitive Surgical da Vinci Xi system performing nephrectomy procedures upon porcine subjects. Each RGB image captured exists as one half of a pair, as the da Vinci Xi captures images using a stereo camera system. Only the left image will be used for training, as only the left image is given a labelled ground truth image from which to calculate training loss.

The ground truth labels include different parts of the surgical instruments, categorized into four groups: shaft, wrist, claspers, and a miscellaneous class for other tools (e.g. ultrasound probes). Labelled ground truth images depict the different segmentation masks as a variety of intensities and are contained within a subfolder indicating their respective instrument types. An example of a surgical scene and the corresponding ground truth label image can be seen in the Fig. 3 below, where the blue, green and red regions represent the shaft, wrist, and clasper classes respectively.

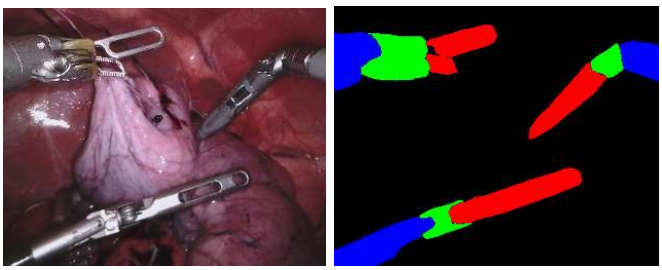


Fig. 3 Example frame and colored segmentation mask from the MICCAI 2017 EndoVis Robotic Instrument Segmentation Challenge Dataset. [1]

The EndoVis 2017 dataset contains 7 different robotic surgical instruments as follows: Large Needle Driver, Prograsp

Forceps, Monopolar Curved Scissors, Cadiere Forceps, Bipolar Forceps, Vessel Sealer and an Ultrasound Probe which is typically held in the jaws of the Prograsp Forceps instrument. The instruments are visualized and labelled in Fig. 4 below.

Fig. 4 Demonstration of the different instrument types in the MICCAI 2017 EndoVis Robotic Instrument Segmentation Challenge Dataset: (a) & (b) Bipolar Forceps (c) Prograsp Forceps (d) Large Needle Driver (e) Vessel Sealer (f) Grasping Retractor (h) Monopolar Curved Scissors (g) Ultrasound probe [1]

B. Data Processing and Augmentation

The ground truth labels provided in the dataset are provided with pixel intensities and a mapping dictionary that matches each intensity to the relevant semantic label. Each instrument type was categorized into its own ground truth image, such that one RGB input image could have multiple ground truth label images if multiple types of surgical images were present in the scene at one time.

To create segmentation masks compatible with the machine learning pipeline developed, new images were generated with composite masks corresponding to the type of segmentation task being performed: binary, part segmentation, or instrument segmentation. Additionally, extra images were generated for combined segmentation of part and instrument types via concatenation and remapping, although to the best of my knowledge no prior studies report exploring this modality of segmentation for this dataset. Combined segmentation involved generating unique labels for each permutation of part and instrument type, i.e. Prograsp Forceps Wrist or Vessel Sealer Clasper. In summary, including the background class, the generated data supports 2 classes for binary segmentation, 5 classes for part segmentation, 8 classes for instrument segmentation, and 21 classes for combined part and instruments segmentation, as not every permutation of part and instrument in the combined segmentation was valid.

To clean the input data and labels, each image was ROI cropped from size 1080x1920 →1024×1280 to remove padding on the perimeter of each image. Although reducing image resolution can impact segmentation performance, each image was downsized to a standardized resolution of 256x320 to ensure dimensional compatibility and computational feasibility. Each of the RGB input images were unit normalized in intensity, and each of the ground truth images were discretized and one-hot encoded to match the number of classes present in the training set. To improve generalizability, the RGB images and their corresponding ground truths were augmented using random horizontal flips and 25% zooms, each applied with 30% probability for any given call of the image during training.

C. Overview of Deep Learning Approach

To systematically examine the effects of incorporating attention mechanisms into segmentation pipelines based on the U-Net, I adopted a three-stage evaluation pipeline designed to replicate reported results from prior studies and critically investigate the impact of the implementation of attention mechanisms on model performance.

The first stage of the evaluation pipeline involved training a Basic U-Net model to serve as a baseline for assessing the impact incorporated attention mechanisms. This model architecture was adapted from the original model proposed by [9] and is illustrated in Fig. 5. The model was initialized with feature channel sizes of 32, 64, 128, 256, and across the successive encoder and decoder layers. The Basic U-net model architecture utilized in this phase was implemented using the Medical Open Network for AI (MONAI) framework [2]. MONAI is an open source library developed for applications of deep learning in medical imaging.

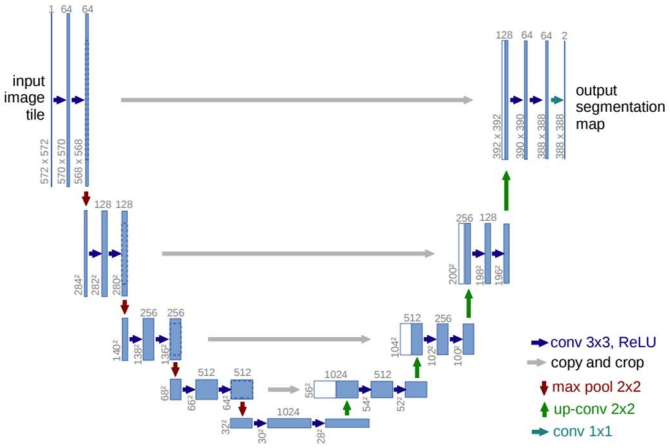


Fig. 5. Basic U-Net Architecture as proposed in [9] utilized in generating the baseline model.

The second stage of the pipeline will train a custom implementation the U-Net. This model, called VGG16-AttnUNet, has been modified to incorporate the VGG-16 CNN architecture as the encoder backbone. The VGG-16 model has been previously trained on large amounts of image data and has been tuned to identify relevant image features, so introducing this as the core convolutional encoder should theoretically reduce the burden of hyperparameter tuning during training and accelerate model convergence. The VGG-16 model is illustrated in Fig. 7. To replace the fully connected layer at the end of the VGG-16 encoder, a 2-layer convolutional bottleneck was introduced to derive finely detailed features before up-sampling the features into the decoder phase. In addition, the second model introduces a simple attention mechanism to the skip connections to learn to identify important scale-invariant features while suppressing the impact of less useful features.

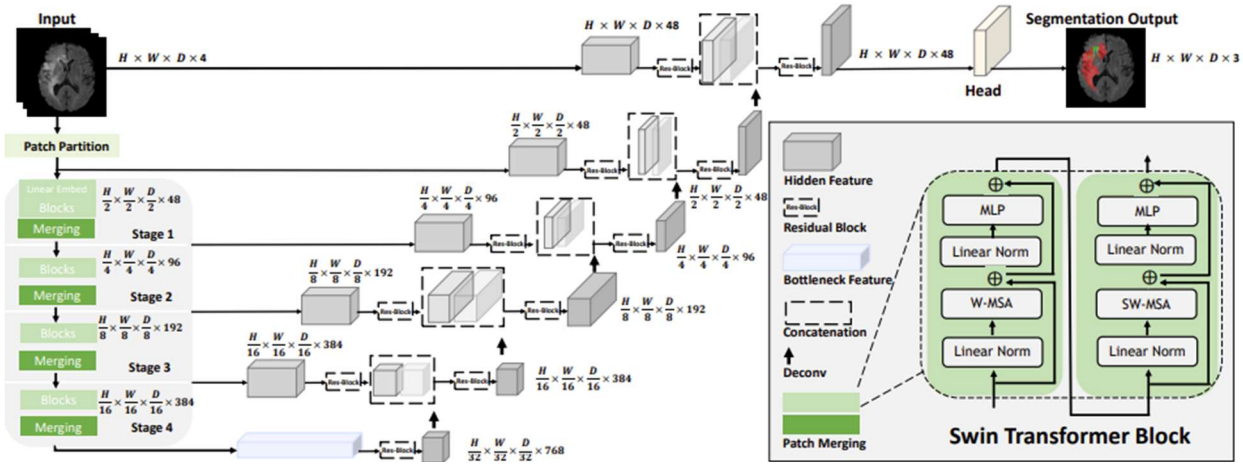


Fig. 6. SwinUNETR model architecture as proposed in [7]

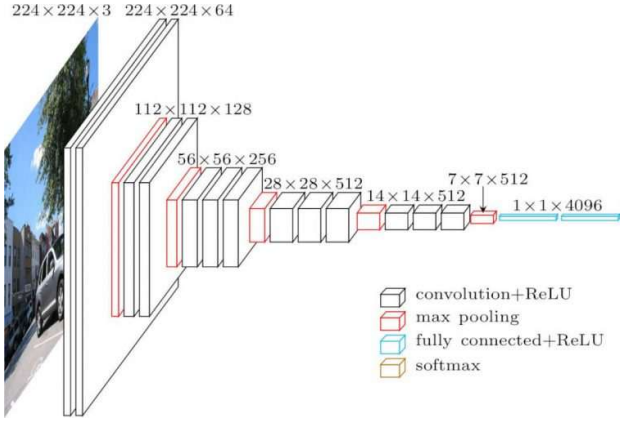


Fig. 7. VGG-16 Model Backbone used in VGG16-AttnUNet [6]

To compare the addition of attention gates at skip connections within the second model to a fully vision transformer-based architecture, the final stage of the evaluation pipeline will assess the performance of the SwinUNETR architecture proposed by [7] for the same segmentation tasks. This architecture, as depicted in Fig. 6, will also be implemented using the MONAI network library and will be adapted to process 2D input images instead of the original 3D image input design depicted in Fig. 6.

The Basic U-Net, VGG16-AttnUNet, and SwinUNETR models will each be independently trained for each of the four generated ground truth label images, producing 12 trained models in total. This evaluation scheme will provide a comprehensive framework for determining the impact of attention mechanisms and transformer-based architectures relative to the baseline model. These results also provide an opportunity to replicate and validate the results of these models against the reported performance observed in the literature.

D. Training Parameterization

The each of the 12 models developed using the pipeline above will be implemented in Python and trained using the PyTorch framework. All experiments were run on a Windows 11 machine equipped with a Nvidia RTX4070 GPU (12GB VRAM) and an Intel Core i9 CPU (13th Gen, 24-core). Due to limits in computational resources and training time, training for each model was limited to 20 epochs, and batch size ranged from 5-20 images depending on the parameter size of the model being trained. AdamW was used for the optimizer with a learning rate of 0.001, and a scheduler was used to apply 50% decay to the learning rate after every 5 epochs. An 80/20 training and validation split was used for training, and a 10% probability dropout was applied to the encoder to improve generalization and prevent overfitting. This dropout probability was increased to 30% for the VGG16-AttnUNet.

To complete training, the loss function used for each segmentation model was the combined Dice and Cross Entropy loss function. This loss function computes a weighted sum of both Dice and Cross Entropy loss and can be useful for combining the strengths and mitigating the weaknesses of both individual loss functions. The Dice and Cross Entropy loss functions are shown below.

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} y_{i,c}\hat{y}_{i,c} + \epsilon}{\sum_{i=1}^{N} y_{i,c} + \sum_{i=1}^{N} \hat{y}_{i,c} + \epsilon} \quad (1)$$

$$\mathcal{L}_{\text{CE}} = -\sum_{j=1}^{N\times C} y_j \log(\hat{y}_j) \quad (2)$$

The combined Dice Cross Entropy loss function with weighting parameters α and β is also shown below. The weighting for the loss function remained equal during training for all models.

$$\mathcal{L}_{\text{Total}} = \alpha \cdot \mathcal{L}_{\text{CE}} + \beta \cdot \mathcal{L}_{\text{Dice}} \quad (3)$$

III. Results

A. Dataset Selection

To assess the performance of the each of the implemented models, a variety of metrics will be tracked so that they can be analysed against each other and results reported in the literature. For semantic segmentation problems, Intersection Over Union (IOU) is a standard performance measure to assess how well the predicted segmentation mask aligns with the ground truth

Metric	Binary Segmentation		
	Basic U-Net	VGG16-AttnUNet	SwinUNETR
Train Time	28.89	44.33	42.99
DCE Loss	0.102	0.114	0.099
DSC	0.959	0.958	0.960
IOU	0.921	0.920	0.925

Metric	Part Segmentation		
	Basic U-Net	VGG16-AttnUNet	SwinUNETR
Train Time	29.88	53.08	47.25
DCE Loss	0.517	0.545	0.510
DSC	0.812	0.798	0.803
IOU	0.733	0.713	0.724

Metric	Instrument Segmentation		
	Basic U-Net	VGG16-AttnUNet	SwinUNETR
Train Time	39.52	47.22	45.86
DCE Loss	0.848	0.980	0.806
DSC	0.514	0.470	0.566
IOU	0.464	0.423	0.511

Metric	Part & Instrument Segmentation		
	Basic U-Net	VGG16-AttnUNet	SwinUNETR
Train Time	40.55	48.9	45.99
DCE Loss	1.020	1.006	0.952
DSC	0.296	0.300	0.323
IOU	0.263	0.262	0.288

Fig. 8 Training Time, DCE Loss, DSC, and IOU metrics gathered at the end of the last training epoch for each model, with binary segmentation on the upper right, part segmentation on the upper left, instrument segmentation on the lower left, and combined segmentation on the bottom right. Time tracked in min.

masks provided in the dataset. The Dice similarity coefficient will also be used to evaluate segmentation mask sets overlap as it is more sensitive to smaller changes in segmentation predictions than IOU. Both of these metrics are commonly reported in published articles on segmentation tasks and will be tracked during model testing to directly compare the performance of the novel model against other models in the literature. The equations for these metrics are shown below, where P represents the mask prediction and G represents the ground truth.

$$\mathrm{DSC} = \frac{2|P \cap G|}{|P| + |G|} \quad (4)$$

$$\mathrm{IoU} = \frac{|P \cap G|}{|P \cup G|} \quad (5)$$

In addition to DSC and IOU, another metric commonly used for segmentation performance assessment is the Hausdorff Distance (HD) metric. The HD measures the greatest distance from a point on the predicted mask boundary contour to the closest point on the ground truth contour and is thus a good measure of alignment of the surface boundary mismatch, as opposed to DSC and IOU which only measure whether pixels between the mask and ground truth match.

Precision, recall, and F1-score will also be tracked during testing to measure quality of classification for each class in the model. Precision measures the proportion of predicted pixels which are classified correctly against the ground truth, while recall measures the proportion pixels of a given class that are successfully identified. These are also standard metrics reported on in multiclass segmentation tasks and are more resistant to overlap metrics like DSC and IOU if the image being tested has high class imbalance within the image, as most of the testing images do. Model training and inference times are tracked as well to provide insight into the feasibility of real-time usage. A real time usage latency benchmark will be selected at ~33ms (30FPS) to match similar reported values of framerates for medical applications. Inference times will be averaged over 200 predictions for each class. All performance metrics will be collected on a previously unseen testing subset comprising 900 images. Since the EndoVis2017 dataset does not include ground truth labels for the combined part and instrument segmentation modality, the models trained on this configuration will only report training performance without a corresponding test set evaluation.

B. Basic U-Net Model Training Results

The Basic U-Net Model consisted of 7.8M trainable parameters, of which the cumulative size was 31.13MB. Excerpts of the same image at the beginning, middle, and end of training for the Basic U-Net model performing each of the four modalities of segmentation are shown in Figures 9 through 12, respectively. The DICECE loss, DSC score, and IOU score were recorded at the end of the last training epoch for each segmentation model and are presented in Figure 8. The Basic U-net consistently outperformed the other two model architectures in training time, likely due to the reduced number of trainable parameters in the more simplistic model architecture.

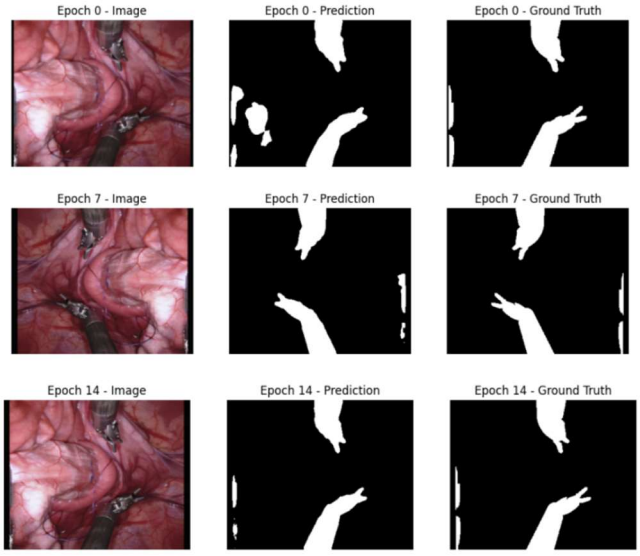


Fig. 9 Prediction progression images from training Basic U-Net for Binary Segmentation

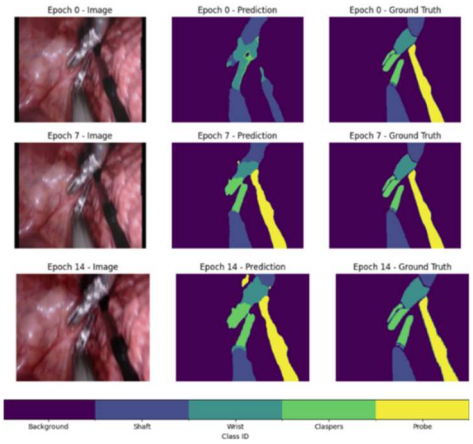


Fig. 10 Prediction progression images from training Basic U-Net for Part Segmentation

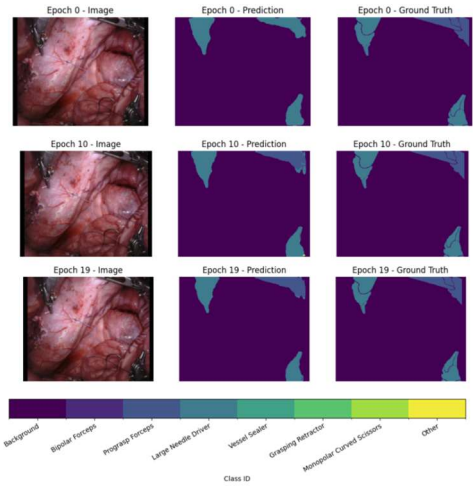


Fig. 11 Prediction progression images from training Basic U-Net for Instrument Segmentation

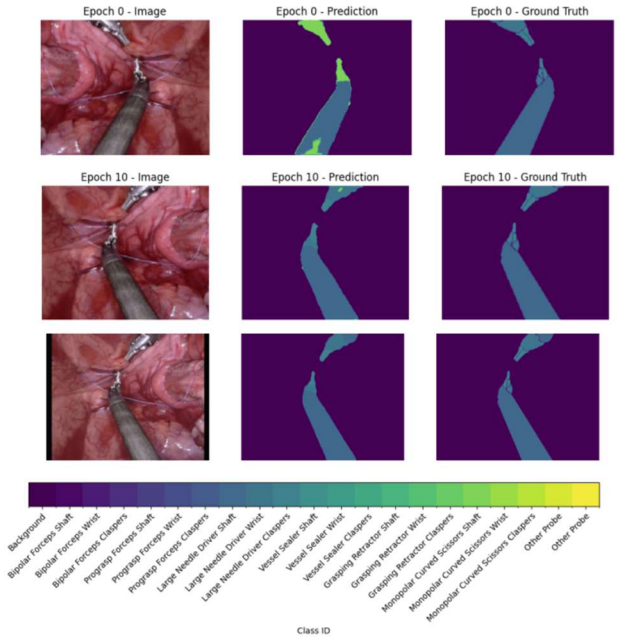


Fig. 12 Prediction progression images from training Basic U-Net for Combined Segmentation

C. *VGG16-AttnUNet Model Training Results*

The VGG16-AttnUNet Model consisted of 31.4M trainable parameters, of which the cumulative size was 125.62. Excerpts of the same image at the beginning, middle, and end of training for the VGG16-AttnUNet model performing each of the four modalities of segmentation are shown in Figures 13 through 16, respectively. The DICECE loss, DSC score, and IOU score were recorded at the end of the last training epoch for each segmentation model and are presented in Figure 8. The VGG16-AttnUNet was consistently outperformed by the other two model architectures every recorded training metric, likely due to the more aggressive dropout introduced to the VGG16-AttnUNet model.

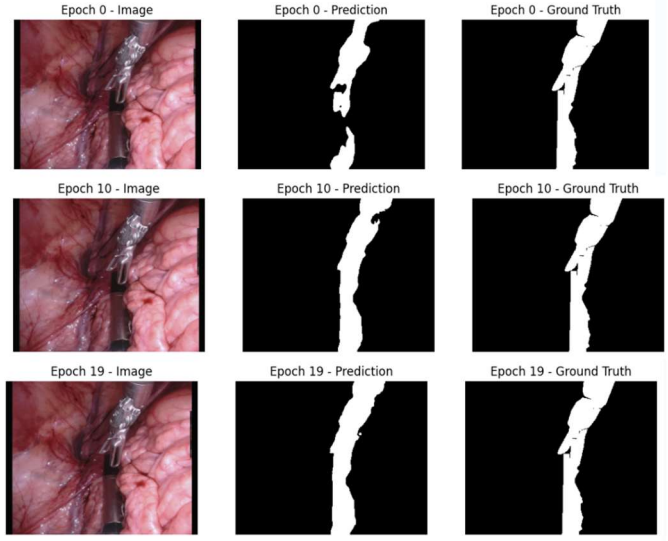


Fig. 13 Prediction progression images from training VGG16-AttnUNet for Binary Segmentation

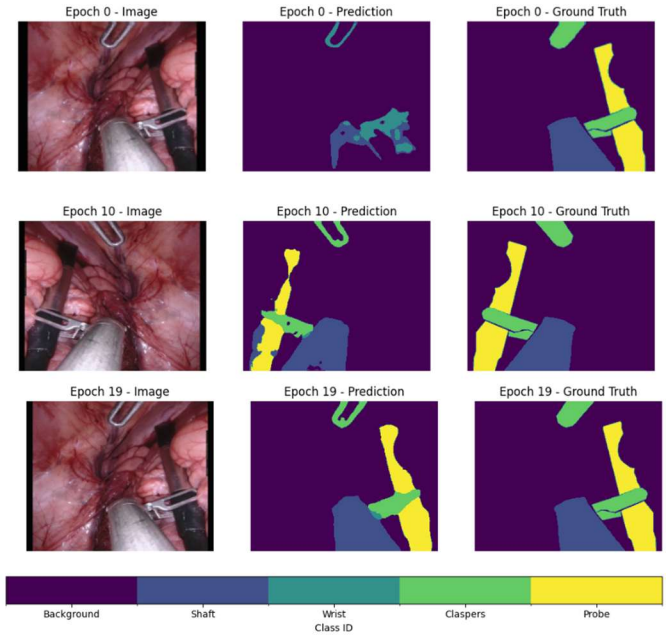


Fig. 14 Prediction progression images from training VGG16-AttnUNet for Part Segmentation

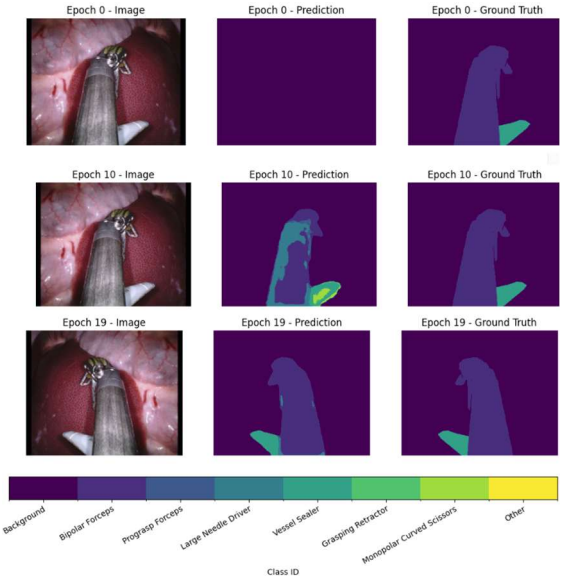


Fig. 15 Prediction progression images from training VGG16-AttnUNet for Instrument Segmentation

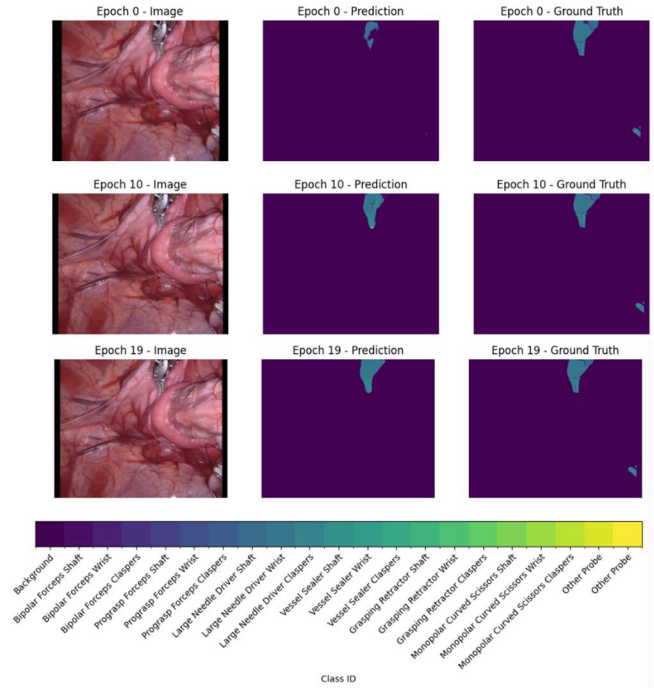


Fig. 16 Prediction progression images from training VGG16-AttnUNet for Combined Segmentation

D. SwinUNETR Model Training Results

The SwinUNETR Model consisted of 25.1M trainable parameters, of which the cumulative size was 100.56MB. Excerpts of the same image at the beginning, middle, and end of training for the SwinUNETR model performing each of the four modalities of segmentation are shown in Figures 17 through 20, respectively. The DICECE loss, DSC score, and IOU score were recorded at the end of the last training epoch for each segmentation model and are presented in Figure 8. The SwinUNETR had the lowest loss and highest DSC and IOU scores at the end of each model training with the exception of the part segmentation modality, where the Basic U-Net achieved better DSC and IOU scores.

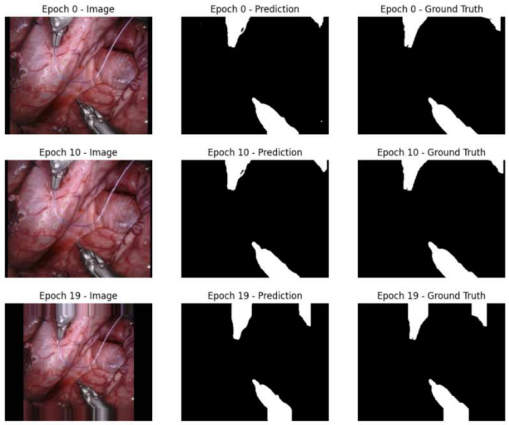


Fig. 17 Prediction progression images from training SwinUNETR for Binary Segmentation

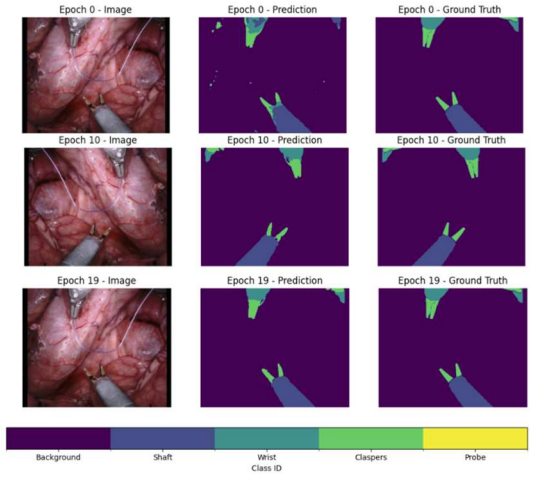


Fig. 18 Prediction progression images from training SwinUNETR for Part Segmentation

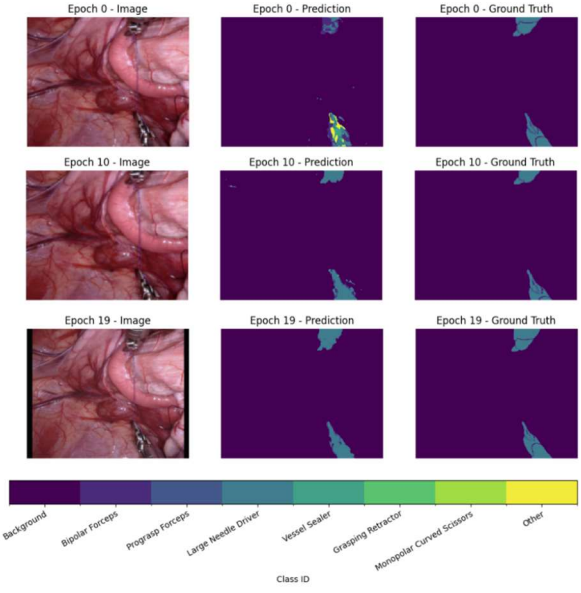


Fig. 19 Prediction progression images from training SwinUNETR for Instrument Segmentation

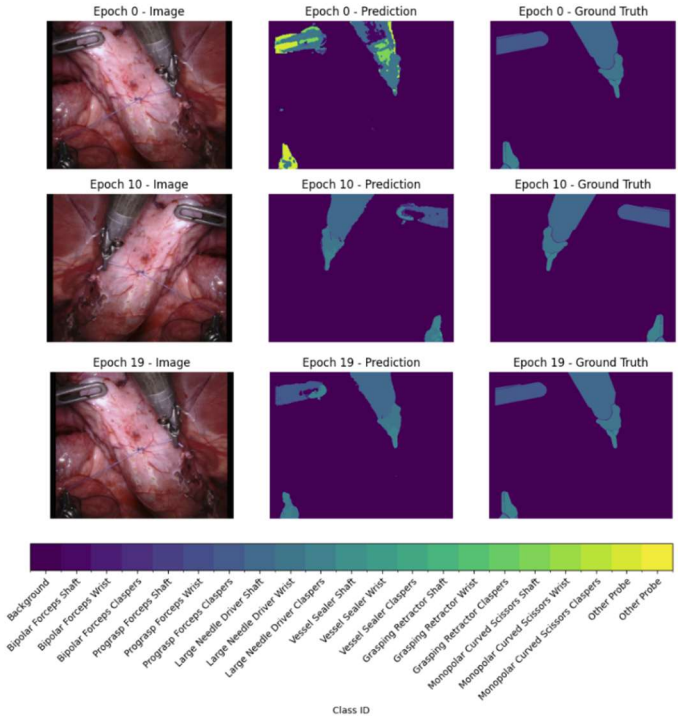


Fig. 20 Prediction progression images from training SwinUNETR for Combined Segmentation

E. Final Model Testing Results

The final testing results the nine models for which ground truth testing data were available are summarized within the table depicted in Fig. 21 below. The model that performed the best at a given metric for a given segmentation modality is highlighted in green. For the binary, part, and instrument segmentation tasks, average inference times ranged between 0.0074-0.0154, 0.0072-0.4515, and 0.02484-0.6451 seconds, respectively. The best binary segmentation performance in DSC score belonged to the VGG16-AttnUNet at 91.60%, the Basic U-Net performed the best part segmentation in DSC score at 79.90%, and the SwinUNETR performed the best instrument segmentation with a DSC score of 53.18%.

IV. Discussion

A. Intra-Study Model Comparisons

The results yielded from training and testing the segmentation models developed in this study were surprisingly similar despite the differences in model architecture. Upon training each of the 12 models, the preliminary indications obtained from the last training epoch (Fig. 8) seemed to suggest that the SwinUNETR model achieved marginally improved performance over the Basic U-Net and the VGG16-AttnUNet models. SwinUNETR outperformed the other two models in validation loss, DSC, and IOU in every segmentation task except for the part segmentation task in which the Basic U-Net performed the best. Despite SwinUNETR performing the best, the differences between each model were generally marginal, with the average variation across models amounting to only 0.073 in loss, 0.035 in DSC, and 0.035 in IOU. VGG16-AttnUNet seemed to consistently perform the worst in validation metrics, failing to outperform either of the two other models in a single validation metric, giving an initial indication that the impact of the custom model architecture was actually a detriment to model performance.

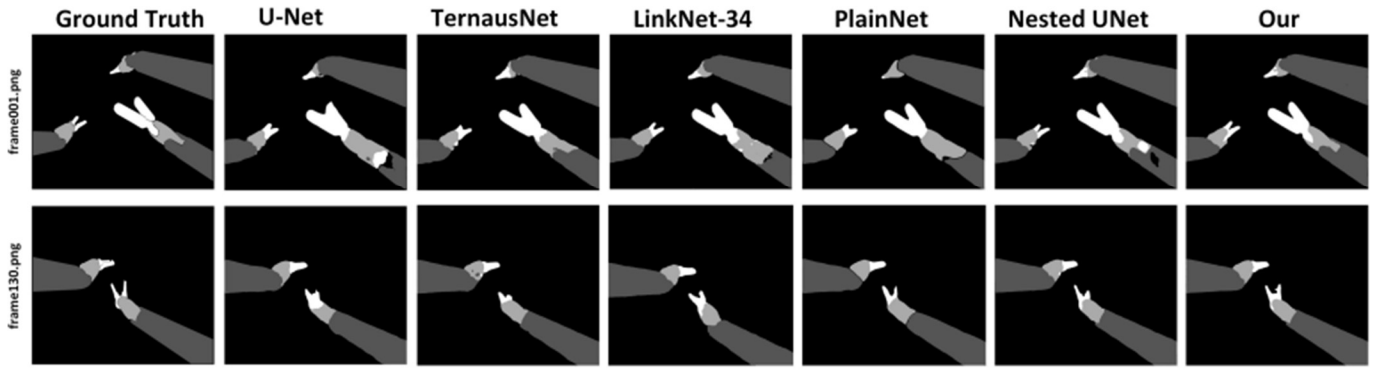
Despite the poor performance in many validation metrics, the masks produced by the models do look generally qualitatively similar to their respective ground truth masks in most visualized cases, despite the presence of misshapen boundaries and hallucinations upon scrutinous inspection. An interesting finding is that the models often appear to learn to mask the hole of the forceps instruments more accurately than the ground truth dataset, which fills the cavity of the forceps claspers to prevent occlusion of the background scene. This phenomenon is especially evident in Figures 14 and 20.

The results of the final model inference metrics (Fig. 21) on the testing dataset produced some key differences from the early indications provided by the metrics collected during training. The average variation across models again was confined to very limited ranges: 0.0117 in DSC, 0.0136 in IOU, 7.0090 in HD, 0.0200 in precision, 0.0276 in recall, and 0.0198 in F1 score. These ranges were confined below these averages for the binary and part segmentation tasks, but the SwinUNETR model began to see larger improvement margins on the other two models when conducting instrument type segmentation.

Interestingly, the trends initially suggested by the validation data were not completely consistent with the testing results. The best performing model for any given metric was largely varied as evidenced by Fig. 21, where each of the architectures excelled at different metrics for binary and part segmentation. The VGG16-AttnUNet likely performed worse in validation

Fig. 21 Averaged 200 Sample Inference Time, DSC Score, and IOU Score, HD Score, Precision, Recall, and F1-Score metrics gathered with the testing dataset for each model trained for Binary, Part, and Instrument Segmentation

Averaged Metrics	Binary Models			Part Seg Models			Instrument Seg Models		
	BUNet	VGG16-AttnUNet	SwinUNETR	BUNet	VGG16-AttnUNet	SwinUNETR	BUNet	VGG16-AttnUNet	SwinUNETR
Inference Time [sec]	0.0074	0.0114	0.0154	0.0072	0.4515	0.0148	0.2484	0.3465	0.6451
DSC	0.9100	0.9160	0.9156	0.7990	0.7980	0.7915	0.5215	0.5103	0.5318
IOU	0.8552	0.8650	0.8632	0.7129	0.7135	0.7053	0.4717	0.4586	0.4814
Hausdorff Distance	28.7077	26.0808	24.0247	43.1655	43.0072	41.3787	47.7621	53.2915	38.7343
Precision	0.8954	0.8957	0.9038	0.7664	0.7786	0.7683	0.6344	0.6138	0.6533
Recall	0.9532	0.9591	0.9528	0.8284	0.8153	0.8137	0.6564	0.6235	0.6854
F1 Score	0.9139	0.9172	0.9187	0.7872	0.7889	0.7826	0.6392	0.6130	0.6613



Methods	Binary segmentation		Parts segmentation		Type segmentation	
	IoU(%) ↑	Dice(%) ↑	IoU(%) ↑	Dice(%) ↑	IoU(%) ↑	Dice(%) ↑
U-Net	75.44 ± 18.18	84.37 ± 14.58	48.41 ± 17.59	60.75 ± 18.21	15.80 ± 15.06	23.59 ± 19.87
TernausNet	81.14 ± 19.11	88.07 ± 14.63	62.23 ± 16.48	74.25 ± 15.55	34.61 ± 20.53	45.86 ± 23.20
LinkNet-34	82.36 ± 18.77	88.87 ± 14.35	34.55 ± 20.96	41.26 ± 23.44	22.47 ± 35.73	24.71 ± 37.54
PlainNet	81.86 ± 15.85	88.96 ± 12.98	64.73 ± 17.39	73.53 ± 16.98	34.57 ± 21.93	44.64 ± 25.16
Nested UNet	82.94 ± 16.82	89.42 ± 14.01	58.38 ± 19.06	69.59 ± 18.66	**41.72 ± 33.44**	**48.22 ± 34.46**
SPP-LinkNet34 (Our)	**83.65 ± 16.47**	**89.80 ± 13.99**	**66.87 ± 17.10**	**76.93 ± 16.08**	15.96 ± 13.78	23.79 ± 18.88

Fig. 22 Segmentation performance results reported for Binary, Parts, and Instrument Type modalities reported by Hayat et al. [8], developers of the SEGSRNet labelled here as SPP-LinkNet34 (Our).

metrics than the other models due to the introduction of a more aggressive dropout probability, which can inhibit validation performance to improve generalizability at testing.

In totality, the intra-study testing data qualitatively suggests that the models each perform generally similarly, with little benefit if at all in aggregate segmentation performance when compared on any one metric besides inference time. The Basic U-Net consistently achieved better performance in average inference time, achieving inference times approximately 1.5-3 faster when compared to the other two models with only marginally reduced segmentation performance. This is likely due to the simpler model architecture of the Basic U-Net and the reduced number of trainable parameters inherent to the model's framework, which reduces the necessary computation burden at inference time. These results would surprisingly indicate that U-Net segmentation models perform only marginally better than models with pre-trained encoder backbones, attention gate mechanisms, or fully validated vision transformer architectures, if at all.

B. Reported Literature Model Comparisons

More interesting findings from this study continue to be revealed when comparing the generated results against segmentation algorithms reported on in recent literature. Fig. 22 displays the testing findings reported by [8], notably including average DSC and IOU scores of 89.80% and 83.65% for the best binary segmentation model, 76.93% and 66.87% for the best part segmentation model, 48.22% and 41.72% for the best instrument segmentation model. Notably, these results were generated on the same EndoVis2017 dataset utilized in this study, except for 300 testing images which were not included due to the presence of an unlabelled class. As evidenced by Fig. 20, the various models developed in this project seem to improve upon the segmentation accuracy in each task. When comparing by DSC score, the VGG16-AttnUNet outperformed the best reported model at binary segmentation by 1.80%, the Basic U-Net outperformed best reported model at part segmentation by 2.97%, and the SwinUNETR outperformed the best reported model at instrument segmentation with a DSC score of 4.96%.

A comparison to the results of another study from [5] reported on in the review seems to reaffirm the performance improvements. As reported in Fig. 23, the VGG16-AttnUNet outperformed the best reported model at binary segmentation by 1.40%, the Basic U-Net outperformed best reported model at part segmentation by 3.64%, and the SwinUNETR outperformed the best reported model at instrument segmentation with a DSC score of 7.11%. These results would seem to indicate that the implementation and tuning of these models are outperforming the state-of-the-art models reported in the literature.

Network	Binary Segmentation		Instrument Part		Instrument Type	
Metric	IoU	DICE	IoU	DICE	IoU	DICE
ToolNetH [6]	74.4	82.2	-	-	-	-
ToolNetMS [6]	72.5	80.4	-	-	-	-
FCN-8s [6]	70.9	78.8	-	-	-	-
CSL [13]	-	88.9	-	87.70 (Shaft)	-	-
U-Net [20]	75.44 (18.18)	84.37 (14.58)	48.41 (17.59)	60.75 (18.21)	15.80 (15.06)	23.59 (19.87)
U-Net + NN	**77.05**** (15.71)	**85.26*** (13.08)	**49.39*** (15.18)	**61.98*** (15.47)	**16.72*** (13.45)	**23.97** (18.08)
TernausNet [24]	83.60 (15.83)	90.01 (12.50)	65.50 (17.22)	75.97 (16.21)	33.78 (19.16)	44.95 (22.89)
U-NetPlus-VGG-11	81.32 (16.76)	88.27 (13.52)	62.51 (18.87)	74.57 (16.51)	**34.84*** (14.26)	**46.07**** (16.16)
U-NetPlus-VGG-16	**83.75** (13.36)	**90.20*** (11.77)	**65.75** (14.74)	**76.26*** (13.54) **94.75(Shaft)**	34.19 (15.06)	45.32 (17.86)

Fig. 23 Segmentation performance results reported for Binary, Parts, and Instrument Type modalities reported by Hasan et al. [5], developers of the U-NetPlus-VGG-11 and 16 models as labelled above.

The best theory I can devise as to the improved performance is due to the resizing of the images performed during data augmentation. To encourage reduce training times, part of my data augmentation pipeline involved resizing each image and mask from 1024×1280 to 256x320. Notably, none of the studies

performing training of models on the EndoVis2017 dataset explicitly report performing this step, likely for the reason that increased image resolution is advantageous when performing surgical tasks that require high precision and spatial resolution. Reducing the image size may have somehow augmented the performance of these models by reducing noise or accelerating convergence, despite likely being bad practice for such high-risk medical applications.

V. Conclusions

In this study, I developed implementations of three encoder decoder deep learning architectures based off of the U-Net architecture to perform binary and multiclass semantic segmentation of endoscopic surgical scenes. These deep learning models explored the impact of attention gates, pretrained convolutional encoders and vision transformers in improving upon the performance of the simplistic Basic U-Net architecture introduced in 2015. My implementation was successful in producing models that compare to and even surpass the state-of-the-art models reported in the literature, which was a confusing finding based on my limited personal experience with implementing deep learning networks.

The performance of the networks in this study has left me with many avenues for continued study. The most pressing question emanating from this work is how I was able to improve upon the results reported in the literature without any kind of novel breakthrough in architecture design. In continued work, the first action I would take to critically assess my results would be to re-train these networks while keeping the dataset images at their original resolution instead of aggressively down-sampling them. This would significantly increase training time, and more capable computational resources would likely be needed to re-train these models efficiently.

Another puzzling finding form my work in this study is the seemingly negligible impact of attention gates and vision transfomers in segmentation performance when compared to the Basic U-Net framework. After re-training these models at the original image size, the next route of investigation would be adjusting the parameterization of the VGG16-AttnUNet and SwinUNETR to make them more feature-rich in an attempt to fully utilize the ability of these more advanced architectures to learn the global and local contexts necessary for effective segmentation. Nonetheless, I think the work conducted in this study provides a solid foundation for future exploration and research into the applications of encoder-decoder frameworks for medical image segmentation.

References

[1] Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y. H., ... & Azizian, M. (2019). 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*.

[2] Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M. Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B. S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P. F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L. A. D., Roth, H. R., Xu, D., Bericat, D., Floca, R., Zhou, S. K., Shuaib, H., Farahani, K., Maier-Hein, K. H., Aylward, S., Dogra, P., Ourselin, S., & Feng, A. (2022).MONAI: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701. https://arxiv.org/abs/2211.02701N.

[3] Fernandes, E. Oliveira and N. F. Rodrigues, "Future Perspectives of Deep Learning in Laparoscopic Tool Detection, Classification, and Segmentation: A Systematic Review," 2023 IEEE 11th International Conference on Serious Games and Applications for Health (SeGAH), Athens, Greece, 2023, pp. 1-8, doi: 10.1109/SeGAH57547.2023.10253772.

[4] Han, Kai & Wang, Yunhe & Tian, Qi & Guo, Jianyuan & Xu, Chunjing & Xu, Chang. (2019). GhostNet: More Features from Cheap Operations. 10.48550/arXiv.1911.11907.

[5] S. M. Kamrul Hasan and C. A. Linte, "U-NetPlus: A Modified Encoder-Decoder U-Net Architecture for Semantic and Instance Segmentation of Surgical Instruments from Laparoscopic Images," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, pp. 7205-7211, doi: 10.1109/EMBC.2019.8856791.

[6] Hassan, M. U. (2018). VGG16 – Convolutional Network for Classification and Detection. Neurohive. Retrieved from https://neurohive.io/en/popular-networks/vgg16/

[7] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., & Xu, D. (2022). Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. arXiv preprint arXiv:2201.01266. Retrieved from https://arxiv.org/abs/2201.01266

[8] M. Hayat, S. Aramvith and T. Achakulvisut, "SEGSRNet for Stereo-Endoscopic Image Super-Resolution and Surgical Instrument Segmentation," 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 2024, pp. 1-4, doi: 10.1109/EMBC53108.2024.10782794.

[9] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* (pp. 234-241). Springer international publishing.

[10] Y. Sun, B. Pan and Y. Fu, "Lightweight Deep Neural Network for Real-Time Instrument Semantic Segmentation in Robot Assisted Minimally Invasive Surgery," in *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3870-3877, April 2021, doi: 10.1109/LRA.2021.3066956.

[11] Xia, Yanjie & Wang, Shaochen & Kan, Zhen. (2023). A Nested U-Structure for Instrument Segmentation in Robotic Surgery.