

Multidisciplinary Surgical Research Annals

<https://msra.online/index.php/Journal/about>

Volume 4, Issue 1 (2026)

<https://doi.org/10.5281/zenodo.18801897>

Microscopic Tissue Boundary Identification in Laparoscopic Videos Using Transformer-Based Segmentation

Maryam Hassan^{*1}, Fouzia Idrees²

Article Details

ABSTRACT

Keywords:

Laparoscopic Video Segmentation; Transformer-Based Segmentation; Tissue Boundary Detection; Hybrid CNN–Transformer; Real-Time Surgical Assistance; Medical Image Analysis

Maryam Hassan*

MPhil Scholar, Department of Computer Science, SBBWU, Peshawar
Email: mariamh0796@gmail.com

Fouzia Idrees

Lecturer, PhD in Multimedia Technology, Department of Computer Science, SBBWU, Peshawar

Accurate soft tissue boundary identification in laparoscopic videos is essential for surgical precision and intraoperative decision support. However, challenges such as smoke, specular reflections, rapid camera motion, occlusions, and low-contrast structures make reliable segmentation difficult. Traditional convolutional neural network (CNN)-based methods often struggle to capture long-range contextual dependencies, resulting in coarse and inconsistent boundary predictions.

This research proposes an efficient hybrid Transformer-guided encoder–decoder architecture for real-time microscopic tissue boundary segmentation in laparoscopic videos. The model integrates convolutional layers for local feature extraction with Transformer-based self-attention mechanisms to capture global contextual information. A multi-scale decoder with skip connections reconstructs high-resolution segmentation maps, while an auxiliary boundary refinement head enhances contour accuracy. The system is trained using a combination of Dice and cross-entropy loss and evaluated using Dice, IoU, Hausdorff Distance, and latency metrics.

The proposed approach aims to achieve improved boundary accuracy over existing CNN and Transformer baselines while maintaining near real-time performance (≤ 100 ms per frame) on a single GPU. This work contributes toward intelligent, reliable, and clinically deployable surgical assistance systems.

INTRODUCTION:

As a new application of Transformer to carry out segmentation in medical image analysis, it has produced a significant impact and good performance related to delineating tissue boundaries from laparoscopic videos. With their capabilities for modeling dependencies by self-attention at local and global levels, Transformers are uniquely positioned to perform this task and this innovation was spurred by the desire for better and more robust segmentation of medical imaging problems. For transformer based methods operating in real time, considering the most accurate identification of tissue boundaries is paramount in laparoscopic surgeries to ensure safe surgical precision is more essential than ever, these methods demonstrate notable advantages over typical convolutional neural networks (CNNs) based segmentation architectures [1].

Over the years, medical imaging technologies have come to be at a point that represents a tremendous improvement over laparoscopy in surgical procedures. Despite this, the dynamic and complex nature of laparoscopic videos have their related challenges in regard to segmenting soft tissues which do not have clear boundaries, occlude frequently, and have high appearance variability associated with their classification. However, there has been challenges of segmenting the complex structures of soft tissues using segmentation models, which include U-shaped Network (U-Net) and Mask Region-based Convolutional Neural Network (RCNN). With computers that can have sequential data processing capabilities, and learn hierarchical representations, Transformers have come to be integrated into medical imaging workflows [1]. Delineating tissue boundaries is the primary capability that contributes to the high accuracy of tissue delineation, furthermore, while the presence of motion artifacts and low contrast are present to help obscure the image quality, the presence of this noise does not diminish the delineation process of the algorithm.

There has been reported success of application of Transformers in different medical imaging contexts. As an example, boundary aware transformers have already demonstrated that inductive bias based on limited receptive field will overcome this limitation to segment ambiguous skin lesion boundaries [2]. Similarly, hybrid transformer models have been employed to increase segmentation accuracy for adaptive optics retinal images [3]. This development extends the limits of what transformers can provide by focusing on the specific needs of laparoscopic video analysis - including precision and flexibility.

The segmentation models are adaptive and resilient to the fast movements (shake), occlusions that will be present during performance of minimally-invasive surgeries where they exist. For example, transformer-based methods utilize multi-head self-attention layers and convolutional blocks, taking advantage of their learning adaptable experience to localized and global spatial. More frequently than not, they are better than other methods for segmenting complex structures in laparoscopic video, and we have also confirmed this. These transformer-based architectures are scalable and can be easily utilized in real-time tasks in a computationally efficient way. This is also true for segments in laparoscopic video, as they are also different scales and resolutions. The SwinPA-Net uses pyramid aggregation networks to help address the variances in lesion sizes and textures of the different types of tissue.

Accuracy to tissue boundary segmentations from laparoscopic video is clinically significant; and therefore, must be accurate. Errors in segmentations can lead to prolonged surgery time and increase patient morbidity due to the stress from unintentional damage to adjacent healthy tissue and structures. In response to this, transformer architecture based models have taken advantage of their inherent accuracy and resiliency with boundary modeling, and can model the dynamic deformation and segmentation of soft tissue, and in a time synchronized capacity. They can incorporate GNN and depth estimation that include segmenting deforming tissue volume, and are applicable to real time surgical scenarios.

Moreover, transformer based segmentations support the rapid transition to automated, and intelligent surgical workflows. Augmented reality (AR) and artificial intelligence (AI) have the potential to JOIN, congruently performed laparoscopic surgery, or in the case of their use, during intraoperative decision making. For instance, as an example, the one research team developed an augmented reality that mapped anatomical structures in laparoscopic video in an AI-based silhouette divide type of model imposed on a 3D

depiction of anatomy structures.

While the transformer-based segmentation has many functional properties, their high computational complexity and that it typically needs a large amount of data to create are considered drawbacks. In the medical imaging context especially, acquiring a large, annotated dataset to train the uni-directional transformers is not easy. However, there are ways to ameliorate this by pre-training the transformer models using transfer learning [8].

The project signifies a new way forward in the development of segmentation models based on Transformers, which is important for medical imaging (and laparoscopic video approaches). Our transformer-based segmentation models capitalize on the power of Transformers, using them to model both-in-the-plane global and local dependencies to segment soft tissue surfaces in real time, and hopefully also in a clinically relevant way. As medical imaging and surgical robotics continue to develop and develop more sophisticated use of transformations, and potentially , augmentative reality, we look to improve ranges of laparoscopic procedures in accuracy and efficiency, and of patient outcomes. Evaluating and translating these models to clinical environments are the first steps toward the larger vision of an intelligent and automated surgical system.

Literature Review

Model architecture containing transformer networks have opened new realms in the medical image segmentation field, and they demonstrate they can overcome the limitations of traditional convolutional neural networks (CNN) with respect to long-range dependency preservation and contextual information.

Natural language processing (NLP) transformer models have been utilized to enhance medical image segmentation accuracy and robustness for challenging anatomical structures.

For some time, traditional CNNs have dominated the medical image segmentation space solely based on their capacity of local feature extraction.

Unfortunately, this limited receptive field prevents them from capturing global dependencies necessary for segmenting complex medical images. However, existing research relies on convolutions that depend on a particular geometry for capturing both local and global features, while the Transformers tackle this problem by employing self-attention mechanisms to mimic how the human visual cortex aggregates all the information throughout the visual field to achieve better global features [10]. Dynamic linear transformers have been shown to be effective for 3D biomedical image segmentation with the speedups and feature extraction on volumetric data.

Recently, the Fully Convolutional Transformer (FCT) is a notable advancement in transformer based medical image segmentation, combining convolutional networks' local feature extraction with global dependency modeling of transformers. Transformer-based architectures have expanded the potential of the medical image segmentation domain and shown they can break the limitations of traditional convolutional neural networks (CNN) with respect to preserving the long-range dependencies and context information.

Transformer models from the natural language processing arena have been harnessed to improve medical image segmentation precision and robustness for complex anatomical structures.

For some time, traditional CNNs have dominated the medical image segmentation space solely based on their capacity of local feature extraction.

An anisotropic transformer model [16] also aims to solve this problem by handling irregular spacing in slice spacing in 3D datasets and improving segmentation accuracy for lung cancer imaging. Correspondingly, the Dilated Transformer employs a new self-attention mechanism to efficiently model long-range dependency and performs accurate CT and MRI data segmentation [17].

As an example, for integration of global and local context modeling in medical image segmentation MissFormer is proposed. MissFormer achieves state of the art performance across various tasks including multiorgan and cardiac segmentation [18]by redesigning transformer feedforward networks with multi scale feature extraction. In addition, weight inflation strategies have been used to adapt pre-trained vision

transformers from the 2D to the 3D, by applying transfer learning on medical imaging datasets [13].

To overcome their limitations the hybrid models blending CNNs and transformer have emerged as a promising approach. For instance, TransConver has a model with a parallel module, which can provide global context and local features in the same prediction and showed promising performance for brain tumor segmentation [19]. The final product is U-Net Transformer [20], which merges the spatial recovery properties of U-Net and the contextual modeling ability of transformers in an attempt to improve performance for abdominal CT segmentation.

A number of developers are performing important collective work in finding transformer based models that alleviate the constraints of usability and efficiency to model experiences in a clinical context. Axial Fusion Transformer UNet (AFTer UNet) redesigned the axial attention operations to disaggregate intra slice and inter slice dependencies in order to make correct predictions [21], while being conscientious to memory. The thesis also introduced automatic segmentation networks that made use of both explicit edges and reverse attention operations which also could improve boundary delineation when trained on a small data set [22].

In summary, transformer architectures have altered the landscape of medical image segmentation, and resilience is the evident showcase that architecture can breed from a traditional CNN, while also revealing where transformers have opened up some gaps to model idioms both globally and locally in the same time. The segmentation models are a much more durable foundation, and could understandably affect clinical practice and clinical education with inconceivable impact.

The explorations in research have only just begun, as there is so much left to understand the computational limits, but the potential has only just begun to expand the possibilities for using transformers in medical imaging. They will continue to make a mark in medical imaging and lend transformative strategies to negotiate the complexities of the segmentation task.

Problem Statement

Precise boundary identification of soft tissues in laparoscopic video is difficult due to smoke, specular highlights, fast camera motion, instrument occlusion and illumination variations. The existing systems often produce coarse masks or display imprecise contours or variable predictions across frames which hinder their ability to serve as navigation aids for intraoperative use.

There is a need for a reliable and efficient segmentation approach that produces clean boundaries at near real time speed and runs on a single workstation class GPU.

Aim and Objectives

Design and evaluate an efficient transformer guided segmentation system that predicts reliable tissue boundaries in laparoscopic video at near real time speed suitable for operating room integration.

The objectives of the study are to accomplish the following:

Build a transformer guided encoder decoder with a boundary refinement head that produces accurate contours.

Achieve means Dice and mean IoU improvements over strong convolutional and transformer baselines on laparoscopic datasets.

Meet a latency target close to one hundred milliseconds per frame on a single RTX class GPU, measured end to end including preprocessing and postprocessing.

Report boundary quality with Hausdorff distance and average surface distance in addition to Dice and IoU.

Methodology

Our work constructs a transformer model that accurately segments tissues from the frames of microscopic laparoscopic videos. The methodology outlines how data is collected, how the model works, and how the model is evaluated. We used the structure of the model as a guide, and the breakdown is as follows:

Data sources

Primary source will be a cholecystectomy video set with organ level masks such as liver, gallbladder, fat, momentum, and background. If desired this can be supplemented with an internally annotated collection of laparoscopic frames

Sampling

Sample frames uniformly across key surgical phases to avoid bias. Maintain patient level splits to prevent leakage.

Annotation workflow

Two trained raters will delineate organ boundaries using a standard tool. Compute inter rater agreement and resolve disagreements through adjudication.

Preprocessing

Resize frames to a fixed square resolution such as 640 by 640. Apply illumination normalization and optional mask for heavy specular spots.

Architecture of Transformer-Based Segmentation Model

The model design is proposed based on the following elements methodically organized to achieve the best possible outcomes:

Input Data: Laparoscopic video frames from datasets of multiple surgical settings

Preprocessing: Normalization, data augmentation, and resolution scaling for stable input quality.

Model Structure: A novel hybrid Transformer-based model built for tissue boundary segmentation:

Self-Attention Mechanism: Adopts long-distance dependencies and relations within frames.

Extract fine-grained local features by convolutional layers.

Encoder-Decoder Architecture: Multi-scale feature extraction (encoder) and reconstruction of the segmentation map (decoder).

Loss Functions: A hybrid of Dice loss and cross-entropy loss for improving accurate segmentation of complex boundaries.

Optimization Techniques: Incorporate linear attention mechanisms and regularization methods like dropout and batch normalization to reduce computational overhead and prevent overfitting.

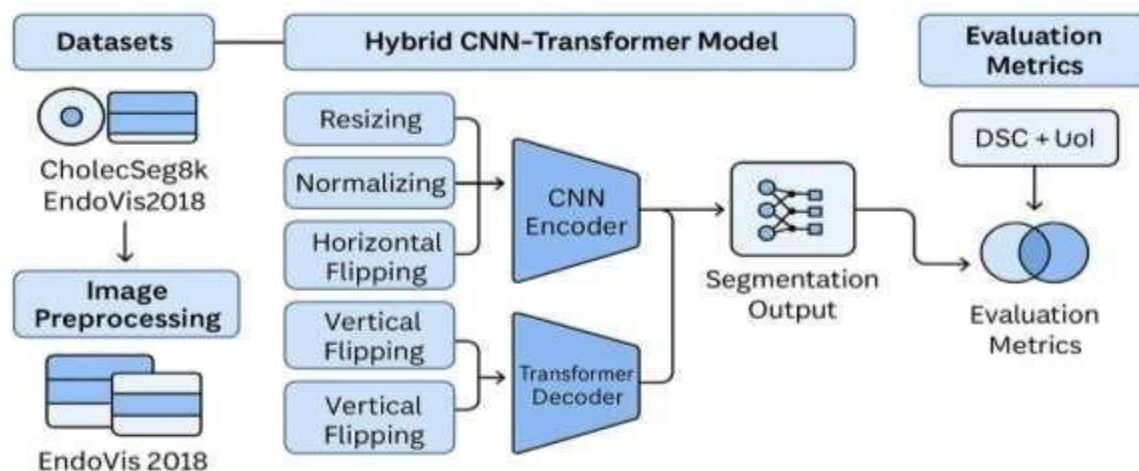


Figure 1 Diagram of the Proposed Model

Model structure

The model is an encoder decoder with skip connections and an auxiliary boundary head.
 Convolutional stem to capture local edges and texture.
 Transformer encoder with windowed self-attention or a linear attention variant to control memory.
 Multi scale feature aggregation to pass rich context to the decoder.
 Decoder with progressive upsampling and skip connections from matching encoder stages.
 Boundary refinement head that learns a contour logit map and combines it with the semantic mask.
 A simple architecture sketch is provided in the diagram above.

Reference configuration

Stage		Output size at 640 input	Details
Conv stem		160 by 160 by C	two 3 by 3 conv blocks with stride two
Transformer one	stage	160 by 160 by 2C	depth two, heads four, window size eight
Transformer two	stage	80 by 80 by 4C	depth two, heads six
Transformer three	stage	40 by 40 by 8C	depth four, heads eight
Transformer four	stage	20 by 20 by 16C	depth six, heads twelve
Decoder		160 by 160 by C	four upsampling blocks with skip connections
Boundary head		640 by 640 by one	one by one conv on fused features plus Sobel inspired refinement
Segmentation head		640 by 640 by K	one by one conv to K classes and softmax

Parameter count is expected in the twenty to thirty five million range depending on C and depth.

Research Plan

The research plan is divided into three main phases:

Background Study

To advance our research, it is essential to study all Transformer model types applied in medical imaging, along with their associated prior work. A particular focus should be placed on video segmentation challenges in laparoscopic procedures. By thoroughly examining existing literature, we can adapt and refine our model design to incorporate key findings and best practices from past research. The proposed model design involves the development of a Transformer based segmentation framework specifically optimized for real-time analysis of laparoscopic videos. This design integrates hybrid attention mechanisms and emphasizes computational efficiency, aiming to effectively manage the complex and dynamic nature of surgical environments.

Implementation

The proposed model will be implemented using Python, leveraging the PyTorch and Hugging Face libraries for the integration of Transformer components. There are multiple aspects that need to be accounted for during the implementation. For datasets, publicly available laparoscopic datasets will be used, and clinical

datasets from collaborating institutions. The datasets will be randomized and divided into training (70%), validation (20%), and testing (10%) datasets for the development and evaluation of the model.

During training, U-shaped architecture will be developed that will have the Transformers in the encoder, which will allow us to have multiple layers and extract nested features hierarchically. To improve computation efficiency, the model will employ linear attentions and regularizations. The model will perform training on GPUs to maximize efficiency. Hyper parameters will be varied according to an internal grid search, and with Bayesian optimization methods for determining the best parameters.

For evaluation, segmentation accuracy will be reported using the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) metrics, and evaluating latency for determining the performance for real-time surgical applications.

Experimental Testing and Evaluation

The model will be further tested within a simulated surgical environment that reflects realistic laparoscopic conditions, including motion artifacts, tissue deformation, and lighting changes. This environment allows us to test the model's robustness and adaptability while the surgical conditions are dynamic. The integration with the surgical simulator will pay special attention to real-time performance with a major performance consideration being to maintain predictions at or below 100 milliseconds per frame during testing. This is important to demonstrate the model's potential for intraoperative support and guiding surgical procedures in real-time.

Socio Economic Impact

The socio-economic impact of developing reliable transformer-based segmentation for laparoscopic surgeries is considerable. From a healthcare perspective, improved boundary detection minimizes surgical errors, reduces operation time, and lowers the risk of complications, which collectively enhance patient safety and recovery outcomes. These improvements directly translate into reduced hospitalization costs, fewer readmissions, and more efficient utilization of healthcare resources. On a broader economic scale, shorter recovery periods allow patients to return to work sooner, thereby decreasing productivity losses for individuals and society. Additionally, the integration of advanced computational methods in surgical practice fosters innovation within the medical technology sector, stimulating investment, job creation, and skills development in both healthcare and computational research fields. In resource-constrained settings, such innovations can optimize the use of limited medical staff and infrastructure, ultimately contributing to more equitable access to quality healthcare.

References

- Z. Wu, X. Zhang, F. Li, and S. Wang, "TransRender : a transformer-based boundary rendering segmentation network for stroke lesions," 2023.
- J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, and J. Qin, "Boundary-Aware Transformers for Skin Lesion Segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12901 LNCS, pp. 206–216, 2021, doi: 10.1007/978-3-030-87193-2_20.
- J. Liu, J. Li, A. Wolde, C. Cukras, and J. Tam, "Hybrid transformer for lesion segmentation on adaptive optics retinal images," 2022, p. 92. doi: 10.1117/12.2612379.
- K. Fang et al., "UMRFormer-net: a three-dimensional U-shaped pancreas segmentation method based on a double-layer bridged transformer network," *Quant. Imaging Med. Surg.*, vol. 13, no. 3, pp. 1619–1630, 2023, doi: 10.21037/qims-22-544.
- Hao Du; Jiazheng Wang; Min Liu; Yaonan Wang; Erik Meijering, "SwinPA-Net: Swin Transformer-Based Multiscale Feature Pyramid Aggregation Network for Medical Image Segmentation," *IEEE Trans.*

- Neural Networks Learn. Syst., vol. 35, no. 4, pp. 5355–5366, 2022, doi: 10.1109/TNNLS.2022.3204090.
- R. Docea et al., “SeeSaw: Learning Soft Tissue Deformation from Laparoscopy Videos with GNNs,” *IEEE Trans. Biomed. Eng.*, vol. 71, no. 12, pp. 3432–3445, 2024, doi: 10.1109/TBME.2024.3424771.
- M. Kasai, H. Uchiyama, T. Aihara, S. Ikuta, and N. Yamanaka, “Laparoscopic Projection Mapping of the Liver Portal Segment, Based on Augmented Reality Combined With Artificial Intelligence, for Laparoscopic Anatomical Liver Resection,” *Cureus*, vol. 15, no. 11, 2023, doi: 10.7759/cureus.48450.
- B. Namazi, G. Sankaranarayanan, and V. Devarajan, “Attention-based surgical phase boundaries detection in laparoscopic videos,” in *Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019, 2019*, pp. 577–583. doi: 10.1109/CSCI49370.2019.00109.
- S. M. K. Hasan, R. A. Simon, and C. A. Linte, “Segmentation and removal of surgical instruments for background scene visualization from endoscopic/laparoscopic video,” vol. 11598, p. 7, 2021, doi: 10.1117/12.2580668.
- Y. Zhang, S. C. Huang, Z. Zhou, M. P. Lungren, and S. Yeung, “Adapting pre-Trained vision transformers from 2d to 3d through weight inflation improves medical image segmentation,” in *Proceedings of Machine Learning Research, 2022*, pp. 391–404.
- A. Tragakis, C. Kaul, R. Murray-Smith, and D. Husmeier, “The Fully Convolutional Transformer for Medical Image Segmentation,” in *Proceedings - 2023 IEEE Winter Conference on Applications of Computer Vision, WACV 2023, 2023*, pp. 3649–3658. doi: 10.1109/WACV56688.2023.00365.
- H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, “nnFormer: Interleaved Transformer for Volumetric Segmentation,” vol. XX, no. Xx, pp. 1–10, 2021, [Online]. Available: <http://arxiv.org/abs/2109.03201>
- Z. Zhang and W. Luo, “Hierarchical volumetric transformer with comprehensive attention for medical image segmentation,” *Math. Biosci. Eng.*, vol. 20, no. 2, pp. 3177–3190, 2023, doi: 10.3934/mbe.2023149.
- A. M. Shaker, M. Maaz, H. Rasheed, S. Khan, M. H. Yang, and F. S. Khan, “UNETR++: Delving into Efficient and Accurate 3D Medical Image Segmentation,” *IEEE Trans. Med. Imaging*, 2024, doi: 10.1109/TMI.2024.3398728.
- S. Tan et al., “SegStitch: Multidimensional Transformer for Robust and Efficient Medical Imaging Segmentation,” vol. 14, no. 8, pp. 1–10, 2024, [Online]. Available: <http://arxiv.org/abs/2408.00496>
- D. Guo and D. Terzopoulos, “A transformer-based network for anisotropic 3D medical image segmentation,” *Proc. - Int. Conf. Pattern Recognit.*, pp. 8857–8861, 2020, doi: 10.1109/ICPR48806.2021.9411990.
- Y. Wu et al., “D-former: a U-shaped Dilated Transformer for 3D medical image segmentation,” *Neural Comput. Appl.*, vol. 35, no. 2, pp. 1931–1944, 2023, doi: 10.1007/s00521-022-07859-1.
- X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, “MISSFormer: An Effective Transformer for 2D Medical Image Segmentation,” *IEEE Trans. Med. Imaging*, vol. 42, no. 5, pp. 1484–1494, 2023, doi: 10.1109/TMI.2022.3230943.
- J. Liang, C. Yang, M. Zeng, and X. Wang, “TransConver: Transformer and convolution parallel network for developing automatic brain tumor segmentation in MRI images,” *Quant. Imaging Med. Surg.*, vol. 12, no. 4, pp. 2397–2415, 2022, doi: 10.21037/qims-21-919.
- O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, “U-Net Transformer: Self and Cross Attention for Medical Image Segmentation,” *Lect. Notes Comput. Sci. (including Subser. Lect.*

Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12966 LNCS, pp. 267–276, 2021, doi: 10.1007/978-3-030-87589-3_28.

X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, and X. Xie, “AFTer-UNet: Axial Fusion Transformer UNet for Medical Image Segmentation,” Proc. - 2022 IEEE/CVF Winter Conf. Appl. Comput. Vision, WACV 2022, vol. c, pp. 3270–3280, 2022, doi: 10.1109/WACV51458.2022.00333.

J. Zhang, F. Li, X. Zhang, H. Wang, and X. Hei, “Automatic Medical Image Segmentation with Vision Transformer,” Appl. Sci., vol. 14, no. 7, 2024, doi: 10.3390/app14072741.