Hierarchical Transformer-Based Semantic Segmentation of Intraoperative Anatomical Structures

by

Anindya Majumder
23241062
Md. Mubashirul Islam
21201428
Nirvik Shaha Rudra
21201241
Shah Samiur Rahman
21201492
Shuddha Sourav Prachi
21201429

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfilment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
June 2025.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing the degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted or submitted for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Anindya Majumder
23241062

Shah Samiur Rahman
21201492

Shuddha Sourav Prachi
21201429

Nirvik Shaha Rudra
21201241

Md. Mubashirul Islam
21201428

Approval

The thesis titled "Hierarchical Transformer-Based Semantic Segmentation of Intra-operative Anatomical Structures" submitted by

1. Anindya Majumder (23241062)
2. Nirvik Shaha Rudra (21201241)
3. Shah Samiur Rahman (21201492)
4. Shuddha Sourav Prachi (21201429)
5. Md. Mubashirul Islam (21201428)

Of Spring, 2025 has been accepted as satisfactory in partial fulfilment of the requirement for the degree of B.Sc. in Computer Science on June 20, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science & Engineering
Brac University

Co-Supervisor:
(Member)

Md. Saiful Islam
Senior Lecturer
Department of Computer Science & Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science & Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Associate Professor & Chairperson
Department of Computer Science & Engineering
Brac University

Abstract

We propose a novel hierarchical Transformer-based model that significantly advances the accuracy of semantic segmentation for intraoperative abdominal organs, outperforming existing methods in both precision and generalizability. Despite the growing prominence of surgical data science, current semantic segmentation models for intraoperative abdominal organ segmentation remain severely limited, both in terms of the number and capabilities. Only one prior approach exists, which fails to generalise beyond dominant anatomical structures due to large-organ bias and poor spatial contextualization. Addressing these critical gaps, this study introduces a hierarchical Transformer-based architecture explicitly tailored for the semantic segmentation of intraoperative imagery. Anchored by a DINOv2 backbone and a custom multi-scale cross-attention decoder, our model captures both boundary-level granularity and global anatomical consistency. The architecture is further supported by a hybrid loss function that combines class-weighted cross-entropy with Dice loss to enhance performance on structurally complex, low-pixel regions. The research objective focuses on enabling high-precision segmentation for surgical environments, overcoming occlusions, spatial overlap, and class imbalance that are endemic to existing datasets. On the Dresden Surgical Anatomy Dataset, our novel architecture achieves a Dice score of 0.9098 and mIoU of 0.8654, outperforming existing benchmarks by up to 28% in challenging classes such as the pancreas and colon. By addressing both the architectural and data-centric limitations of existing literature, this work establishes a new frontier in surgical precision.

Acknowledgement

Firstly, we would like to humbly thank the Almighty, whose boundless mercy, strength, and wisdom have guided us through every step of this journey. Without his divine blessings, this achievement would not be possible. We also extend our heartfelt gratitude to our respected supervisor, Md. Golam Rabiul Alam, whose invaluable guidance, insightful feedback, and unwavering support have been instrumental throughout our research. Our sincere thanks also go to our co-supervisor, Md. Saiful Islam, for his thoughtful suggestions and constant encouragement, which have deeply enriched our work.

Table of Contents

Declaration i

Approval ii

Abstract iii

Acknowledgement iv

Table of Contents v

List of Figures vii

List of Tables viii

1 Introduction 1
- 1.1 Introduction . . . 1
- 1.2 Motivation . . . 1
- 1.3 Problem Statement . . . 2
- 1.4 Research Objective . . . 2
- 1.5 Methodology in Brief . . . 3
- 1.6 Scopes and Challenges . . . 3
- 1.7 Report Organisation . . . 4

2 Literature Review 5
- 2.1 Preliminaries . . . 5
- 2.2 Review of Existing Research . . . 5
- 2.3 Summary of Key Findings . . . 17

3 Methodology 18
- 3.1 Methodology Overview . . . 18
- 3.2 Data Preparation . . . 19
  - 3.2.1 Data Collection . . . 19
  - 3.2.2 Exploratory Data Analysis . . . 19
- 3.3 Data Preprocessing . . . 20
  - 3.3.1 Data Cleaning . . . 20
  - 3.3.2 Data Transformation . . . 21
  - 3.3.3 Annotations . . . 21
- 3.4 Architecture Design . . . 22
  - 3.4.1 Pre-analysis of Existing Architectures . . . 22
  - 3.4.2 Model Architecture . . . 23

4 Implementations & Result Analysis **28**
4.1 Implementations 28
4.1.1 Requirements 28
4.1.2 Model training 28
4.1.3 Hyper-parameters 30
4.2 Performance Evaluation 31
4.2.1 Key Metric 31
4.2.2 Statistical Analysis 34
4.2.3 Comparisons and Relationships 35
4.2.4 Discussion 36

5 Conclusion **38**
5.1 Conclusion 38
5.2 Summary of Findings 38
5.3 Limitations 39
5.4 Recommendations for Future Work 39

Bibliography **42**

List of Figures

3.1 Top level overview of the proposed system 18
3.2 Dynamic visualisation of annotation mask and ground truth image . . 20
3.3 Pascal VOC 2012 Annotation, where each RGB values indicate a distinguished class 21
3.4 Model Architecture 23
3.5 Backbone Architecture 24
3.6 Decoder Architecture 26

4.1 Gradient Convergence Curve with Validation Loss 29
4.2 Model Learning Visualisation through epochs 30
4.3 Validation Loss, Pixel Accuracy and mIOU VS Epoch 31

List of Tables

4.1 Performance metrics for semantic segmentation (mean over 6 runs), including 95% confidence intervals (CI) and p-values based on a one-sample t-test. 34
4.2 Class-wise semantic segmentation performance 34
4.3 Comparison of Dice scores with existing study [21]. Our model shows consistent improvements across all organs, including up to 95.8% relative gain. 35

Chapter 1

Introduction

1.1 Introduction

Surgical data science is a rapidly growing field that focuses on analysing surgical data to enhance the precision of surgical procedures. In recent years, surgical scene segmentation has become a growing field because of the emerging remote surgeries, particularly robot-assisted surgery. Remote surgery is a kind of surgical procedure where the surgeon doesn't need to be present physically at the operating theatre; rather, he could perform the surgery using robotic arms like the da Vinci Surgical System, where he could view the organs using high-definition surgical cameras. For the remote surgical procedure, it is a vital issue for surgeons to identify the internal organs with the highest precision; hence, it will be helpful if the footage itself shows the internal organs with precision. Considering this issue, we have developed a novel hierarchical transformer-based semantic segmentation architecture which could perform the real-time segmentation of internal organs with the highest precision. This study explores how our architecture leverages multi-modal inputs to enhance the segmentation and visualisation of colour-graded internal organ images. Our goal is to develop a system capable of accurately and efficiently segmenting major abdominal anatomical structures in real-time.

1.2 Motivation

The evolution of robotic-assisted telesurgery represents a transformative leap in modern surgery, enabling expert surgeons to operate remotely with high precision. However, the effectiveness of such systems hinges critically on the clarity and interpretability of the visual feedback provided to the operator. In the absence of direct tactile interaction, the real-time segmentation and accurate visualisation of internal anatomical structures become not only desirable but essential for surgical safety and efficacy. Despite recent advances, existing segmentation methods often fall short in handling the complexity and variability inherent in surgical scenes, especially for abdominal organs. This is particularly problematic in telesurgical contexts, where latency, visual ambiguity, and anatomical similarity can compromise both situational awareness and procedural accuracy. Motivated by these challenges, this study introduces a hierarchical transformer-based semantic segmentation architecture designed specifically for telesurgical environments, which is able to segment 11 distinct abdominal organs with precision. By leveraging multi-modal inputs and contextual

attention, our approach enhances organ boundary delineation and delivers high-fidelity, real-time visualisation, directly supporting the demands of next-generation telesurgery or robotic surgery.

1.3 Problem Statement

Abdominal surgery is performed worldwide but still carries significant intraoperative risks: in a study of 9,288 abdominal procedures, 183 cases (about 2%) experienced adverse events—predominantly bowel (44%) and vessel (29%) injuries—which were associated with over a threefold increase in 30-day mortality and morbidity [3]. A primary factor in these complications is the surgeon's limited ability to recognise and delineate anatomical structures in real time, especially under variable lighting, occlusions, and complex tissue appearances. This scenario is getting worse when it is a telesurgery or robot-assisted surgery, where the surgeon is highly reliant on the visual output of the device. Existing computer-aided anatomy recognition algorithms show moderate to good precision in intrathoracic and abdominal surgery but are hindered by small, heterogeneous annotated datasets and challenges in capturing both fine-grained boundary details and global organ context [11]. To address these limitations, we aim to develop a novel hierarchical transformer-based architecture that can perform state-of-the-art real-time semantic segmentation of internal abdominal organs. By leveraging the self-attention mechanism and multi-level feature extraction, our proposed model is designed to help surgeons better visualise anatomical structures during surgery, enhancing precision and reducing the risk of complications.

1.4 Research Objective

Abdominal surgery carries significant risks if anatomical structures are not accurately recognised during the surgical procedure, potentially leading to inadvertent injury or bleeding. Transformer models can capture global and local context via self-attention, which is crucial for reliable segmentation in challenging intraoperative settings. We propose a hierarchical transformer model for the semantic segmentation of internal abdominal organs. The model uses hierarchical transformer blocks with a DINO-V2[23] backbone and a custom decoder. Early layers capture fine details and boundaries, while deeper layers capture overall organ shape and context. Positional encodings keep spatial structure intact. We will monitor real-time segmentation of surgical footage to evaluate feasibility in robotic and semi-automated surgery environments, measuring inference latency, segmentation accuracy under intraoperative challenges, and the potential to improve surgical precision and patient safety. By providing accurate anatomical guidance, our approach aims to support semi-automated surgeries and help surgeons achieve better precision while recognising anatomical structures. Finally, we outline future implementations such as augmented reality overlays and AI-assisted segmentation enhancements to support interactive surgical training and move toward fully autonomous surgical systems.

1.5 Methodology in Brief

Based on the methodology, this study employed a deep learning approach for abdominal organ segmentation using the Dresden Surgical Anatomy Dataset[17], comprising 11 organ classes from laparoscopic surgery videos of 32 patients. The research methodology involved:

- **Data Collection & Preprocessing**: The dataset of 1920×1080 dimensions of images was restructured, split into training (67%), validation (22%), and test sets (11%), resized to 1024×1024 with zero-padding, and normalised. Class imbalance was addressed via ROI extraction for small organs and extensive augmentation (rotations, flips, noise injection, etc.) for multilabel images.
- **Architecture**: A hybrid model was developed by fine-tuning the DINOv2[23] (ViT-L/14) backbone for feature extraction, integrated with a custom decoder. The decoder used object queries and multi-scale cross-attention blocks to generate masks, producing a 12-channel output (organs + background), upsampled to 1022×1022.
- **Training & Analysis**: The model was trained using a combined Cross-Entropy and Dice loss with class-weighted emphasis (1.2–1.4 for organs, 0.05 background). Optimisation used AdamW (LR=1e-5), batch size=2, and early stopping. Performance was evaluated via mIoU, Dice score, and visual inspection, with hyperparameters tuned iteratively based on validation metrics.

The end-to-end process encompassed exploratory analysis, iterative model refinement, and rigorous evaluation to ensure robust segmentation for surgical applications.

1.6 Scopes and Challenges

There are several real-world applications where our model can be used. To begin with, our model can assist in Surgical Assistance by segmenting organs in real time during surgery. For example, it can automatically highlight the liver, intestines, etc., during laparoscopy to help avoid accidental damage. Next, we can integrate our model into telesurgery or robotic surgery. It can be added to the robot's vision system so that the robot can track specific organs and automatically adjust the camera view. Lastly, our model can be used in Distant Surgery, which demands precise visual feedback and real-time performance to let surgeons safely operate from remote locations using robotic arms. Our segmentation model can help visualise and label organs, guide robotic arms, or warn about proximity to vital areas.
We mainly faced challenges with the *Dresden Surgical Anatomy* dataset [17], where we noticed several imbalances. First, the dataset does not contain equal quantities of all 11 classes. Some organs are overrepresented, while others appear in far fewer images. In addition, multilabelled images are much fewer than single-labelled ones, creating issues in our experiments. Second, organs vary greatly in size within video frames. Some, like intestinal veins, are very small, while others, like the stomach, occupy larger areas. Large organs cover more pixels and are easier for models to segment—indeed, models segment them with higher accuracy. Third, most dataset

images are background-dominated, with 40%-50% of each image being background on average. We also struggled to find an ideal backbone and decoder architecture that could segment abdominal organs with higher precision. We tested over 10+ architectures, including OneFormer and SAM2 [18]. Ultimately, we chose DinoV2 [23] as our backbone for its ability to learn spatial features effectively, and we developed a decoder architecture inspired largely by *Mask2Former*.

1.7 Report Organisation

This report is organised into five chapters such as Introduction, Literature Review, Methodology, Implementation & Result Analysis and Conclusion. These chapters perfectly outline our esteemed research on Hierarchical Transformer-Based Semantic Segmentation of Intraoperative Anatomical Structures.
The Structure is as follows:

1. **Introduction**:
   This chapter overviews our research motivation, along with the research objective, and clearly articulates our research problem and the objectives we aim to achieve.

2. **Literature Review**:
   This chapter includes preliminaries and the existing research related to our working domain. In this chapter, we jot down numerous transformer architectures and current trends in surgical data science, as well as, highlight the key contributions of prior work and research gaps that our research aims to address. A summary of key findings from the research papers is also included in this chapter.

3. **Methodology**:
   This chapter covers a methodology overview with the design process and workflow diagram. In addition, this chapter contains the data preparation segment, including dataset introduction, exploratory data analysis and data preprocessing. Pre-analysis of existing architectures and our novel architecture is also covered in this chapter.

4. **Implementations & Result Analysis**
   This chapter outlines different parts of the model training process as well as different performance metrics, IoU, F1-score, SSIM, etc, to evaluate the performance of our proposed model. Moreover, hyperparameter tuning that we will perform, like learning rate, batch size, dropout rate, etc., is also included in this chapter.

5. **Conclusion**:
   This chapter gives an overview of the future implementation of our model and summarises the outcome of our research and potential future applications of the proposed model, like fully autonomous robotic surgery and the use of augmented reality in robotic surgery.

Chapter 2

Literature Review

2.1 Preliminaries

Most of the papers we have covered for the literature review are from 2020-2024, and we found these papers mainly through Google Scholar, where a few of the papers were suggested by our supervisor. The papers contain different methodologies and models used for vision-related tasks such as feature extraction and segmentation of images. We mainly read about different types of Transformers, along with deep learning models. The detailed literature review below contains different models and methodologies, experiments carried out with those models, different datasets used for those experiments, results of experiments, comparison with other models and limitations of models with scope for improvements.

2.2 Review of Existing Research

We have studied numerous transformer architecture-related papers and surgical data science application papers focused on applied computer vision tasks. The summary of the papers we studied is given below with proper citations:

Han et al.[8] (2021) came up with a new architecture called Transformer iN Transformer (TNT), which provides a solution to a limitation of ViT, which misses local details due to patches being too coarse. TNT views local patches as "visual sentences" and breaks them down into sub-patches("visual words") for fine-grained attention. TNT architecture is built by stacking the TNT blocks, in which an outer transformer block is used for processing the sentence embeddings, and an inner transformer is used to model the relation among word embeddings. For training the TNT model, ImageNet ILSVRC 2012 was used. Compared to CNNs, TNT-B can outperform the widely used ResNet and RegNet by achieving 96.3% accuracy in Top-5. From experiments, the authors found that TNT outperformed all other visual transformer models. In Particular, surpassing the DeiT-S model by 1.7%, TNT-S achieved 81.5% top-1 accuracy. This indicates that TNT architecture can preserve local structure information inside the patch. The TNT model showed superior results in image classification, surpassing older vision transformer models. TNT integrates with DETR for object detection tasks. For instance, the results on

COCO val2017 showed that DETR+TNT-S outperformed DETR+PVT-Small by 3.5 AP with equivalent parameters. One of the challenges of TNT is to tune the model so that it provides optimal results for small datasets as well.

Wang et al.[9] (2021) introduced a Pyramid Vision Transformer(PVT)which, in contrast to ViT, is highly capable of doing dense prediction tasks. Training PVT on dense partitions of an image helps to produce high output resolution. Moreover, computations of large feature maps can be reduced with the use of a progressive shrinking pyramid in PVT. Pyramid Vision Transformers can be used as a backbone for dense prediction without convolutions, making it a replacement for CNN backbones. Tasks such as object detection, image classification and semantic segmentation were used by the authors to validate the performance of PVT. To generate feature maps of different scales, the proposed method used four stages, with each stage having similar encoding layers and a patch embedding layer. Experiments have been performed by the authors to compare PVT with ResNet and ResNeXt architecture. The ImageNet 2012 dataset was used in experiments of image classification tasks. The results show that with almost the same number of parameters and computational budgets, the PVT models outperformed conventional CNN backbones. Object detection experiments were conducted on the COCO benchmark. The effectiveness of the PVT backbone has been compared with RetinaNet and Mask R-CNN. The results showed the remarkable performance of PVT-based models over other models. For example, PVT-Tiny had 4.9 points more in Average Precision(AP) than the ResNet18. Similar results were also found while conducting experiments based on Mask R-CNN. The authors used the ADE20K dataset to validate the performance of PVT. Pyramid Vision Transformer backbones were evaluated based on Semantic FPN. Results show that ResNet or ResNeXt was completely outperformed by the PVT-based models. For instance, PVT-Tiny/Small/Medium achieved 2.8 more points than ResNet with almost the same number of parameters. In addition, pure transformer pipelines were built for experimenting with semantic segmentation and object detection. PVT was combined with DETR(a Transformer-based detection head) for object detection tasks and for semantic segmentation, PVT was combined with Trans2Seg [72], a Transformer-based segmentation head. Both these models showed very good performances compared to other conventional models like lResNet50-basedDETR and ResNet50-d8+DeeplabV3+, respectively, for object detection and semantic segmentation.

Han et al.[13] (2022) proposed a new model named PyramidTNT(TNT with pyramid architecture). In order to extract both local and global representations, the Transformer-in-Transformer (TNT) architecture utilises an inner transformer and an outer transformer. The authors here introduced two advanced designs to present new TNT baselines: the pyramid architecture and the convolutional stem. The pyramid architecture is used to extract multi-scale representations while the convolutional stem improves the patchify stem and stabilises the training process. The new PyramidTNT has established hierarchical representations, thus bringing significant improvement to the original TNT. An image classification experiment was carried out on the ImageNet-1K dataset, which consists of 1.28M and 50K training and validation images, respectively, with 1000 classes using PyramidTNT. Results

showed that PyramidTNT achieved higher accuracy in image classification compared to the TNT model. It has been observed that compared to TNT-S, PyramidTNT-S achieved 0.5% higher top-1 accuracy using 1.9B fewer FLOPs. The COCO 2017 benchmark was used in experiments related to object detection. Object detection frameworks used were RetinaNet, Cascade Mask R-CNN and Mask R-CNN. For instance, PyramidTNT-S-based RetinaNet surpassed the models with the Swin Transformer by 0.5 AP and 2.2 APL, respectively, indicating that the pyramid architecture of TNT can capture better global information for large objects. In addition, PyramidTNT-S surpasses Hire-MLP-S by 0.9 APb on Mask R-CNN with less FLOPs constraint.

Qin et al.[24] (2024) proposed a transformer-based model called Pyramid Fusion Transformer (PFT) to extract semantic information across the feature pyramid. This method is capable of extracting multi-scale information from the feature pyramid which helps to provide better results for semantic segmentation with higher accuracy. This efficiency is achieved through a novel cross-scale interquery attention mechanism. To validate the performance of the PFT model, the authors carried out experiments on the following segmentation datasets: COCO-Stuff-10K, ADE20K and PASCAL-Context dataset. For evaluation, Swin Transformer and ResNet (ResNet-50, ResNet-101 and ResNet-101c) were used as network backbones. Results of ADE20K showed that PFT has surpassed MaskFormer by 1.6 mIoU with the Swin-B backbone. The best achievement of the PFT model was 57.4 mIoU. For the COCO-Stuff-10K dataset, the PFT model outperformed the MaskFormer by 2.1 mIoU while achieving 52.2 mIoU with the Swin-L backbone. Results with the PASCAL-Context dataset showed that the PFT model outperformed other models by achieving 57.3 mIoU using the ResNet-101c backbone.

In this paper[25], Zhao et al. (2024) came up with a design called Multiscale Pyramid Transformer to overcome the limitations of traditional deep learning methods in image colourisation. Multiscale Pyramid Transformer is capable of capturing both local and global contextual information without using reference images. The dual-attention module and Colour-Attention decoder are two key aspects of this method for increasing its efficiency. The dual-attention module increases computational efficiency by reducing the attentional complexity to linear terms. On the other hand, the Colour-Attention decoder is capable of learning colourisation patterns from any type of image (grayscale/coloured image) and develops the model's ability to generalise. Datasets used for conducting experiments using the Multiscale Pyramid Transformer were ImageNet, COCO-Stuff, CelebA-HQ and ADE20K. Evaluation metrics used for the experiments were Frechet inception distance (FID) and Colourfulness Score (CF). It is observed from the experiments that the Multiscale Pyramid Transformer exhibits the lowest FID on all datasets, which signifies that this model can produce authentic and natural images. The authors also found that the ΔCF values of this model outperformed all other models, thus indicating that this model is capable of producing accurate images with realistic colour distribution. A user study was carried out with 95 volunteers participating it. Four other methods, alongside the Multiscale Pyramid Transformer, were used to apply colourisation

techniques to 50 images picked from ImageNet. The participants were then told to select random output images, and the majority of them picked the ones where the Multiscale Pyramid Transformer was used.

Architectures like CNN, especially U-Net, have successfully managed many medical image segmentation tasks before, but their failure in exhibiting long-range interactions and spatial dependencies has caused their performance drop in medical image segmentation with different shapes and structures. In this paper[10], Azad et al. (2022) came up with a novel idea for TransNorm, a segmentation framework that merges a Transformer module into both the encoder and skip-connections of the standard U-Net. TransNorm architecture integrates CNN and Transformer modules for improving semantic feature extraction. Two key features of this model are the Spatial Normalisation Mechanism and the Attention Mechanism. Spatial Normalisation Mechanism improves feature fusion and the attention mechanism helps to detect boundaries more precisely and increase segmentation accuracy. Experiments were carried out to validate the method on the following datasets: SYNAPSE MULTI-ORGAN SEGMENTATION dataset, SKIN LESION SEGMENTATION dataset and MULTIPLE MYLOMIA SEGMENTATION (MM) dataset. The performance metrics used for Synapse multi-organ segmentation were the average Dice Similarity Coefficient (DSC) and the average Hausdorff Distance (HD). For skin lesion segmentation and MM, performance evaluation metrics used were accuracy (AC), sensitivity (SE), specificity (SP), F1-Score, and mean Intersection over Union (mIoU). Results on the Synapse multi-organ segmentation dataset show the TransNorm model achieving the best result: 78.40% DSC and 30.25% HD. In the skin lesion segmentation experiment, the proposed model surpassed the SOTA approaches in every metric. This model shows improved generalisation performance and increases the F1 score by 8%, showing better results than U-Net.In MM segmentation, the TransNorm model is found to describe myeloma much more precisely and produce more fitting segmentation masks.

Liu et al.[19] (2023) initially proposed a hybrid neural network for medical image segmentation called FuzzyTransNet. For preprocessing medical images, a fuzzy neural network was developed using fuzzy logic supplemented by wavelet transformation to improve image quality. By applying multi-correlated convolution, multi-feature aggregation, and attention mechanisms, FuzzyTransNet accomplishes feature extraction. This model integrates Swin Transformer V2 with ResNet50 for extracting global semantic information without losing underlying features. For Skin Lesion Experiments, the ISIC-2017 and ISIC-2018 datasets were used. Evaluation criteria used here included Dice, mIoU, the Jaccard Index (JI) and accuracy (ACC). On the ISIC-2017 dataset, FuzzyTransNet outperformed U-Net by 3.6%, 3.9%, and 2.7% respectively in terms of the JI, Dice coefficient, and ACC. This model also surpassed other models with a much more complex ISIC-2018 dataset. For Polyp Segmentation, CVC-ColonDB, Kvasir, and CVC-linicDB datasets were used in the experiments. The evaluation metrics used here were Dice and mIoU. In the Kvasir and ClinicDB datasets, FuzzyTransNet achieved 1.1% and 0.7% higher Dice and mIoU values, respectively, compared to the TransFuse model. In addition, FuzzyTransNet showed improvement in the Dice Coefficient metric compared to the TransFuse model.

In clinical settings, marginal segmentation errors can cause a poor user experience. For example, the exclusion of the subtle speculation patterns around a nodule from segmentation masks can reduce model credibility from a clinical viewpoint. In addition, inaccurate segmentation, such as an error in the measurement of nodule growth, can lead to erroneous assignment of the Lung-RADS category. To address such issues, Zhou et al.[7](2018) proposed a new powerful architecture for medical image segmentation called UNet++. It consists of a deeply-supervised encoder-decoder network and is redesigned with a series of nested, dense skip pathways connecting the sub-networks. This model is capable of capturing fine-grained details of the foreground objects much more effectively with the help of redesigned skip connections in its architecture. The pattern of skip connections has made it different from the usual U-Net. The datasets used for experiments in this paper are the cell nuclei dataset, colon polyp dataset, liver dataset, and lung nodule dataset. U-Net and a customized wide U-Net architecture were used to compare with the UNet++ model. Dice coefficient and IoU were used as evaluation metrics in the experiments along with the Adam optimizer with a relatively low learning rate of 3e-4. Results show that UNet++ outperformed U-Net and wide U-Net, generating an average IoU gain of 3.9 and 3.4 points respectively over those two models. It is also mentioned in this paper that pruning the UNet++ model reduced the interference time but also caused model accuracy to decrease significantly at the same time.

In a paper, Hussain et al.[14](2022) wrote about how deep learning-based image processing models have contributed to the development of Robot-Assisted Surgery(RAS). Intensive pre-processing and feature engineering were time-consuming, laborious and cost-intensive with previous models but DL models have made these less tedious. Due to the high success rate of DL models in CAD systems, the idea of using these models in robotic surgery was proposed. Here, the authors carried out the survey and categorized the use of DL models into four segments: Surgical Tools, Surgical Processes, Surgical Surveillance, and Surgical Performance. The category of surgical tools is further divided into two categories: tool detection and tool segmentation. Surgical tool detection includes detecting or locating any tool during surgery. In surgical tool detection, the most applied DL method was CNN followed by LSTM, RNN and autoencoder architectures. It has been found that CNN models with different variants(few architectural modifications) performed better in most tool detection cases. For the task of tool detection, the datasets used were The Endoscopic Vision (EndoVis) challenge and m2cai16-tool datasets in addition to ATLAS Dione. Performance evaluation parameters most widely used for this purpose were accuracy, precision and Area Under the Curve(AUC). On the other hand, surgical tool segmentation differentiates tools from organs or other tools. In tool segmentation, the most frequently used dataset is the EndoVis robotic instrument segmentation challenge dataset with the Dice score used as the most common performance metric. The surgical process section is subdivided into Gesture Recognition, Trajectory Segmentation, and Tissue Segmentation. For gesture recognition, widely applied DL methods were LSTM, CNN and RCN. Common datasets used for this segment included the JIGSAWS dataset and other in-house datasets with accuracy being the major performance metric along with several other metrics. Results show better accuracy for datasets that had a fusion of video and kinematic data compared to

datasets with either image or video data. Trajectory segmentation involved motion analysis and pattern recognition of the surgical tools used. Kinematic data was incorporated with video data to improve the result of segmentation. It has been found that combining different DL model architectures such as CNN and LSTM produced better performance output. Tissue segmentation comprises all the studies performed on tissues including vessel segmentation, edge detection, healthy and cancerous tissue classification, uncertainty inference segmentation, and tissue retraction. In addition, binary classification of healthy and cancerous tissues was also included. Commonly used architectures used for this purpose included U-Net, LSTM, GAN and other CNN variants. Mostly in-house data were used due to a lack of large-scale public datasets. Accuracy, Intersection over Union (IoU), Dice and AUC were used as performance metrics. Surgical surveillance is subdivided into surgical planning, phase and state estimation and activity recognition. In surgical planning category, mostly in-house datasets were used which included CT scan images and MRI images. Dice had been used as a performance parameter in most cases. In the phase and state estimation category, frequently used DL methods are CNN and LSTM with accuracy used as the most dominant performance metric. ImageNet data was used to train DL models in the activity recognition category with accuracy as a performance parameter. For this category, CNN and LSTM were integrated for better results. Finally, the surgical performance category was divided into skill assessment and workflow recognition categories. For the skill assessment part, the CNN model was used with JIGSAWS being the most common dataset and accuracy and AUC were used for performance evaluation. For workflow recognition, LSTM and CNN were the commonly used methods with accuracy as the performance metric. Several technical and ethical challenges were proposed in using image-driven deep learning methods in robot-assisted surgery. Some technical challenges included the limited availability of datasets, the black-box nature of DL methods and image artefacts. Ethical challenges included cybersecurity, patient data privacy, and legal frameworks.

In this paper, Kirillov et al.[18](2023) came up with a foundational model for image segmentation called the Segment Anything Model(SAM). It has three components: an image encoder, a flexible prompt encoder, and a fast mask decoder. Due to higher efficiency and better runtime performance, SAM can be used in real-time interactive tasks. The limitation of segmentation masks on the internet has provoked the authors to build a data engine for the collection of the 1.1B mask dataset, SA-1 B. In the SA-1B dataset, there are 11M diverse, high-resolution, licensed, and privacy-protecting images and 1.1B high-quality segmentation masks collected with the data engine. The images of this dataset have higher resolution than images of some renowned datasets, such as the COCO dataset. Experiments were carried out to determine the mask quality produced by the data engine, and the results confirmed that the masks are better than masks of some currently existing datasets.SAM uses an MAE pre-trained ViT-H image encoder, and it is trained on the SA-1B dataset. Zero-shot transfer experiments were carried out with SAM. SAM was applied to newly compiled suites of 23 datasets. Results showed that SAM yielded higher performance on 16 out of 23 datasets. RITM was also used to make a comparison with SAM, and it was observed that SAM outperformed RITM on all

datasets. There still exist some limitations of SAM. It can miss fine structures and cannot always produce boundaries with high accuracy.

The paper named 'Segment anything in medical images[22] explores the Segment Anything Model in the medical domain to capture a wide variety of anatomical structures and lesions across diverse medical imaging modalities, and demonstrates substantial capabilities. To fine-tune the Segment Anything Model, authors used a promptable 2D segmentation approach as well as the network architecture includes an image encoder, a prompt encoder, and a mask decoder. The architecture has been trained on a diverse dataset of 1,570,263 medical image-mask pairs from 10 modalities and evaluated on various segmentation tasks. MedSAM obtained median DSC scores of 87.8% on the nasopharynx cancer segmentation task, demonstrating 52.3%, 15.5%, and 22.7% improvements over SAM, U-Net, and DeepLabV3+, respectively. MedSAM achieves state-of-the-art performance on various segmentation tasks by outperforming specialist models like Unet and DeepLabV3+, and achieves consistent performance across both internal and external validation sets.

Martinez-Escobar et al.[1] (2012) proposed the idea of colourising grayscale CT scan images before performing segmentation on them so that the contrast between tumour and healthy tissues was more prominent. The methodology involved mapping tissue densities into an RGB colour space based on HU values. Segmentation techniques like thresholding provided better results with colourised images than with normal grayscale images, greatly decreasing the number of false positives and false negatives. Ten different datasets were used to test this colourisation method: Cystic Teratoma on the left ovary, immature Teratoma, Mucinous Cystadenoma on the right ovary, and Neuroblastomas. These datasets were divided into three categories based on their characteristics. Category A included tumours having similar densities, category B contained inhomogeneous tumours with fuzzy edges, and Category C had heterogeneous tissues. Accuracy was measured based on the values of False positives and False negatives. The datasets were segmented with both colourisation and without colourisation. A t-test was carried out on both grayscale and colourised image datasets. Colourised images showed better results than the grayscale images. For category B, a grayscale method was incapable of detecting tumours with fuzzy edges, but the coloured method was capable of segmenting most of the actual tumours. Even with category C, which contained different pixel values of tumour and healthy tissues, the grayscale method did not work well. On the other hand, the colourisation technique showed significant improvement even in the real-time adjustment of the tissue density range from the region of interest. Finally, the authors suggested some future improvements that can be added to make the colourisation method much more efficient. Some of these include using advanced segmentation algorithms like the probabilistic method algorithm and intensity inhomogeneity correction (IIC) to improve colourisation mapping.

Previously, laparoscopic image datasets had limited usage. They were mostly used for general data classification and semantic segmentation of surgical instruments.

Carstens et al.[17] (2023) came up with the Dresden Surgical Anatomy Dataset that can be used for abdominal organ segmentation in Surgical Data Science. This dataset is comprised of 13195 laparoscopic images with detailed segmentations of multiple anatomical structures and thus ML can use it for tasks like organ recognition. Images used in this dataset were taken from 32 robot-assisted rectal resections and included annotations like weak labels to make the organ images more visible. The Staple algorithm was used for merging annotations and performance metrics used for comparison were the F1 score and the Hausdorff distance. Results showed a high F1 score and low value for Hausdorff distance, which indicates there were no over- or under-segmentations. It has applications in the development of AI-based surgical assistance systems with enhanced safety and precision. There is scope for further research on integrating such datasets with real-time AI applications to produce better surgical results.

The paper 'InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions propose a new large-scale CNN-based[20] foundational model that achieves comparable performance to existing state-of-the-art ViTs and CNNs on various vision tasks, including Image Classification, Object Detection and segmentation by introducing deformable convolution and adaptive spatial aggregation. The architecture extended the existing DCNv2 operator, called DCNv3 and scaled up to large parameters and data, as a result, achieved fewer training data, shorter training time and improved the efficiency of the architecture. Authors explored scaling rules for their architecture to achieve state-of-the-art performance and designed a new block which includes the DCNv3 operator, layer normalisation and GELU as an activation function. InternImage-T achieves 83.5% top-1 accuracy, outperforming ConvNeXt-T, and InternImage-B surpasses the hybridViT CoAtNet-2 by 0.8 points. Additionally, InternImage-H achieves 89.6% top-1 accuracy, which is better than any other state-of-the-art CNN architectures and ViTs. Moreover, IntrernImage-H outperformed Mask2Former in semantic segmentation, achieving mIOU of 62.9. This paper takes feature extraction to the next level, which makes it a state-of-the-art architecture as the backbone of other multimodal architectures.

This paper presents a novel and effective SAR (Synthetic Aperture Radar) ship detection method with Swin Transformer as its backbone and introduces an enhanced Feature Pyramid Network (FEFPN)[15]. The main objective of this work is to overcome the limitations of CNN-based approaches such as VGG-16 and ResNet-50, which have difficulty in designing long-term models and often cannot draw specific models. This limitation shows the best performance, especially for small ships with complex histories. To address these issues, the authors proposed the Swin Transformer, which is good at capturing long-term expectations due to the existence of a network of self-tracking and parallel numbers with irregular quadratic complexity of the Transformer model. Furthermore, FEFPN aims to transfer semantic information by multiple top-down methods with gradually decreasing weights, thus improving the quality of maps, especially in shallow layers. The design is based on two key insights: The semi-translatable structure of Swin Transformer, which makes it better at capturing semantic information rather than geometric details and evolution.

CNN. These experiments are conducted using the widely used SAR Ship Detection Dataset (SSDD), which is split into training and testing in a ratio of 8:2. The performance of this method is measured using the mean probability (mAP) calculated based on the ground truth and inverse. Precision measures the accuracy of detected vessels while repeatability measures the ability to detect the entire ground truth. The proposed method is compared with various state-of-the-art models such as Faster R-CNN, RetinaNet, DCN (Deformable Convolutional Network) and PANet which use CNN-based neural networks. The results show that the proposed method outperforms all other comparable models, achieving the highest mAP of 93.08%. For example, DCN achieves a mAP of 92.60%, while Faster RCNN and RetinaNet achieve 90.30% and 87.10%, respectively. The results show that mAP can be improved by 2.21% compared to ResNet 50 by using only Swin Transformer, proving its effectiveness in long-range modelling. Furthermore, FEFPN adds another 0.57% improvement, demonstrating its ability to improve the graphs, especially for small vessels. Although this method is powerful, it also has some limitations. For example, it is based on a two-stage detection pipeline, which may involve lower costs compared to single-stage detectors such as RetinaNet. In addition, the experiments were limited to the SSDD dataset, and the generalisation of the approach to other SAR data or multidimensional data (e.g., combining SAR with optical images) is not yet available. This paper does not discuss the computational complexity or the rapid thinking of the plan, which is important for emergency use. It has good potential for ship research due to its integrated structure, which can be explored in future studies. Further research can focus on optimising the process, especially for emergency applications, and extend the evaluation to increase the power and efficiency of SAR. Overall, this paper significantly contributes to ship SAR research by combining new architectural options with robust experimental results, paving the way for future improvements in this field.

The paper titled "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks"[6] by Mingxing Tan and Quoc V. Le introduces an innovative method for scaling Convolutional Neural Networks (ConvNets). The authors highlight that traditional scaling techniques, which adjust depth, width, or resolution independently, are not the most effective. Instead, they propose a unified scaling strategy that optimises all three dimensions simultaneously through a simple compound coefficient. This balanced approach significantly improves performance. The study introduces a new series of models known as EfficientNets, created using neural architecture search and scaled with the proposed method. These models achieve state-of-the-art performance on ImageNet and other datasets while being more compact and faster than previous architectures. For instance, EfficientNet-B7 reaches 84.3% top-1 accuracy on ImageNet, outperforming the previous leading model, GPipe, while being 8.4 times smaller and 6.1 times quicker. The most lightweight model, EfficientNet-B0, delivers 77.1% top-1 accuracy with just 5.3 million parameters, showcasing impressive efficiency. The core contribution of the paper is the compound scaling technique, which adjusts depth, width, and resolution together using fixed ratios. This approach yields higher accuracy and efficiency compared to traditional scaling methods. The baseline model, EfficientNet-B0, is designed through neural architecture search and then scaled to produce models ranging from EfficientNet-B1 to B7.

These models are fine-tuned for both accuracy and computational efficiency, achieving outstanding results on ImageNet and transfer learning datasets. EfficientNets demonstrate top performance on five out of eight benchmark datasets, including CIFAR-100 (91.7%) and Flowers (98.8%), while using significantly fewer parameters. In performance comparisons, EfficientNets surpass well-known models such as ResNet, DenseNet, Inception, NASNet, AmoebaNet, and GPipe. For example, EfficientNet-B3 outperforms ResNeXt-101 with 18 times fewer FLOPS. The experiments focus primarily on the ImageNet dataset, with additional evaluations on datasets like CIFAR-10, CIFAR-100, Birdsnap, Stanford Cars, Flowers, FGVC Aircraft, Oxford-IIIT Pets, and Food-101. Key performance indicators include top-1 and top-5 accuracy, alongside measures like parameter count and FLOPS to assess efficiency. The authors also report inference latency on CPUs, showing that EfficientNet-B1 runs 5.7 times faster than ResNet-152, while EfficientNet-B7 is 6.1 times faster than GPipe. While the paper presents strong results, there are certain limitations. The compound scaling method requires a small grid search to identify optimal scaling coefficients (α, β, γ), adding some complexity to the process. Additionally, the focus is mainly on image classification tasks, with limited evaluation of object detection or segmentation. Since the models are tested on specific hardware, performance may differ on other devices like GPUs. Future work could explore applying the compound scaling method to other architectures, such as transformers or recurrent neural networks, and refining scaling coefficients using more advanced optimisation techniques. Expanding the application of EfficientNets to areas like object detection and semantic segmentation also presents promising research opportunities. In summary, this paper introduces a significant advancement in ConvNet scaling through the compound scaling method and the EfficientNet model family. The authors demonstrate that a balanced scaling approach across depth, width, and resolution results in models that are both highly accurate and computationally efficient. EfficientNets set new performance standards on ImageNet and various transfer learning datasets, outperforming existing models in accuracy, parameter efficiency, and speed. Despite some limitations, the study paves the way for future research in efficient neural network design and model scaling.

The Retinexformer[16] introduces a novel approach to low-light image enhancement, addressing limitations in traditional Retinex-based models and convolutional neural networks (CNNs). By proposing a One-stage Retinex-based Framework (ORF) integrated with an Illumination-Guided Transformer (IGT), the model significantly enhances image quality while efficiently managing computational resources. The framework aims to estimate illumination, rectify corruptions, and improve visibility in low-light conditions, outperforming state-of-the-art methods. Retinexformer simplifies the image enhancement process by eliminating the need for multi-stage pipelines through the ORF, which estimates illumination to brighten images and restores corrupted regions in a unified manner. The IGT enhances the model's ability to capture long-range dependencies and manage non-local interactions across different lighting conditions, with its core component, Illumination-Guided Multi-head Self-Attention (IG-MSA), leveraging illumination information for precise self-attention computations. The end-to-end training process is streamlined, reducing computational overhead while maintaining high performance. Robust corruption

modelling is achieved by incorporating perturbation terms to address noise, artefacts, and colour distortions inherent in low-light images and introduced during enhancement. Comprehensive experiments were conducted to validate the efficacy of Retinexformer. The model was tested on thirteen benchmarks, including LOL-v1, LOL-v2 (real and synthetic subsets), SID, SMID, SDSD (indoor and outdoor), and FiveK, with additional datasets such as LIME, NPE, MEF, DICM, and VV used for qualitative analysis. Evaluation metrics included Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), alongside qualitative assessments and user studies. Retinexformer was compared against leading models like DeepUPE, RetinexNet, KinD, RUAS, SNR-Net, MIRNet, and Restormer, consistently outperforming competitors and demonstrating superior enhancement quality and efficiency. Retinexformer achieved remarkable improvements across various datasets, with PSNR improvements of over 6 dB on challenging datasets like SID and SDSD, significantly surpassing previous state-of-the-art methods. Visual results highlighted enhanced brightness, reduced noise, and preserved colour fidelity, outperforming models that either introduced artefacts or failed to correct exposure issues. High user satisfaction scores and improved object detection accuracy in low-light conditions underscored the model's practical benefits. While Retinexformer sets new benchmarks in low-light image enhancement, potential limitations include computational costs, as the Transformer architecture may still pose challenges for real-time applications on resource-constrained devices, and the need for enhanced performance in extremely adverse lighting environments and across diverse image types. Future work could focus on refining the model's efficiency, exploring real-time deployment scenarios, and extending its applicability to other computer vision tasks like video enhancement and autonomous driving systems.

In this paper, Chen et al.[4] present a seminal deep learning framework for semantic segmentation called DeepLab. In order to control the resolution of computed features without increasing the number of parameters or computational cost, the authors introduced a technique called atrous convolution. This method lets the model see a larger part of the image at once while still keeping the details sharp and clear. Next, the authors proposed Atrous Spatial Pyramid Pooling (ASPP), which is an efficient alternative to multi-scale image processing. ASPP applies multiple atrous convolutions with different dilation rates in parallel, enabling robust multi-scale object recognition. Furthermore, the authors augmented DCNN outputs with a dense CRF to improve object boundary accuracy. The CRF is particularly effective in smoothing edges of objects. This framework was evaluated using different datasets. It set a new benchmark with the PASCAL VOC 2012 dataset by achieving 79.7% mean IoU. With other datasets like PASCAL-Context, PASCAL-Person-Part, and Cityscapes, DeepLab demonstrated strong generalisation with significant performance gains over prior work. The architecture of the DeepLab models was set up based on both VGG-16 and ResNet-101. ResNet-101 provided better localisation and segmentation accuracy due to its deeper architecture and identity mappings. DeepLab has improved spatial resolution by replacing standard max-pooling with atrous convolutions. In addition, CRFs led to 3-5% improvement in segmentation accuracy across models. Despite its benefits, DeepLab still struggles with precise boundary segmentation in very fine structures such as bicycles and chairs. As part

of future improvement, the authors suggested the use of encoder-decoder architectures to recover finer details.

In this paper, Oquab et al.[23] proposed a scalable self-supervised framework which can learn general-purpose visual features from images by itself, unlike other weakly supervised architectures. It can be used in different types of tasks such as classification, segmentation, and depth estimation using the "frozen features" produced by itself. For experimental purposes, the authors created a dedicated dataset called LVD-142M, which consisted of 142 million curated images. DINOv2 is trained with different Vision Transformers (ViT) architectures and applied on different datasets. Results show that DINOv2 ViT-g achieves 86.5%, surpassing state-of-the-art self-supervised (iBOT) and even weakly-supervised (OpenCLIP) models. DINOv2 outperforms all prior self-supervised models on ImageNet-A, -R, and -Sketch. DINOv2 matches or exceeds prior state-of-the-art using only frozen features and simple linear heads on tasks like semantic segmentation (ADE20k) and monocular depth estimation (NYUd, KITTI). In addition, it shows massive gains in instance-level benchmarks (Oxford, Paris) with +30–40% mAP over iBOT and OpenCLIP.
In this paper, Cheng et al. [12] presented a new unified transformer-based architecture called Masked-attention Mask Transformer (Mask2Former) that is capable of addressing any image segmentation task (panoptic, instance, or semantic). Mask2Former achieves state-of-the-art (SOTA) results across all these tasks while being more efficient and flexible. It consists of a CNN or ViT backbone for extracting low-resolution features, a Pixel Decoder for upsampling low-resolution features to high-resolution spatial features, and a Transformer Decoder for refining object queries using attention mechanisms and producing masks. The decoder processes a fixed number of object queries and outputs binary masks and class labels. It consists of multiple layers, where each layer contains Masked Cross-Attention, Self-Attention, Feed Forward Network (FFN), and LayerNorm & Residual Connections. Each decoder layer updates the queries based on image features and mask information. A novel form of cross-attention in the Transformer decoder, where attention is restricted to the foreground region of each mask prediction, improves both convergence speed and accuracy. For reducing computational cost and balancing accuracy with efficiency, features are fed in a round-robin manner across decoder layers. On the COCO panoptic dataset, it achieves a state-of-the-art 57.8 PQ, outperforming the previous best model by over 5 points, along with higher AP and mIoU scores. For instance segmentation, Mask2Former reaches 50.1 AP on COCO val2017, outperforming leading models like HTC++ and significantly improving boundary quality, especially on large objects. In semantic segmentation, it attains 57.7 mIoU on ADE20K, beating previous specialised models such as BEiT and FaPN-MaskFormer.
Jenke et al. [21] proposed a novel method to train a multi-class organ segmentation model for laparoscopic surgery. The authors combined multiple complementary, partially annotated datasets for this experiment. Lack of fully annotated datasets is the main challenge in surgical scene segmentation. To address this issue, the authors introduced an approach called *Implicit Labelling* (IL). It leverages the idea of mutual exclusivity in anatomical segmentation: if a pixel is annotated as belonging to one class, it is implicitly not part of any other class. This principle treats positive labels of one class as negative samples for all other classes and ignores background pixels

that might contain other unlabelled structures. The Dresden Surgical Anatomy Dataset (DSAD), which contains binary segmentations for individual anatomical structures, was used to test this method. The authors trained a DeepLabV3 model using their IL method and compared it against Ensemble (EN) and Fully Supervised (FS) baselines. Results show that IL outperformed the EN approach by up to 7% in Dice score. Despite using only partially labelled data, IL showed similar performance to FS. IL reduced confusion between anatomically similar classes like stomach and colon. It helps to solve the problem of not having enough fully labelled data in surgical videos and offers a faster, more efficient way to train models compared to older methods.

2.3 Summary of Key Findings

We have found some research gaps in the papers we have covered. We have found only one model[21] that performs semantic segmentation for abdominal organs, where they only addressed 6 anatomical structures. Most existing transformer models perform well on generic segmentation but are not optimised for surgical settings with multiple organs. Models like U-Net[2] and DeepLab[4] struggle with small or overlapping anatomical structures due to their limited receptive field and poor context modelling. Moreover, Dinov2[23] architecture is perfectly outlined for extensive spatial features extraction using the self-attention mechanism, and since it is self-supervised hence it's completely free from class bias. For the decoder layer, we have found the mask2former[12] as the current state-of-the-art model, although architecturally it requires high computation, as a result, we have modified some modules for our use case. Finally, we have come up with a novel hierarchical-transformer-based architecture for semantic segmentation of Intraoperative Anatomical Structures to fulfil some of these research gaps.

Chapter 3

Methodology

3.1 Methodology Overview

Top level overview of the proposed system

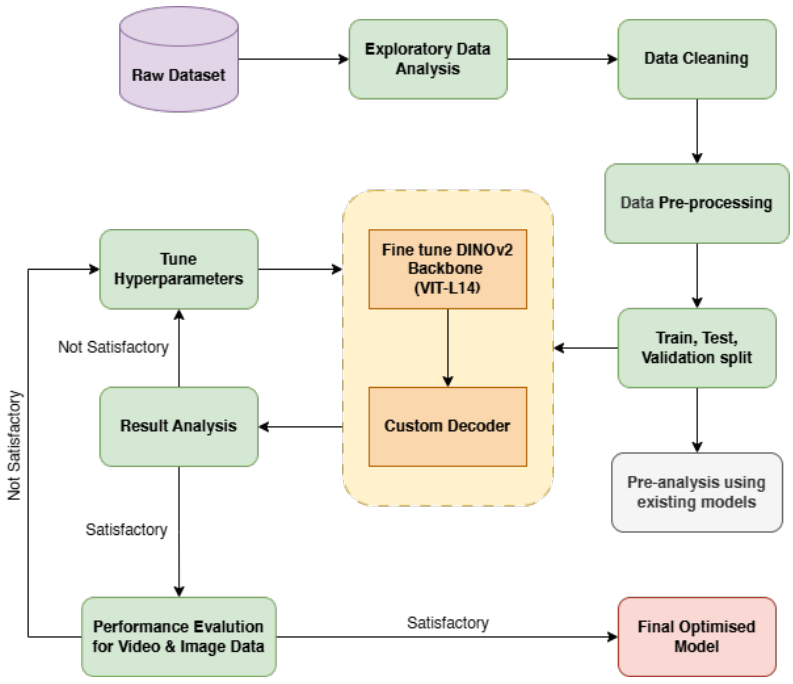

Figure 3.1: Top level overview of the proposed system

We first collected the appropriate dataset from the secondary source, named The Dresden Surgical Anatomy Dataset for Abdominal Organ Segmentation in Surgical Data Science[17] and performed exploratory data analysis on the raw dataset, followed by Data Cleaning and Data Pre-processing. Later, we split the data into training, validation, and testing sets and conducted preliminary analysis using existing models. The core of our model is built by fine-tuning the DINOv2[23] backbone and integrating it with a custom decoder. During training, results are analysed, and if found unsatisfactory, the hyperparameters are tuned and the process is repeated. Once satisfactory results are obtained, a performance evaluation is conducted. If the results are satisfactory, the final optimised model is delivered for future works.

3.2 Data Preparation

3.2.1 Data Collection

We have selected **The Dresden Surgical Anatomy Dataset for Abdominal Organ Segmentation in Surgical Data Science**, which is the secondary source of data. The Dresden Surgical Anatomy dataset[17] is a specialised dataset for applications of surgical data science that particularly focuses on the segmentation of abdominal organs. According to the paper, this dataset includes semantic segmentation for 11 abdominal organs, such as the pancreas, colon, spleen, liver, two vascular structures, small intestine, ureter, abdominal wall, stomach, and vesicular glands as seen in laparoscopic views. The dataset was collected from 32 surgeries where all patients had clinical reasons for surgical procedures. Patients had a mean BMI of 26.75 kg/m^2, and the average age was 63 years. Using a Da Vinci® Xi/X Endoscope with an $8mm$ diameter, 30 angled camera (Intuitive Surgical, Item code 470057)[17], surgeries were conducted and recorded in MPEG-4 format at 1920×1080 pixel resolution, with each surgery lasting between two and ten hours. With two years of experience in robot-assisted rectal surgery (MC, FMR), a medical student used the Surgery Workflow Toolbox version 2.2.0 to annotate the surgical processes. Videos from at least 20 surgeries were selected for each anatomical structure, with up to 100 equidistant frames randomly chosen per organ, ensuring diversity in data. Consequently, the dataset contains at least $1,000$ annotated images for each organ or structure, covering at least 20 patients. We are using this dataset as our primary dataset for this research.

3.2.2 Exploratory Data Analysis

The data is organised in a 3-level folder structure[17]. The first level consists of twelve subfolders, one for each organ or anatomical structure (colon, pancreas, abdominal_wall, inferior_mesenteric_artery, intestinal_veins, stomach, liver, small_intestine, spleen, vesicular_glands, and ureter), and one for the multi-organ dataset (multilabel). Each folder contains 20 to 23 subfolders corresponding to the different surgeries from which the images were extracted, with subfolder names derived from the individual index number of each surgery. These folders contain two versions of 5 to 91 PNG files: one raw image extracted from the surgery video file and one image with the mask of the expert-reviewed semantic segmentation (black for the background and white for the segmentation). The raw images are named imagenumber.png (e.g., image23.png), while the masks are named masknumber.png (e.g., mask23.png). In the multilabel folder, there are separate masks for each of the considered structures visible in the individual image (e.g., masknumber_stomach.png). The image indices always match for associated raw images and masks. Each surgery- and organ-specific folder also contains a CSV file named weak_labels.csv, which includes information about the visibility of the eleven regarded organs in the respective images. The columns in these CSV files are ordered alphabetically: Abdominal wall, colon, inferior mesenteric artery, intestinal veins, liver, pancreas, small intestine, spleen, stomach, ureter, and vesicular glands.
In our dataset, we found the class imbalance in folders containing different organ images, which was balanced during data augmentation. In addition, data inconsistencies were present in our dataset. For instance, there were supposed to be 32

surgery folders present according to the referred dataset paper, but in the actual dataset, some of the surgery folders were missing. As a result, we got fewer images than the number of images mentioned in the dataset paper. Moreover, an organ-wise folder was present in the dataset folder, and for each organ image, there was one mask. The only exception was for the case of the stomach: there were multiple abdominal organs present in one image, and so a separate multimodal folder was present in the dataset for this case. For each image in the multimodal folder, 7 organ masks were present along with an annotation, and we had to handle this part in our data pre-processing. An image of masking is provided below:

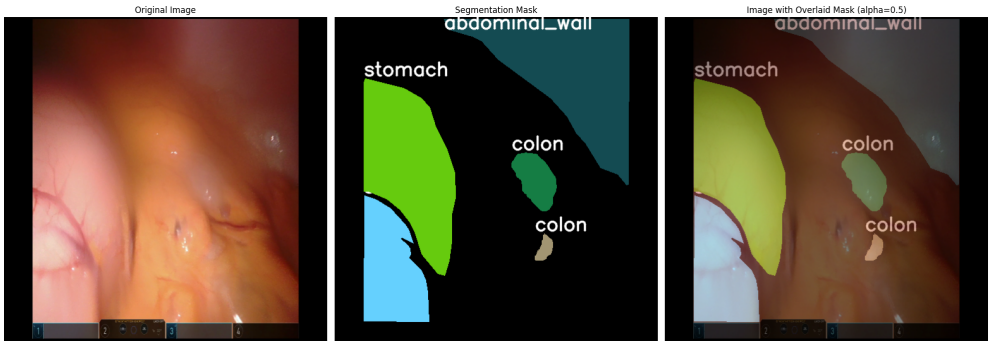

Figure 3.2: Dynamic visualisation of annotation mask and ground truth image

3.3 Data Preprocessing

3.3.1 Data Cleaning

In our dataset, three different binary masks were present, contributed by three different surgeons, and one merged mask folder was present that had been created by another surgeon merging those three binary mask folders. We used only the merged mask folder in our experiment. Data augmentation was performed for some specific organs and multilevel images. We have applied some data preprocessing techniques to resolve some of the issues of our dataset. During initial training, the model struggled to segment organs occupying smaller areas, while it performed better on larger organs. Hence, we decided to extract the ROI of the smaller area occupied organs and load in into the training so that our model could learn more effortlessly. We set a threshold (0.001% - 10%) for pixel percentage covered by the organ. If the annotated pixel value is within this range, we create a square box and take its ROI. Let's say an organ only takes up about 4% of the full image (which is 1024x1024 in size). We take the ROI of the area around the organ and crop it into a smaller 512x512 image. In this smaller image, the organ now fills about 25% to 35% of the space. Then we scale the image back up to the original size. From our analysis, we noticed that even the biggest organs usually don't take up more than 35% of the full image. In case of images containing multiple organs, we follow the same procedure, but now the number of images extracted after ROI is equal to the number of organs present in the original image. For instance, if the original image contained two different organs, then we will get two separate images for those two organs after ROI extraction. All the images (original and extracted ones) are kept in the dataset folder.

3.3.2 Data Transformation

The original image resolution was 1920×1080. To ensure consistency in input size, all images and their corresponding masks were resized to 1024×1024. This was achieved by first applying zero-padding to convert the rectangular images into square dimensions without cropping any part of the content. Following this, a linear interpolation method was used for resizing. To enhance generalisation, supervised data augmentation techniques were applied. The images were also normalised to be compatible with the requirements of the proposed architecture. For multi-level data augmentation, a variety of transformations were used, including rotations of $+30°$ and $-30°$, horizontal and vertical flips, random brightness and contrast adjustments, Gaussian noise injection, elastic deformation, grid distortion, random gamma correction, and combined shift-scale-rotate operations.

3.3.3 Annotations

To maintain consistency and compatibility with widely accepted benchmarks, we adopted the **PASCAL VOC 2012** annotation format for our semantic segmentation task. This standard is extensively used in computer vision research, particularly in video-related tasks. Given our goal of achieving real-time segmentation, PASCAL VOC 2012 was deemed a suitable annotation standard for our dataset.

The images in our dataset were annotated into 12 distinct classes, each represented by a unique RGB value. The classes and their corresponding RGB encodings are as follows: Background ($[0, 0, 0]$), Abdominal wall ($[128, 0, 0]$), Colon ($[0, 128, 0]$), Inferior mesenteric artery ($[0, 0, 128]$), Intestinal veins ($[128, 128, 0]$), Liver ($[128, 0, 128]$), Pancreas ($[0, 128, 128]$), Small intestine ($[255, 0, 0]$), Spleen ($[0, 255, 0]$), Stomach ($[0, 0, 255]$), Ureter ($[255, 255, 0]$), and Vesicular glands ($[255, 0, 255]$). These RGB codes allow for precise pixel-level semantic segmentation of each anatomical structure.

Figure 3.3: Pascal VOC 2012 Annotation, where each RGB values indicate a distinguished class

3.4 Architecture Design

3.4.1 Pre-analysis of Existing Architectures

To form an ideal multi-modal architecture, we have done some pre-analysis using various existing models of instance segmentations and image enhancement. For semantic segmentation, we have used Mask2Former[12], OneFormer, Retinexformer and Yolov11 with SAM2**sam**. We fine-tuned these architectures using our primary dataset[17].

Rationale Behind Existing Model Selection:

- **Mask2Former and OneFormer:** These transformer-based models capture global context efficiently and deliver high-precision segmentation, which is critical for identifying complex anatomical structures.
- **YOLOv11 with SAM2:** Prioritises real-time performance, ensuring that the system can be effectively deployed in a surgical setting where timely feedback is essential.
- **Retinexformer:** Used for analysing the image enhancement under challenging lighting conditions and in areas with intricate boundaries.

Key Observations: The Mask2Former model, implemented using the Detectron2 framework, achieved a mean Average Precision (mAP) of **0.47**. Its transformer-based segmentation head effectively delineated surgical instruments and anatomical structures, particularly excelling in scenarios with overlapping objects. This makes it suitable for the complex visual scenes typical in surgical environments.
OneFormer demonstrated a competitive performance with an mAP of **0.49**. Although slightly lower than Mask2Former, OneFormer offers a unified architecture that can handle both segmentation and detection tasks simultaneously. This unified approach is particularly advantageous when dealing with the multi-class challenges presented by the surgical dataset.
The YOLOv11 model, enhanced with the SAM2 module, achieved an mAP of **0.44**. Despite a modest decrease in accuracy compared to the transformer-based models, YOLOv11 with SAM2 boasts real-time inference capabilities. Its faster processing speed makes it an attractive option for intraoperative applications where low latency is crucial.
By leveraging multi-scale feature representations and advanced mask refinement techniques, Retinexformer significantly improved boundary delineation and overall segmentation quality. Its robust performance under varying lighting conditions is especially beneficial for the dynamic environment of surgical procedures.

This analysis allowed us to harness the strengths of each architecture and helped us to design a robust and versatile architecture for surgical instance segmentation and enhancement.

3.4.2 Model Architecture

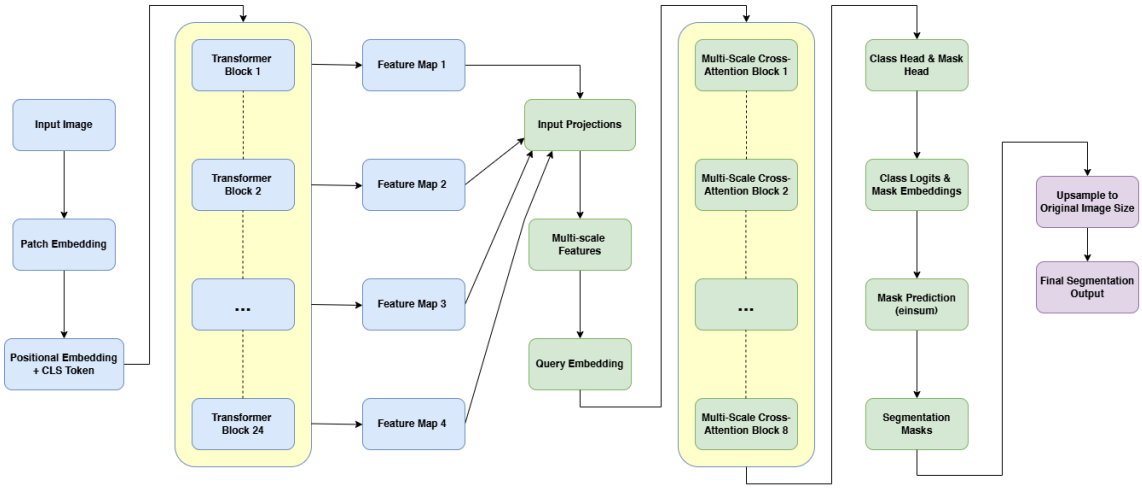

Figure 3.4: Model Architecture

1. **Input**: Our architecture takes input image as shape of (Channels, Height, Width) where for RGB images, total channel is 3, height and width are 1022. Our default image size is 1024×1024, but to match the Vit-L patch size, we have to resize the image to the closest dimension which is a multiple of 14, hence take 1022 as our default dimensions.

2. **Backbone Feature Extraction**: The process begins with **Patch Embedding**, where the input RGB image is divided into small, non-overlapping square patches. Using the ViT-L/14 configuration, each patch is of size 14×14, and since the input image resolution is 1022×1022, this results in 73 patches along each side, totalling $73 \times 73 = 5329$ patches. Each patch, having a shape of $3 \times 14 \times 14$, is flattened into a $3 \times 14 \times 14 = 588$ dimensional tensor and then projected linearly into a 1024-dimensional embedding space, where 1024 is the default embedding dimension of the Vit-L/14. The output of this stage is a patch embedding tensor of shape $\mathbb{R}^{5329 \times 1024}$. Patch embedding transforms the image into a sequence of patch vectors that can be effectively processed by subsequent transformer modules.

   Next, in the **Positional Embedding** stage, spatial context is added into the model by adding learnable positional encodings. A special classification token CLS_token $\in \mathbb{R}^{1 \times 1024}$, intended to capture global image semantics, is prepended to each patch sequence. Concatenating this with the patch embeddings forms a sequence of $(5329 + 1) = 5330$ tokens per image, giving a tensor of shape $\mathbb{R}^{5330 \times 1024}$. A learnable positional embedding of shape $\mathbb{R}^{5330 \times 1024}$ is then added to this sequence. This step compensates for the lack of inherent positional awareness in transformers, ensuring the model understands where each patch is located within the image.

   The enriched token sequence is subsequently passed through a **Stack of Transformer Blocks**, consisting of 24 identical attention blocks. Each block contains two sub-modules: a Multi-Head Self-Attention (MHSA) mechanism and a Multi-Layer Perceptron (MLP) as the feed-forward network. Initially, Layer

Normalisation is applied to the input tensor of shape $\mathbb{R}^{5330\times1024}$. The MHSA layer computes the query (Q), key (K), and value (V) projections of the same shape as the input. These are divided across 16 heads, the default number in ViT-L/14, with each head having a dimensionality of $d_k = \frac{1024}{16} = 64$, where 1024 is the default embedding dimension and 16 is the number of heads in Vit-L/14. The self-attention operation for each head follows the equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{64}}\right)V \tag{3.1}$$

which allows each token to attend to every other token, capturing both local and global context.

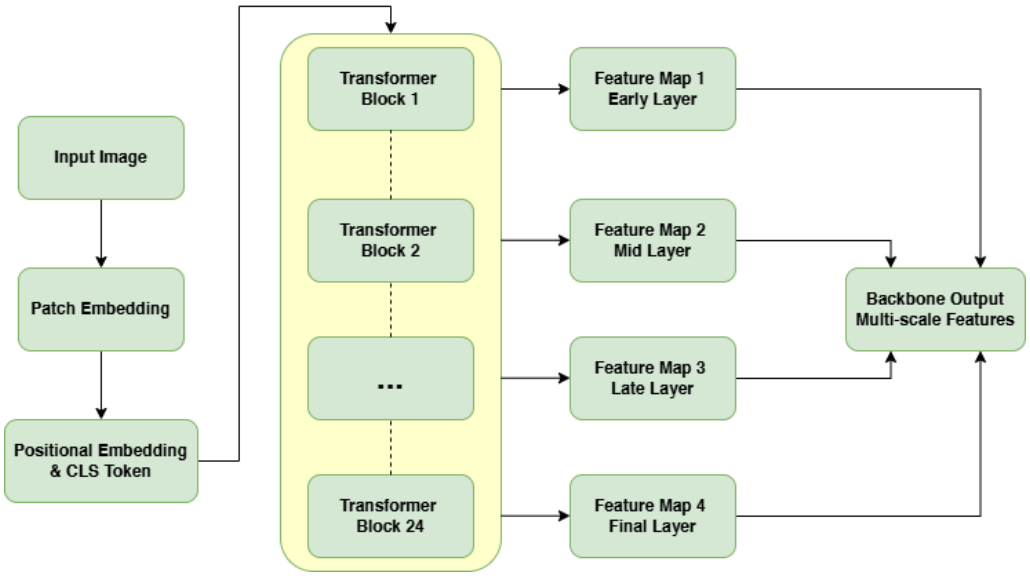

Figure 3.5: Backbone Architecture

After the attention layer, the output is passed through a two-layer MLP (feed-forward network) with dimensions $1024 \rightarrow 4096 \rightarrow 1024$. This further transforms the token representations through nonlinear operations. Residual connections are added after both the MHSA and MLP layers to stabilise training and retain gradient flow. Throughout the transformer stack, intermediate features are extracted at specific depths—layers 6, 12, 18, and 24—yielding tensors of shape $\mathbb{R}^{5330\times1024}$ at each selected stage. These extracted features provide multi-scale representations of the image, from fine-grained edges and textures to higher-level semantic content. All extracted features are LayerNorm-normalised before being saved to ensure consistent statistical behaviour across training.

In the final phase, **Reshaping Features**, the sequence output from each selected transformer layer is converted back to a spatial grid format for further use in decoders or downstream tasks. This is done by first removing the class token, resulting in a shape of $\mathbb{R}^{5329\times1024}$, and then reshaping the sequence back to image-like dimensions of $\mathbb{R}^{1024\times73\times73}$, corresponding to the original patch grid. This transformation restores the spatial topology required for

dense prediction tasks such as segmentation. As a result, we obtain a list of four tensors, each of shape $\mathbb{R}^{1024\times73\times73}$, containing multi-scale hierarchical features ready for decoding.

3. **Custom Decoder**: The decoding process begins with the **Input Projection** step, where each backbone feature map is transformed into a shared hidden dimension to facilitate multi-scale feature interaction. Specifically, each input feature map of shape $\mathbb{R}^{1024\times73\times73}$ undergoes a 1×1 convolution to reduce the channel dimension from 1024 to 256. This is followed by group normalisation with 32 groups across the 256 channels, which helps maintain feature consistency and stabilises training. The spatial dimensions are then flattened using flatten, producing tensors of shape $\mathbb{R}^{256\times5329}$. A final permutation reorders the tensor to $\mathbb{R}^{5329\times256}$. This procedure is applied to all four input feature maps, resulting in a list of four tensors, each of shape $\mathbb{R}^{5329\times256}$, ready for cross-scale interaction.

   Following this, the **Object Query Embedding** stage introduces a set of learnable object queries designed to represent candidate objects or semantic regions in the image. These queries are initialised as a tensor of shape $\mathbb{R}^{200\times256}$, where 200 is the number of object queries, and 256 is the hidden embedding dimension. These object queries serve as the input tokens to be refined throughout the subsequent transformer-based attention mechanism.

   The next component is the **Multi-Scale Cross-Attention Blocks**, where the object queries iteratively gather information from the projected multi-scale features. This module consists of 8 stacked transformer blocks. Within each block, the object queries attend to each of the four projected feature maps through multi-head attention mechanisms. For each scale, the queries of shape $\mathbb{R}^{200\times256}$ are used as the input to the multi-head attention layer, while the keys and values are taken from the corresponding projected feature tensor of shape $\mathbb{R}^{5329\times256}$. The attention output for each scale is added to the queries. After aggregating information from all four scales, a residual connection followed by LayerNorm, a feed-forward network (FFN) with the structure $256 \to 1024 \to 256$, and another LayerNorm completes the update process for each transformer block.

   Unlike the Mask2Former decoder [12], which relies on deformable attention for feature sampling and includes a heavier mask head with convolutional fusion layers, our decoder adopts a streamlined design. We replace deformable attention with a straightforward multi-scale cross-attention mechanism that is more transparent, lightweight, and easier to modify. Furthermore, we avoid architectural components like iterative refinement and deep supervision, which—while beneficial in some contexts—can introduce unnecessary complexity and hinder rapid experimentation or debugging during inference. All input features in our decoder are projected to a common dimension for consistency, whereas Mask2Former uses more flexible and adaptive projections. This deliberate simplification makes our model more interpretable and efficient, particularly during development and deployment in constrained environments.

   Once the attention process is complete, the updated queries proceed to the **Classification and Mask Heads**. Each query is passed through a linear

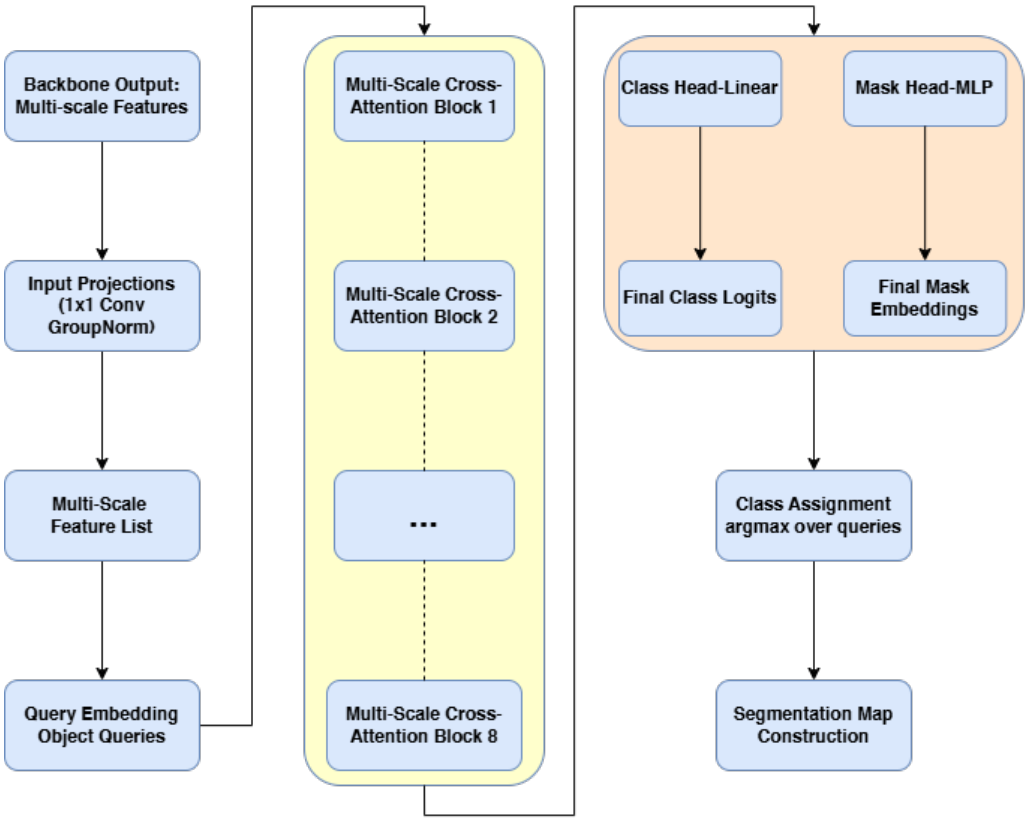

Figure 3.6: Decoder Architecture

classification layer that maps the hidden embedding of size 256 to class logits of size 12, resulting in an output tensor of shape $\mathbb{R}^{200\times12}$. In parallel, the same query embeddings are forwarded through a multi-layer perceptron (MLP) consisting of three hidden layers and ReLU activations, maintaining the hidden size of 256 throughout. This yields a mask embedding tensor of shape $\mathbb{R}^{200\times256}$, where each embedding is designed to be spatially interpreted in the subsequent stage.

In the **Mask Prediction** phase, the mask embeddings are combined with the highest-resolution feature map to produce spatial masks for each object query. The feature map used here has a shape of $\mathbb{R}^{256\times73\times73}$. The combination is carried out using an Einstein summation, where for each query q, and spatial location (h, w), the output is computed as:

$$\text{masks}_{q,h,w} = \sum_{c=1}^{256} \text{mask\_embed}_{q,c} \cdot \text{mask\_feat}_{c,h,w} \tag{3.2}$$

This operation results in a mask tensor of shape $\mathbb{R}^{200\times73\times73}$, representing the predicted spatial activation of each object query.

Finally, the **Segmentation Map Construction** stage aggregates the predictions into a semantic segmentation output. From the class logits of shape $\mathbb{R}^{200\times12}$ and the corresponding masks of shape $\mathbb{R}^{200\times73\times73}$, the model selects the query with the highest class confidence for each class. Formally, for each

class c, the best query q^* is selected using:

$$q^* = \arg\max_q \text{class\_logits}_{q,c} \tag{3.3}$$

Then, the final segmentation map is constructed by assigning the mask of the selected query to the corresponding class:

$$\text{seg}_{c,:,:} = \text{masks}_{q^*,:,:} \tag{3.4}$$

The resulting segmentation tensor has the shape $\mathbb{R}^{12\times73\times73}$, where each class is represented by a single, most confident mask.

4. **Output Postprocessing**: In this part, the segmentation map is resized back to the original input image size, ensuring the output aligns with the original image for visualisation or evaluation. The final output is $\mathbb{R}^{12\times1022\times1022}$.

For this architecture total parameter count is $319,709,280$, where $15,340,640$ are trainable, and the total multihead size is 20.30 GB. The forward and backwards pass size is 14136.13 MB, where the parameter size is 1239.33 MB and the estimated total size is 15387.99 MB.

Chapter 4

Implementations & Result Analysis

4.1 Implementations

4.1.1 Requirements

To support the training and evaluation of our proposed hierarchical transformer-based model, we used a high-performance local workstation running Ubuntu 24.10 (64-bit) powered by an Intel Core *i*7-14700 processor with 28 threads, 64 GB of RAM, and a Gigabyte Z690 AORUS ELITE motherboard. All experiments were conducted in a Python 3.12 environment with the latest versions of PyTorch 2.6.0, torchvision and xformers, ensuring full compatibility with GPU-accelerated libraries such as CUDA Toolkit 12.6 and cuDNN.
For GPU acceleration, which is crucial for handling large datasets and complex attention mechanisms, we deployed an NVIDIA RTX 4070 Ti SUPER graphics card with 16 GB of GDDR6 memory and driver version 570.133.20. During peak training periods, GPU utilisation reached 100%, allowing us to use full parallelism and minimise training time without compromising model convergence.

4.1.2 Model training

The dataset was split into training, validation, and test sets with a distribution of 67%, 22%, and 11% respectively. From the raw dataset, specific surgeries were selected to ensure adequate representation across anatomical structures. The **training set**, which includes at least 12 surgeries per organ, consists of Surgeries 1, 4, 5, 6, 8, 9, 10, 12, 15, 16, 17, 19, 22, 23, 24, 25, 27, 28, 29, 30, and 31. The **validation set**, ensuring 5 surgeries per structure, includes Surgeries 2, 7, 11, 13, 14, 18, 20, and 32. The **test set**, with 3 surgeries per structure, comprises Surgeries 3, 21, and 26. These train, test and validation splits are suggested by the authors of Dresden Surgical Anatomy Dataset[17].
The training process began with the data preparation using the dataloader that loads the image and mask file paths for both training and validation sets. Later on, it resizes the data according to the desired image size. For our case, we use 1022 as the image size for training because this one is the largest size that our current hardware could handle. The datasets are wrapped in PyTorch DataLoaders, which

handle batching and shuffling (for training), and the first validation sample is set aside for visualisation purposes after each epoch.
Once the data is perfectly loaded, it is sent to the model architecture. This model is configured with the number of output classes and other hyperparameters with xformers acceleration. The used loss function is a combination of both Cross Entropy and Dice Loss. Cross-Entropy Loss measures the pixel-wise classification error, while Dice Loss evaluates the overlap between the predicted and ground truth segmentation masks, which is especially important for handling class imbalance. The loss function is further customised with class weights to emphasise or de-emphasise certain classes, and a minimum background weight to reduce the influence of the background class. We have applied weighted training for better feature extraction. Initially, we applied two different weights to the training data: background weight = 0.05 and all other organs weight = 1.2. Later, we inspected that our model could segment the larger area occupied organs more precisely than the smaller area occupied organs. As a result, we introduced 3 categories from 11 organ classes where the weights were distributed in a way that: smallest organs (Intestinal veins, spleen, ureter) = 1.4, medium organs (Interior mesentric artery, pancreas, vesicular glands) = 1.3 and largest organs (Stomach, abdominal wall, colon, liver, small intestine) = 1.2.

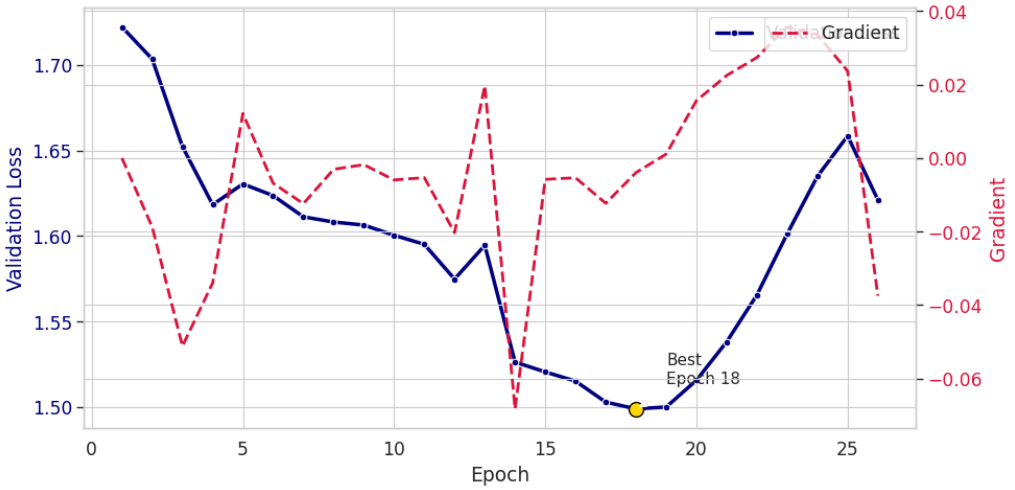

Figure 4.1: Gradient Convergence Curve with Validation Loss

During each training step, a batch of images and their corresponding ground truth masks are loaded from the DataLoader. The images are passed through the model, producing logits, which are the raw, unnormalized predictions for each class at each pixel. These logits are then used to compute the loss: Cross-Entropy Loss is calculated directly from the logits and ground truth, while Dice Loss requires converting the ground truth mask into a one-hot encoded format and the logits into probabilities using the softmax function. The total loss is the sum of these two components. The model's predictions are obtained by taking the class with the highest probability for each pixel, and metrics such as mean Intersection over Union (mIoU) are computed to monitor performance.
Backpropagation is the core of the learning process. After the loss is computed for the current batch, PyTorch's automatic differentiation engine calculates the gradients of the loss concerning each model parameter. These gradients indicate how

each parameter should be adjusted to reduce the loss. The optimiser, in this case AdamW, uses these gradients to update the model's weights. This process is repeated for every batch in the epoch, allowing the model to gradually learn better representations for segmenting the images.
At the end of each epoch, the average training loss and metrics are computed and logged. The model is then evaluated on the validation set, following a similar process: images are passed through the model, losses and metrics are computed, and predictions are generated. For qualitative assessment, the model's prediction on a sample validation image is masked and overlaid on the original image, then saved for visual inspection. The model's weights are also saved at the end of each epoch, and the best-performing model (based on validation loss) is exported for later use. This cycle of data loading, forward pass, loss computation, backpropagation, and evaluation repeats for the specified number of epochs or until the early stopping criteria are met, resulting in a trained segmentation model.

Figure 4.2: Model Learning Visualisation through epochs

4.1.3 Hyper-parameters

To achieve the highest efficiency of the proposed architecture, we have performed hyperparameter tuning that could enhance real-time organ segmentation.

- **Learning Rate**: We have explored the different learning rates and finally, we chose 0.00001 as the starting value of learning rate with a *Reduce LR On Plateau* scheduler, which reduces the learning rate by a factor of 0.1 if the validation loss plateaus for 3 epochs.
- **Batch Size**: We have tested various batch sizes (e.g., 16, 32, 64) to balance between faster convergence and maintaining data variability in each batch, and finally chose a batch size of 2, considering our hardware specifications.
- **Optimiser**: We have applied AdamW as an optimiser to ensure faster and more stable convergence. This process is repeated for each batch in the epoch.
- **Epochs**: We have set the number of required epochs as 100 with early stopping with patience value as 8. Hence, if the validation loss does not improve for 8 consecutive epochs, training would stop early, even if the maximum number of epochs has not been reached, and two checkpoints will be saved, one is the minimum loss and another one is the maximum mIOU.

- **Loss Function**: We have used a custom loss function, a combination of Dice Loss and Cross Entropy Loss.

$$\text{Cross Entropy} = -\sum_{c=1}^{C} y_c \log(p_c) \tag{4.1}$$

Here, C denotes the total number of classes, y_c is the ground truth label for class c, and p_c is the predicted probability for class c.

$$\text{Dice} = \frac{2 \times \text{Intersection}}{\text{Cardinality} + 1 \times 10^{-7}} \tag{4.2}$$

In this formula, Intersection refers to the element-wise multiplication between the predicted and ground truth masks, while Cardinality is the sum of the predicted and ground truth masks. A small constant 1×10^{-7} is added to avoid division by zero.

$$\text{Total Loss} = \text{Cross Entropy} + \text{Dice} \tag{4.3}$$

The Total Loss is computed as the summation of Cross Entropy Loss and Dice Loss, ensuring both pixel-wise accuracy and overlap consistency.

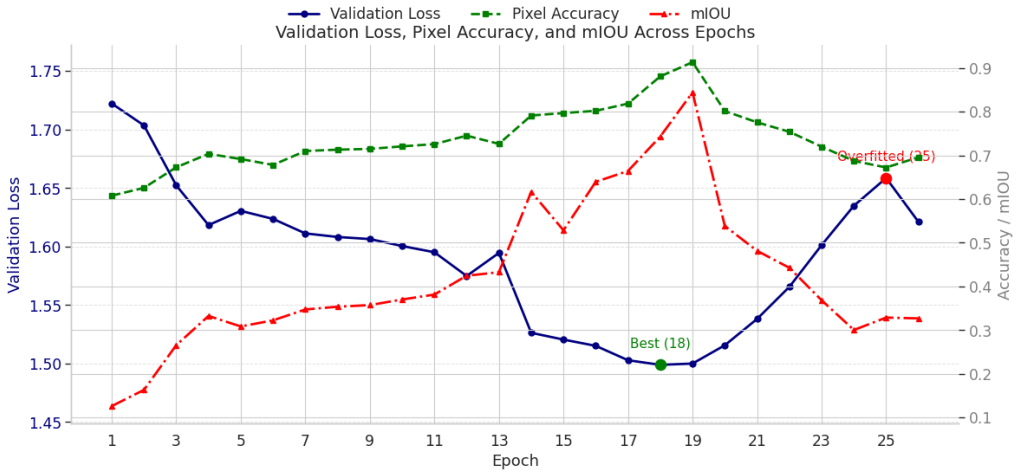

Figure 4.3: Validation Loss, Pixel Accuracy and mIOU VS Epoch

4.2 Performance Evaluation

4.2.1 Key Metric

The performance evaluation of the proposed surgical image processing pipeline is conducted using a comprehensive suite of metrics tailored to assess the accuracy, efficiency, and usability of each component. The evaluation focuses on two key aspects: segmentation performance and organ classification accuracy. These metrics ensure that the pipeline meets the rigorous demands of real-world surgical applications.

- **Precision**: Indicates how many of the predicted surgical regions are actually relevant, helping avoid false identification of non-organ tissues.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{4.4}$$

For class c:

$$\text{Precision}_c = \frac{\text{True Positive}_c}{\text{True Positive}_c + \text{False Positive}_c} \tag{4.5}$$

- **Recall**: Measures how well the model detects all relevant surgical regions, critical for not missing vital anatomical structures.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{4.6}$$

For class c:

$$\text{Recall}_c = \frac{\text{True Positive}_c}{\text{True Positive}_c + \text{False Negative}_c} \tag{4.7}$$

- **F1 Score**: Balances precision and recall to provide a single robust indicator of segmentation accuracy under surgical constraints.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.8}$$

For class c:

$$\text{F1}_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{4.9}$$

- **Confidence Score**: Represents the model's estimated probability for assigning each pixel to a specific class, capturing uncertainty in pixel-wise classification.

$$\text{Confidence Score}(x_i) = \max_{c \in C} P(y_i = c \mid x_i) \tag{4.10}$$

Where x_i is the feature vector for pixel i, y_i is the predicted label for pixel i, C is the set of all semantic classes, and $P(y_i = c \mid x_i)$ is the predicted probability for class c at pixel i.

- **IoU (Intersection over Union)**: Quantifies the overlap between predicted and actual surgical areas, crucial for evaluating spatial accuracy.

$$\text{IoU} = \frac{|\text{Prediction} \cap \text{Ground Truth}|}{|\text{Prediction} \cup \text{Ground Truth}|} \tag{4.11}$$

For class c:

$$\text{IoU}_c = \frac{\text{True Positive}_c}{\text{True Positive}_c + \text{False Positive}_c + \text{False Negative}_c} \tag{4.12}$$

Hence,

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \frac{\text{True Positive}_c}{\text{True Positive}_c + \text{False Positive}_c + \text{False Negative}_c} \tag{4.13}$$

- **Dice Coefficient**: Highlights the spatial agreement of predicted organ boundaries with ground truth, important for surgical precision.

$$\text{Dice} = \frac{2 \cdot |\text{Prediction} \cap \text{Ground Truth}|}{|\text{Prediction}| + |\text{Ground Truth}|} \tag{4.14}$$

For class c:

$$\text{Dice}_c = \frac{2 \cdot \text{True Positive}_c}{2 \cdot \text{True Positive}_c + \text{False Positive}_c + \text{False Negative}_c} \tag{4.15}$$

- **Pixel Accuracy**: Measures the proportion of correctly classified pixels over the total number of pixels.

$$\text{PixelAcc} = \frac{\sum_{i=1}^{C} n_{ii}}{\sum_{i=1}^{C} \sum_{j=1}^{C} n_{ij}} \tag{4.16}$$

Here, C denotes the total number of semantic classes. The term n_{ij} represents the number of pixels whose ground truth label is class i but are predicted as class j. The diagonal elements n_{ii} indicate the number of correctly classified pixels for class i. The numerator $\sum_{i=1}^{C} n_{ii}$ thus gives the total count of correctly predicted pixels, while the denominator $\sum_{i=1}^{C} \sum_{j=1}^{C} n_{ij}$ represents the total number of labeled pixels across all classes.

- **SSIM (Structural Similarity Index)**: Evaluates structural similarity between predicted and ground truth masks, vital for preserving anatomical consistency in medical imaging.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4.17}$$

where:

  - μ_x and μ_y are the mean pixel values of images x and y
  - σ_x^2 and σ_y^2 are the variances of x and y
  - σ_{xy} is the covariance between x and y
  - C_1 and C_2 are constants to stabilise the division (based on the data range)

This equation is applied to the grayscale versions of the predicted and ground truth segmentation masks.

In the context of semantic segmentation for surgical environments, traditional metrics like accuracy (except pixel-wise) are not ideal because they can be misleading in the presence of class imbalance, a common scenario where background pixels vastly outnumber critical organ pixels. A high accuracy might simply reflect the model's ability to predict the dominant background class correctly, while failing to capture small but clinically important structures. Similarly, metrics like specificity or overall pixel accuracy do not adequately penalise false negatives, which are particularly critical in medical applications where missing an organ boundary can lead to severe consequences. Therefore, we rely on more robust and class-sensitive metrics like Precision, Recall, F1 Score, IoU, Dice Coefficient, and SSIM, which provide a more balanced and structure-aware evaluation of segmentation performance, ensuring both localisation accuracy and overlap fidelity in delicate surgical contexts.

4.2.2 Statistical Analysis

To rigorously assess the performance and reliability of our proposed hierarchical Transformer-based semantic segmentation model, we conducted six independent training runs, each initialised with random seeds and shuffled datasets. This enabled us to capture variability introduced by stochastic optimisation, data sampling, and initialisation, thereby offering a robust statistical evaluation of model consistency. Table 4.1 summarises the aggregate performance across global metrics including Intersection over Union (IoU), Dice coefficient, Precision, Recall, F1-score, Structural Similarity Index Measure (SSIM), and Pixel Accuracy. The model attained a mean Dice score of 0.9098 and a mean IoU of 0.8654, with tight 95% confidence intervals of ±0.0749 and ±0.0996, respectively. All performance metrics demonstrated statistical significance against a conservative baseline threshold (Dice = 0.75), with p-values ranging from 0.0009 to 0.0021. These results strongly indicate that the proposed model's improvements are not attributable to random chance but reflect genuine performance gains achieved through architectural innovation.

Metric	IoU	Dice or F1	Precision	Recall	SSIM	Pixel Acc	Confidence
Mean	0.8654	0.9098	0.9426	0.9004	0.8887	0.9004	0.9132
Median	0.9068	0.9421	0.9700	0.9327	0.9169	0.9327	0.9174
Std Dev	0.1229	0.0925	0.0728	0.0939	0.0931	0.0939	0.0614
Variance	0.0151	0.0086	0.0053	0.0088	0.0087	0.0088	0.0038
95% CI	±0.0996	±0.0749	±0.0589	±0.0760	±0.0755	±0.0760	±0.0489
p-value	0.0021	0.0013	0.0009	0.0018	0.0016	0.0017	0.0011

Table 4.1: Performance metrics for semantic segmentation (mean over 6 runs), including 95% confidence intervals (CI) and p-values based on a one-sample t-test.

We have also evaluated the model's class-wise performance.

Class	IoU	Dice or F1	Recall	SSIM	Pixel Acc.	Precision	Confidence	95% CI	p-value
Abdominal Wall	0.8427	0.9145	0.8900	0.8912	0.8530	0.9415	0.8790	±0.0480	0.0017
Colon	0.8271	0.8809	0.8715	0.8657	0.8715	0.9270	0.8704	±0.0451	0.0026
Inferior Mesenteric Artery	0.9351	0.9590	0.9497	0.9396	0.9396	0.9771	0.9348	±0.0380	0.0005
Intestinal Veins	0.8794	0.9312	0.8831	0.8681	0.8681	0.9915	0.9070	±0.0425	0.0010
Liver	0.8421	0.9147	0.9100	0.9023	0.8425	0.9197	0.8498	±0.0462	0.0018
Pancreas	0.9383	0.9623	0.9537	0.9396	0.9396	0.9765	0.9479	±0.0357	0.0004
Small Intestine	0.8641	0.9136	0.9082	0.8898	0.8898	0.9399	0.8892	±0.0433	0.0019
Spleen	0.9431	0.9623	0.9581	0.9504	0.9504	0.9696	0.9200	±0.0396	0.0004
Stomach	0.8434	0.8996	0.8642	0.8483	0.8683	0.9512	0.8840	±0.0470	0.0021
Ureter	0.9634	0.9780	0.9676	0.9579	0.9579	0.9897	0.9376	±0.0335	0.0002
Vesicular Glands	0.9278	0.9524	0.9508	0.9364	0.9364	0.9641	0.9407	±0.0374	0.0006
Multilabel	0.7347	0.8033	0.8335	0.8343	0.8343	0.8276	0.8840	±0.0535	0.0042

Table 4.2: Class-wise semantic segmentation performance

To further elucidate the segmentation robustness across anatomical structures, we conducted class-wise analysis as shown in Table 4.2. Notably, critical vascular and glandular structures such as the *Ureter*, *Pancreas*, *Spleen*, and *Inferior Mesenteric Artery* achieved Dice scores above 0.95, with exceptionally low confidence intervals (±0.0335 to ±0.0396) and p-values below 0.001. This confirms not only the high accuracy of the model but also its stability in segmenting small, low-pixel, and clinically sensitive regions—an enduring challenge in surgical image analysis.

Even under the complex scenario of multilabel segmentation, where multiple overlapping organs are simultaneously present, the model maintained competitive performance (Dice = 0.8033, CI = $\pm$0.0535, p = 0.0042). This highlights the model's robustness in real-world surgical scenes, which are inherently occluded and spatially congested.
Overall, the consistently low standard deviations and narrow confidence intervals across both global and class-specific metrics demonstrate the stability of the proposed architecture. Coupled with uniformly low p-values across all evaluated categories, these results provide compelling evidence for the model's reliability and statistical robustness. Such consistency is critical for deployment in high-stakes applications like intraoperative navigation and robotic-assisted surgery, where precision and trustworthiness are paramount.

4.2.3 Comparisons and Relationships

We found only one study that performs semantic segmentation of anatomical structures in a surgical setting. In "One Model to Use Them All: Training a Segmentation Model with Complementary Datasets"[21], the authors segment six organs: the abdominal wall, colon, liver, pancreas, small intestine, and stomach. We compare the dice score of their fully supervised multiclass segmentation models with our approach. Despite the scarcity of prior research explicitly focused on intraoperative organ segmentation, we identified one relevant benchmark study: *"One Model to Use Them All: Training a Segmentation Model with Complementary Datasets"* [21]. In this work, the authors proposed a fully supervised multiclass segmentation model trained on complementary surgical datasets and evaluated on six abdominal organs: abdominal wall, colon, liver, pancreas, small intestine, and stomach using DeepLab[5] like architecture.
We compare the **dice score** of their fully supervised multiclass segmentation with our approach.

Organs	**Theirs (Dice)**	**Ours (Dice)**	**Absolute Δ**	**Relative % Gain**
Abdominal Wall	0.800	0.915	+0.115	14.4%
Colon	**0.450**	**0.881**	**+0.431**	**95.8%**
Liver	0.862	0.915	+0.053	6.1%
Pancreas	0.679	0.962	+0.283	41.7%
Small Intestine	0.913	0.914	+0.001	0.1%
Stomach	0.719	0.900	+0.181	25.2%

Table 4.3: Comparison of Dice scores with existing study [21]. Our model shows consistent improvements across all organs, including up to 95.8% relative gain.

Table 4.3 presents a comparative analysis of Dice scores between their approach and our proposed hierarchical Transformer-based model. Our method consistently outperforms theirs across all organ classes, with absolute Dice improvements ranging from +0.001 to +0.431. The most significant gains are observed in the segmentation of the colon and pancreas. For the colon, our model improved the Dice score from 0.450 to 0.881, representing an absolute gain of +0.431 and a relative improvement of approximately 95.8%. In the case of the pancreas, we achieved an increase from 0.679 to 0.962, equivalent to a 41.7% relative improvement.

For the stomach, the Dice score improved from 0.719 to 0.900, a relative increase of 25.2%. The abdominal wall improved from 0.800 to 0.915, indicating a 14.4% gain, while the liver advanced from 0.862 to 0.915, marking a 6.1% improvement. The small intestine showed a marginal increase from 0.913 to 0.914, which, though minor in absolute terms (+0.001), demonstrates the stability of our model even in already high-performing cases. On average, our model achieved an absolute Dice improvement of 13.0 percentage points across the six overlapping classes, corresponding to a relative increase of approximately 17.2% over the existing benchmark.
In addition to superior performance, our model offers broader segmentation capabilities. While the previous study[21] was limited to six organ classes, our approach successfully segments 11 anatomically distinct structures, including smaller and more complex organs such as the inferior mesenteric artery, intestinal veins, ureter, and vesicular glands. This expanded coverage underscores the strong generalisation ability of our model and its potential utility in comprehensive surgical scene understanding.
We attribute these improvements to the design of our hierarchical Transformer architecture, which integrates a DINOv2[23] backbone with a multi-scale cross-attention decoder. This configuration enables the model to preserve both fine boundary details and holistic spatial context—capabilities that are critical for robust segmentation under challenging intraoperative conditions characterised by occlusion, variability in lighting, and anatomical overlap.

4.2.4 Discussion

Our proposed hierarchical Transformer-based model exhibits strong overall performance in the semantic segmentation of intraoperative anatomical structures. High average scores across global evaluation metrics—including a mean IoU of 0.8654 and a Dice coefficient of 0.9098—indicate precise anatomical delineation and spatial consistency. The model also achieved a high mean precision (0.9426) and recall (0.9004), suggesting its capability to accurately identify organ boundaries while maintaining low rates of false positives and false negatives.
Despite these promising aggregate results, the metric distributions reveal non-trivial variability. Median values exceed corresponding means for all metrics, and standard deviations such as 0.1229 for IoU indicate the presence of more challenging cases in the test set. This suggests that while many segmentations are near-perfect, performance can degrade in a subset of difficult scenarios, potentially due to visual occlusion, low contrast, complex lighting, or anatomical anomalies.
Class-wise segmentation results provide further insight. Structures such as the *Inferior Mesenteric Artery* (Dice: 0.9590) and *Ureter* (Dice: 0.9780) achieve near-ceiling performance, likely owing to their well-defined visual and spatial characteristics. In contrast, the *Multilabel* category underperforms relative to others, with a Dice score of 0.8033. This class presents increased complexity due to the simultaneous presence of multiple organs, which introduces spatial ambiguity, inter-organ occlusion, and similar visual textures, making it inherently harder to disambiguate individual structures. These challenges can hinder the pixel-wise classification performance of segmentation models, especially in densely packed anatomical regions.
A direct comparison with Jenke et al. [21] further illustrates the strengths of our approach. Our model outperforms theirs across all six shared organ categories, with

particularly notable improvements for the *colon* (Dice: 0.881 vs. 0.450; +95.8%) and *pancreas* (0.962 vs. 0.679; +41.7%). Such improvements are clinically significant, given the surgical importance and segmentation difficulty of these organs. High segmentation accuracy in these cases may support more precise image-guided interventions and enhance intraoperative decision-making.

Moreover, our model demonstrates broader anatomical coverage than previous work that segments 11 structures compared to 6, extending to vascular and glandular regions such as the *intestinal veins*, *vesicular glands*, and *inferior mesenteric artery*. This expanded anatomical scope highlights the generalisation ability and robustness of the proposed architecture, which is essential for deployment in the diverse and dynamic settings typical of minimally invasive surgery.

In summary, the statistical consistency across most structures, substantial performance gains over previous models, and wider organ coverage underscore the effectiveness of our approach in intraoperative semantic segmentation. Nonetheless, the variability in difficult cases and the reduced accuracy in complex multilabel contexts indicate areas where segmentation remains challenging and must be interpreted with appropriate clinical caution.

Chapter 5

Conclusion

5.1 Conclusion

This paper presents a comprehensive review of the implementation of hierarchical transformers for the semantic segmentation of intraoperative anatomical structures—an emerging focus in medical image analysis. Our study demonstrates that the proposed architecture maintains both efficiency and accuracy in real-time applications, significantly improving the precision of segmentation during surgical procedures. The hierarchical structure of our model enables more refined and efficient image segmentation, positioning it as an invaluable tool for medical practitioners for telesurgery. Furthermore, our ongoing research aims to integrate AI-driven enhancements into the real-world system, offering more accurate and context-specific outputs that can assist in performing robotic surgeries with greater precision. This advancement holds the potential to be a transformative development in semi-automated surgical procedures, with the capability to substantially improve clinical outcomes. As research in hierarchical transformers for real-time medical image analysis progresses, we believe it will usher in a new era of innovation, moving us closer to the realisation of fully autonomous robotic surgery.

5.2 Summary of Findings

Based on the evaluation of a semantic segmentation model on a surgical dataset, key findings are: Overall performance metrics were strong, with a mean Dice of 0.9098 and IoU of 0.8654. Median scores were higher (Dice 0.9421, IoU 0.9068), indicating occasional outliers. Standard deviations were moderate (e.g., 0.0925 for Dice). Class-wise analysis showed excellent results for Ureter (Dice 0.9780), Spleen (0.9623), and Pancreas (0.9623), while Multilabel performed weakest (Dice 0.8033). Compared to existing research, the model substantially outperformed in segmenting abdominal organs, notably improving Dice for Colon (0.881 vs. 0.450), Stomach (0.900 vs. 0.719), and Pancreas (0.962 vs. 0.679). Small intestine segmentation was comparable (0.914 vs. 0.913). The model demonstrates state-of-the-art anatomical segmentation in surgical contexts.

5.3 Limitations

Our proposed model gave promising results, but there are several limitations that remain open for future improvements. In real surgical environments, the effectiveness of handling depth, motion, or temporal continuity is restricted because the architecture is currently limited to processing 2D surgical images. Moreover, although it is designed for real-time use, the model has not been tested in a live surgical system; therefore, the inference speed and responsiveness under clinical conditions remain unverified. The model also requires a high-end GPU for training and deployment, which may be difficult for many hospitals due to the unavailability of such systems. Additionally, the dataset is limited to a specific type of surgery, so the model's generalizability across different procedures or anatomical variations remains uncertain. We also observed occasional false positives in low-contrast or overlapping regions, indicating that further refinement in segmentation accuracy is needed. Lastly, the model is currently limited to detecting only 11 predefined abdominal organs and cannot identify additional anatomical structures outside the dataset's scope, which requires a larger dataset to train our model to make it more generalised. Therefore, overcoming these limitations can lead to practical clinical adoption and broader applicability.

5.4 Recommendations for Future Work

To overcome the current limitations, we plan to incorporate 3D or video-based segmentation, which will enhance robustness in dynamic surgical scenes by leveraging temporal and spatial continuity. Moreover, we aim to apply model distillation techniques to improve and reduce deployment costs. We also plan to train our model on a more diverse dataset of surgeries to improve its generalizability by employing domain adaptation techniques. Lastly, expanding the model to support a wider range of organ classes can enhance its effectiveness in complex surgical scenarios.

Bibliography

[1] M. Martinez-Escobar, J. Leng Foo, and E. Winer, "Colorization of ct images to improve tissue contrast for tumor segmentation," *Computers in Biology and Medicine*, vol. 42, no. 12, pp. 1170–1178, 2012, ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2012.09.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482512001540.

[2] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015. arXiv: 1505.04597 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1505.04597.

[3] J. D. Bohnen, M. N. Mavros, E. P. Ramly, *et al.*, "Intraoperative adverse events in abdominal surgery," *Annals of Surgery*, vol. 265, no. 6, pp. 1119–1125, 2016. DOI: 10.1097/sla.0000000000001906.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018. DOI: 10.1109/TPAMI.2017.2699184.

[5] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep. 2018.

[6] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, 2020. arXiv: 1905.11946 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1905.11946.

[7] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020. DOI: 10.1109/TMI.2019.2959609.

[8] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21, Red Hook, NY, USA: Curran Associates Inc., 2021, ISBN: 9781713845393.

[9] W. Wang, E. Xie, X. Li, *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 548–558. DOI: 10.1109/ICCV48922.2021.00061.

[10] R. Azad, M. T. AL-Antary, M. Heidari, and D. Merhof, *Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model*, 2022. arXiv: 2207.13415 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2207.13415.

[11] R. B. D. Boer, C. De Jongh, W. T. E. Huijbers, *et al.*, "Computer-aided anatomy recognition in intrathoracic and -abdominal surgery: A systematic review," *Surgical Endoscopy*, vol. 36, no. 12, pp. 8737–8752, 2022. DOI: 10.1007/s00464-022-09421-5.

[12] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," 2022.

[13] K. Han, J. Guo, Y. Tang, and Y. Wang, *Pyramidtnt: Improved transformer-in-transformer baselines with pyramid architecture*, 2022. arXiv: 2201.00978 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2201.00978.

[14] S. M. Hussain, A. Brunetti, G. Lucarelli, R. Memeo, V. Bevilacqua, and D. Buongiorno, "Deep learning based image processing for robot assisted surgery: A systematic literature survey," *IEEE Access*, vol. 10, pp. 122 627–122 657, 2022. DOI: 10.1109/ACCESS.2022.3223704.

[15] X. Ke, X. Zhang, T. Zhang, J. Shi, and S. Wei, "Sar ship detection based on swin transformer and feature enhancement feature pyramid network," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 2163–2166. DOI: 10.1109/IGARSS46834.2022.9883800.

[16] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 12 504–12 513.

[17] M. Carstens, F. M. Rinner, S. Bodenstedt, *et al.*, "The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science," *Scientific Data*, vol. 10, no. 1, p. 3, 2023, ISSN: 2052-4463. DOI: 10.1038/s41597-022-01719-2. [Online]. Available: https://doi.org/10.1038/s41597-022-01719-2.

[18] A. Kirillov, E. Mintun, N. Ravi, *et al.*, "Segment anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003. DOI: 10.1109/ICCV51070.2023.00371.

[19] R. Liu, S. Duan, L. Xu, L. Liu, J. Li, and Y. Zou, "A fuzzy transformer fusion network (fuzzytransnet) for medical image segmentation: The case of rectal polyps and skin lesions," *Applied Sciences*, vol. 13, no. 16, 2023, ISSN: 2076-3417. DOI: 10.3390/app13169121. [Online]. Available: https://www.mdpi.com/2076-3417/13/16/9121.

[20] W. Wang, J. Dai, Z. Chen, *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 14 408–14 419.

[21] A. C. Jenke, S. Bodenstedt, F. R. Kolbinger, M. Distler, J. Weitz, and S. Speidel, "One model to use them all: Training a segmentation model with complementary datasets," *International Journal of Computer Assisted Radiology and Surgery*, vol. 19, pp. 1233–1241, 2024. DOI: 10.1007/s11548-024-03145-8. [Online]. Available: https://doi.org/10.1007/s11548-024-03145-8.

[22] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024. DOI: 10.1038/s41467-024-44824-z. [Online]. Available: https://doi.org/10.1038/s41467-024-44824-z.

[23] M. Oquab, T. Darcet, T. Moutakanni, *et al.*, *Dinov2: Learning robust visual features without supervision*, OpenReview preprint, 2024. [Online]. Available: https://openreview.net/forum?id=a68SUt6zFt.

[24] Z. Qin, J. Liu, X. Zhang, *et al.*, "Pyramid fusion transformer for semantic segmentation," *IEEE Transactions on Multimedia*, vol. 26, pp. 9630–9643, 2024. DOI: 10.1109/TMM.2024.3396281.

[25] T. Zhao, G. Li, and S. Zhao, "End-to-end image colorization with multiscale pyramid transformer," *IEEE Transactions on Multimedia*, vol. 26, pp. 11 332–11 344, 2024. DOI: 10.1109/TMM.2024.3453035.