International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 8, 2025
https://theaspd.com/index.php

Explainable Semantic Segmentation Of Organs In Laparoscopic Hysterectomy Using Transfer Learning, Ensembles, And Vision Transformers

Mamdouh Gomaa[1 2], Ahmed Rabie[1], Al Hussien Seddik Saad[1], Osman Ali Sadek[1], Moheb Ramzy Girgis[1]
[1]Department of Computer Science, Faculty of Science, Minia University, 61511, Minia, Egypt.
[2]Department of Computer Science, Faculty of Information Technology, Amman Arab University, 11953, Amman, Jordan.

Abstract
Purpose: To develop and evaluate a clinically oriented semantic-segmentation framework that delineates the ureter, uterine artery, and pelvic nerves in laparoscopic hysterectomy. Methods: We analyzed the publicly available, de-identified UD Ureter–Uterine Artery–Nerve dataset (586 RGB images) with expert multiclass masks. Images were resized to 128×128 and intensities normalized. We trained six transfer-learning U-Net backbones (VGG16, ResNet50, MobileNetV2, EfficientNet-B0, Inception-ResNetV2, MultiResUNet), a Vision Transformer (ViT), and a weighted soft-probability ensemble of OrganFocus U-Net, a baseline U-Net, and the ViT. Augmentation (flips, small rotations, mild intensity changes, elastic deformations) was applied to the training split only; all augmented variants of an image remained within the same split. An 80/10/10 image-level split (train/validation/test) with class balance was used. The primary metric was mean Intersection-over-Union (mIoU). Ninety-five percent confidence intervals (95% CI) were computed by non-parametric bootstrap over test images (≥5,000 resamples; percentile method). Results: The best transfer-learning backbone (VGG16) achieved mIoU 76.85% (95% CI: 73.71–79.76%). The ViT achieved 84.00% (95% CI: 81.52–85.35%). The weight-optimized ensemble reached 86.57% (95% CI: 84.49–87.99%). Grad-CAM heatmaps showed anatomically coherent focus across models. Conclusions: Combining complementary inductive biases from convolutional encoders and Transformers via a weighted ensemble yields high-accuracy segmentation on clinically acquired laparoscopic images, while Grad-CAM supports case-level interpretability. Future work will profile compute/throughput and validate on multi-institutional data and surgical videos.
Keywords: *Laparoscopic Hysterectomy; Deep Learning; Semantic Segmentation; Transfer Learning; Ensemble Learning; Vision Transformer; Explainable Artificial Intelligence.*

INTRODUCTION

Laparoscopic hysterectomy requires reliable identification of critical pelvic structures to avoid iatrogenic injury. Accurate, real-time delineation of the uterine artery, pelvic nerves, and especially the ureter is therefore clinically important.

The Role of Transfer Learning in Enhancing Medical Image Segmentation

Transfer learning (TL) is a powerful machine learning paradigm that enables knowledge gained from solving one problem to be applied to a different, but related, problem [1,2]. In computer vision, this typically involves pre-training a deep neural network on a large, general dataset and then fine-tuning it for a specific medical imaging task. This approach has proven particularly effective in medical image analysis, where annotated data can be scarce and costly to obtain [3,4]. Comprehensive reviews have highlighted how TL facilitates the adaptation of deep models for segmentation and other clinical applications, thereby improving performance and generalization in medical image tasks [5].

Various formulations and TL approaches have been explored, such as a study on carotid intima-media boundary segmentation that found the usefulness of TL with a model pre-trained on natural images [6]. However, these studies apply TL to non-Fully Convolution Network (FCN) classification architectures that include fully connected layers, whereas our approach integrates TL into the encoder (downsampling path) of an FCN-based segmentation model, excluding any fully connected layers.

A study by Valindria et al. [7] showed that a model trained for liver and kidney segmentation on a dataset of 35 MRI images, from the MALIBO study, performed poorly on a target dataset of 45 MRI images, from the UK Biobank, due to the differences being image size, resolution, and acquisition center. Fine-tuning the source-trained model on the target domain achieved performance comparable to training from scratch. To address this domain shift, the authors introduced Reverse Classification Accuracy (RCA) [8] to identify the most informative images for annotation in the target domain, demonstrating that annotating as few as five images could achieve accuracy comparable to using all 45 images, in both fine-tuning and training-from-scratch scenarios.

Vision Transformers in Medical Image Analysis

Vision Transformers (ViTs) have shown proficiency in learning long-range dependencies and have recently demonstrated remarkable representational learning capabilities in computer vision and medical image applications [9-11]. Unlike CNNs, ViTs achieve improved long-range information by tokenizing images into 1D sequences and using self-attention blocks for global communication [10]. However, ViTs may be less adept at capturing local positional information compared to CNNs, lacking the locality inductive bias inherent to CNNs [9,12]. Despite their capabilities, ViTs often require substantial amounts of training data, which can be costly to obtain [10,13].

Recent advances have highlighted the effectiveness of ViTs in medical image segmentation tasks, owing to their ability to capture global contextual information for high-resolution images [9]. Unlike CNNs, which primarily focus on local features, Transformers model broader spatial dependencies, enhancing segmentation performance.

Models such as U-NETR, VT-U-Net, and SwinU-NETR use transformer blocks as the main encoder and CNNs as decoders, forming U-shaped architectures that benefit from both global and local feature modeling [10, 13, 14]. In this setup, a CNN encoder is followed by transformer blocks to capture broader dependencies while maintaining computational efficiency. Examples include TransU-Net and its extensions, which have demonstrated strong results in CT and 3D medical image segmentation [15, 16].

Models, such as TransFuse and FusionNet, utilize both transformer and CNN encoders in parallel, merging global and local features for enhanced segmentation accuracy [17, 18].

Yu et al. [19] proposed UNesT, a hierarchical Transformer architecture designed for efficient volumetric medical image segmentation. The model employs a nested tokenization strategy that captures both local spatial details and long-range global dependencies, enabling scalability to high-resolution volumetric data. Their approach achieved state-of-the-art results on whole-brain segmentation across 133 tissue classes and on fine-grained renal substructure tasks. By releasing both code and pretrained models, UNesT provides a practical and reproducible benchmark, illustrating the growing impact of Transformer-based designs in advancing beyond conventional CNN and U-Net architectures.

Peng et al. [20] introduced a U-attention nested U-Transformer for multi-organ segmentation. A U-shaped encoder–decoder with channel–spatial attention is paired with a convolutional Transformer module that uses 1D convolution for feature fusion instead of multi-head self-attention. On two public clinical datasets, DSC increased from 91.36% to 95.11%, while average symmetric surface distance dropped from 1.68 to 0.40.

Pak et al. [21] compared three CNN architectures DeepLabV3, MANet, and U-Net++ against three vision transformers SegFormer, BEiT, and DPT for semantic segmentation during robot-assisted radical prostatectomy. Using 3,000 intraoperative images, CNNs generally outperformed transformers on organ masks: MANet achieved the highest mean DSC of 94.40% (mean IoU 87.60%), followed by DeepLabV3 (DSC 93.80%, IoU 85.60%) and U-Net++ (DSC 93.00%, IoU 83.50%). Among transformers, DPT reached DSC 94.00% (IoU 75.00%), SegFormer 91.90% (IoU 76.10%), and BEiT 91.60% (IoU 69.00%).

Wang et al. [22] addressed long-range dependency and detail preservation by unifying CNN and Transformer streams in CTUnet. A CTBlock fuses CNN and Transformer features, and a CTFusion module integrates local and global information while replacing standard U-Net skips. On BUSI and DDTI, CTUnet improved segmentation while better retaining semantics and spatial detail.

Sun et al. [23] proposed DA-TransUNet, which embeds dual attention blocks and Transformer modules into a U-shaped architecture to optimize positional and channel features. Dual attention in each skip connection helps filter irrelevant information. Validated on five datasets, DA-TransUNet improved segmentation accuracy, DSC, and Hausdorff distance over strong baselines.

The integration of transformer-based architectures and CNNs continues to advance the state-of-the-art in medical image segmentation, improving accuracy and reliability in various clinical contexts.

Principles and Architectures of Ensemble Techniques

An ensemble method is understood as a procedure that combines a set of individual learners through a specified aggregation rule to tackle a particular problem [24-26]. The foundations of ensemble learning can be traced to the early 1990s, with landmark studies by Hansen and Salamon [27] and by Schapire [28]. In essence, the base learners within an ensemble are trained independently on the same task, and a final decision is obtained by aggregating their diverse predictions to capitalize on complementarity and reduce variance, as illustrated in Figure 1.

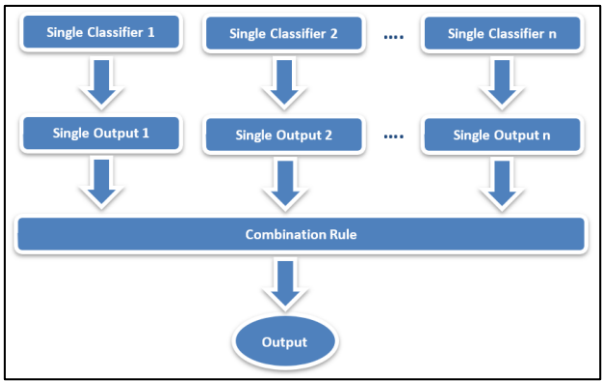


Figure 1. The common architecture of the ensemble system. The single techniques can be either classification or regression techniques.

Early formulations showed that averaging diverse neural networks can improve generalization, laying a foundation for modern ensemble methods [27]. The design of an ensemble system typically proceeds in two phases: generation and integration. In the generation phase, one may construct either a homogeneous collection multiple instances of the same learning algorithm trained under distinct configurations or data resamplings or coupled with meta-schemes such as Bagging [29], Boosting [28], or the Random Subspace method [30] or a heterogeneous collection different algorithms trained on the same data to encourage complementary inductive biases [24,31]. The integration phase then applies an explicit combining rule (e.g., majority voting, (weighted) averaging, or stacking) to fuse member outputs into a single prediction that is typically more reliable than any individual constituent [25,26].

Ensemble learning has also seen renewed momentum with deep, layer-wise constructions. gcForest [32] arranges a cascade whose layers include Completely Random Tree Forests and Random Forests, concatenating class-distribution vectors with original features to feed the next layer. Utkin et al. [33] extend this framework via weighted averaging over class-probability subsets. Nguyen et al. [34] propose MULES, a deep ensemble that performs classifier and feature selection at each layer under a bi-objective optimization. Qi et al. [35] describe a deep ensemble whose layers comprise ensembles of Support Vector Machines, with parameters determined via AdaBoost.

Applying Gradient-weighted Class Activation Mapping for Interpretable Laparoscopic Organ Segmentation

Artificial intelligence (AI), particularly its deep learning branch, has enabled remarkable progress in complex tasks such as image segmentation. However, these models often function as "black boxes," making it difficult for humans to interpret how predictions are made [36], This lack of transparency can undermine trust and hinder the adoption of AI in sensitive domains like medicine, where understanding the reasoning behind predictions is as important as the results themselves. Although recent developments in explainable AI (XAI) have aimed to address this issue, most methods still struggle with the unique demands of segmentation models, especially in medical imaging where explanations must be interpretable at both pixel and region levels [37].

One notable XAI technique is Gradient-weighted Class Activation Mapping (Grad-CAM) [38], which produces visual heatmaps to highlight the regions in an input image that most influence a model's prediction. In this research, Grad-CAM was applied to the outputs of the utilized segmentation models, providing intuitive visual explanations that allow researchers and clinicians to verify whether the model's focus aligns with anatomically relevant regions. This approach enhances the transparency and trustworthiness of the model's predictions for laparoscopic organ segmentation, supporting clinical validation and interpretability, and is especially valuable in high-stakes medical contexts where understanding model reasoning is critical.

This study aimed to develop and evaluate pixel-wise segmentation of the ureter, uterine artery, and pelvic nerves during laparoscopic hysterectomy using transfer-learning backbones, a Vision Transformer (ViT), and a weighted-average ensemble. This work extends our earlier study [39] by adding a weight-optimized heterogeneous ensemble, reporting explicit image-level train/validation/test splits, and providing 95% bootstrap confidence intervals (95% CI). We also enforced stricter experimental control and systematic explainability via Gradient-weighted Class Activation Mapping (Grad-CAM) to enhance clinical interpretability and reproducibility. The ureter is particularly challenging due to its thin, tortuous course, frequent occlusions by instruments/tissue, and specular highlights.

MATERIALS AND METHODS

Selection and Description of Participants.

This study uses the UD Ureter-Uterine Artery-Nerve Dataset [40], a publicly available, clinically acquired and de-identified collection of laparoscopic images from the University of Debrecen. No subjects were prospectively recruited and no direct patient intervention occurred. Eligibility/exclusion and source population details are inherent to the dataset curation; ethics and confidentiality details are provided in the Ethics Statement. The dataset's objective is to enable automatic organ segmentation and differentiation during laparoscopic hysterectomy, targeting ureter, uterine artery, and nerves.

Study Design and Protocol

According to the protocol, we implemented a semantic segmentation pipeline using: (i) transfer learning (TL) within a U-Net [41], (ii) a Vision Transformer (ViT) for long-range context modeling [42], and (iii) a weighted-average ensemble combining a baseline U-Net, OrganFocus U-Net [39], and the ViT [42]. Ensemble weights were selected by grid search over the simplex to maximize validation mean Intersection-over-Union (IoU); the final weights were (0.3, 0.1, 0.6). The dataset was split into training, validation, and test partitions and evaluated primarily by IoU [43].

Technical Information

Dataset and Preprocessing.

The dataset contains 586 RGB images from 38 laparoscopic surgeries, each with expert multiclass masks [40]. Images were resized to 128×128, intensities normalized, and labels encoded as in Table 1. The working corpus comprised 1,218 images (586 originals plus augmented variants). An 80/10/10 image-level split into training, validation, and testing sets, respectively, was applied with class-balance maintenance. Data augmentation applied only to the training split included random horizontal flips, limited rotations, mild intensity perturbations, and elastic deformations within clinically plausible bounds to improve robustness while avoiding artifacts. Augmented variants were generated after partitioning, and all variants of a given original were kept within the same split to prevent leakage.

Table 1. Organ classes encoding

Class Name	Label (Pixel Value)
Background	0
Uterine artery	1
Ureter	2
Nerves	3

Transfer Learning

Transfer learning within a U-Net encoder–decoder [41] leverages weights pre-trained on a source domain and adapts them to laparoscopic organ segmentation. We evaluated six backbones: VGG16 [44], ResNet50 [45], MobileNetV2 [46], MultiResUNet [47], InceptionResNetV2 [48], and EfficientNet-B0 [49], by integrating them one by one into the U-Net encoder. The workflow is shown in Figure 2, and the U-Net architecture for transfer learning models is shown in Figure 3.

Figure 2. Workflow of transfer learning models for organ segmentation, including preprocessing, training, and evaluation steps.

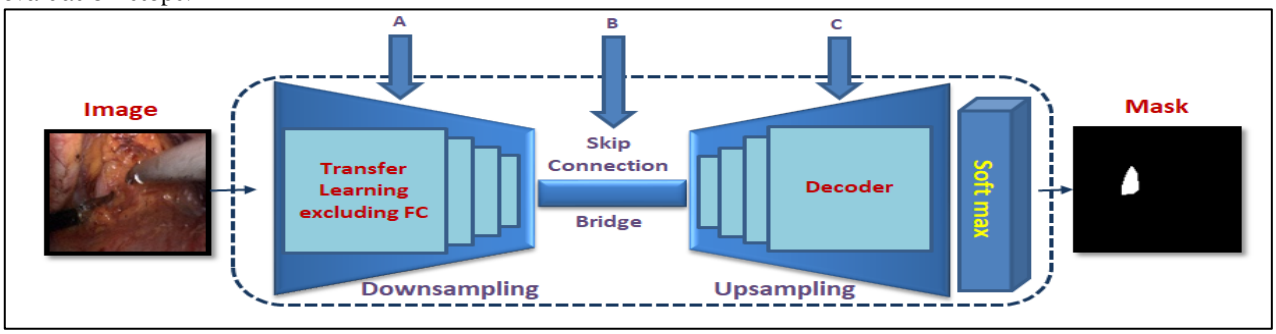


Figure 3. The U-Net architecture of transfer learning models.

Vision Transformers.

To better capture global anatomical context, we integrated a vision transformer [42] into the pipeline, shown in Figure 4. The architecture tokenizes each image into fixed-size patches, applies positional encodings, and processes the sequence with transformer encoders; the outputs are then reshaped into segmentation maps, as shown in Figure 5.

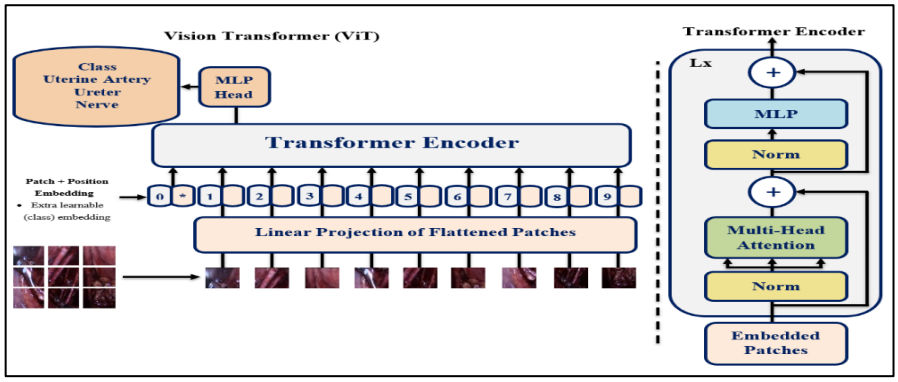


Figure 4. ViT and Transformer encoder schematic. Adapted from [42].

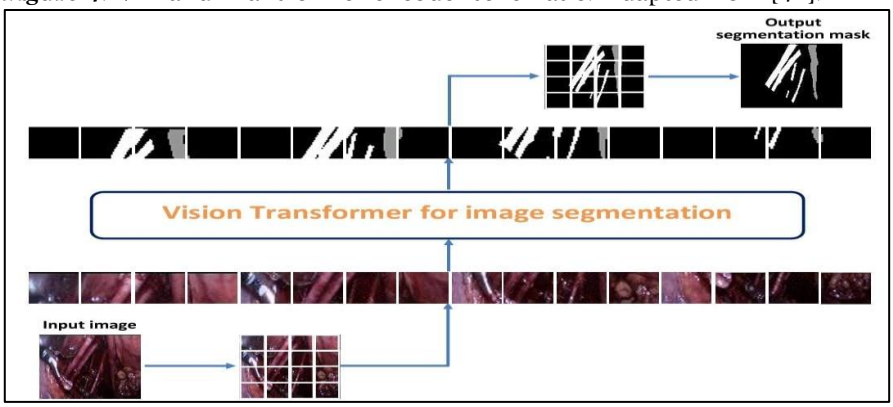


Figure 5. Example input/output for ViT-based segmentation. Adapted from [42].

Ensemble Weighted Average

In this work, three pre-specified base learners OrganFocus U-Net [39], a baseline U-Net [50], and a vision transformer [42] were fused by per-pixel weighted averaging of softmax probability maps. A grid search over the 3-simplex on the validation split was conducted to maximize mean IoU [51]; the selected weights were (0.3, 0.1, 0.6) and were then fixed for test evaluation. Final masks were obtained by applying argmax to the fused probabilities. Figure 6 summarizes the ensemble workflow. Performance was quantified using mean IoU [43], and the IoU definition is shown in Figure 7.

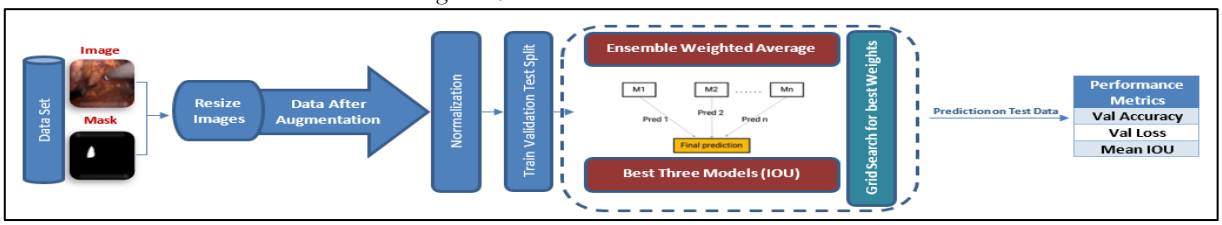


Figure 6. Workflow of the ensemble weighted average technique, highlighting data preprocessing, model selection, ensemble prediction, weight optimization, and performance evaluation.

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Figure 7. Formula for calculating the IoU metric used in semantic segmentation evaluation.

Gradient-weighted Class Activation Mapping (Grad-CAM) was applied systematically across (i) transfer-learning U-Net models, (ii) the vision transformer, and (iii) the weighted ensemble, to generate heatmaps highlighting regions that most influenced predictions. This provided case-level interpretability and verified focus on anatomically relevant structures.

Computing environment.

Training/inference was performed on an Intel® Core™ i7-10750H (Intel Corp., Santa Clara, California, USA), NVIDIA GeForce RTX 2060 (NVIDIA Corp., Santa Clara, California, USA), and 32 GB of RAM.

Software.

Windows; TensorFlow 2.3.0; scikit-learn 2.4.0; NumPy 1.21.6; OpenCV 4.8.0; pandas; SciPy; Matplotlib; PyTorch.

Statistics

Performance was evaluated using mean Intersection-over-Union (mean IoU). We computed per-image confusion matrices and aggregated them to obtain dataset-level estimates. Ninety-five percent confidence intervals (95% CI) were estimated via non-parametric bootstrap over test images (≥5,000 resamples; percentile 2.5–97.5%). All metrics are reported as percentages (×100).

RESULTS

Overall performance

Quantitative results (mean IoU with 95% CI) are summarized in Table 2. Across models, the transformer outperformed individual CNN baselines, and the weighted ensemble attained the highest mean IoU on the test set.

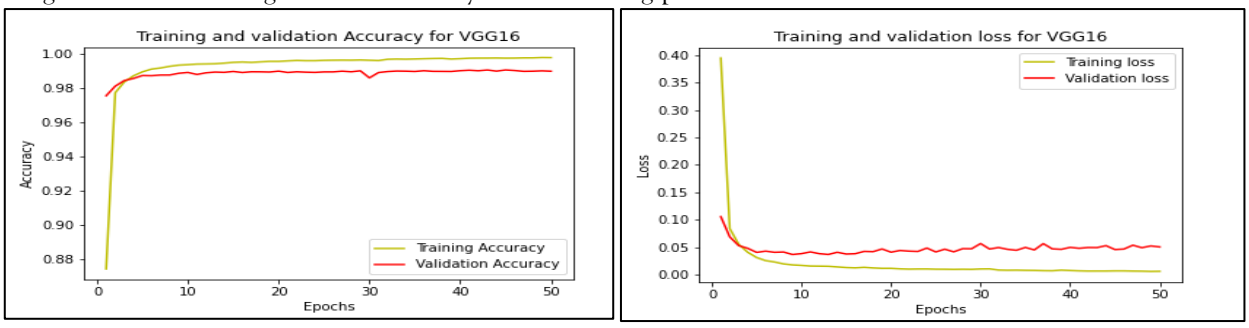
Table 2. Mean IoU (%) with 95% confidence intervals on the test set

Model	Mean IoU (%)	95% CI (lower – upper)
Ensemble	**86.57%**	**84.49 – 87.99**
ViT	84.00%	81.52 – 85.36
VGG16	76.85	76.85 - 73.71
ResNet50	76.07	76.07 - 72.40
MobileNetV2	74.92	74.92 - 71.25
MultiResUNet	74.68	74.68 - 71.26
InceptionResNetV2	71.94	71.94 - 67.89
EfficientNetB0	62.75	62.75 - 58.72

Transfer Learning Performance Analysis

Among transfer learning models, VGG16 achieved the highest performance across the key evaluation metrics, as shown is Table 2. Accordingly, the subsequent subsections will present a detailed analysis and discussion of the VGG16 model's results.

The training and validation performance of the VGG16 model, is depicted in Figure 8. The left plot illustrates the training and validation accuracy across 50 epochs, showing the model's ability to learn and generalize from the data. The right plot displays the training and validation loss over the same number of epochs, providing insight into the convergence and stability of the learning process.

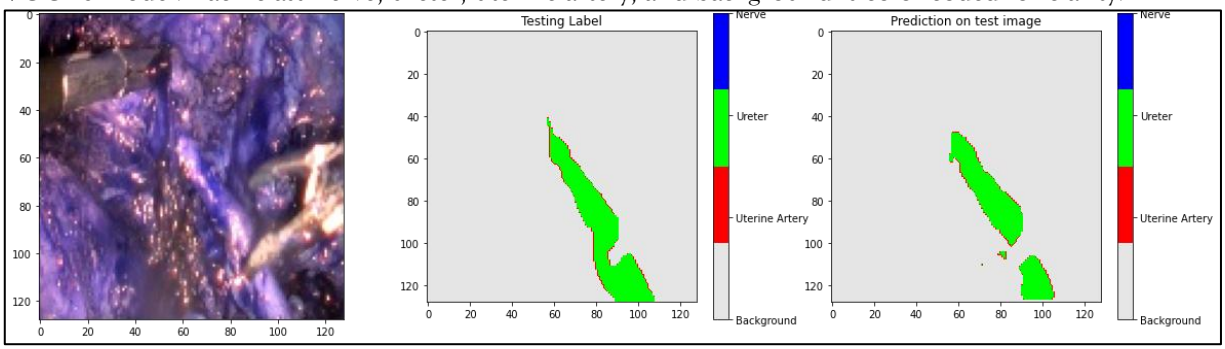


(a) (b)

Figure 8. (a) Training and validation accuracy curves for the VGG16 model, (b) Training and validation loss curves for the VGG16 model.

To qualitatively assess the segmentation performance of the VGG16 model, Figure 9 presents a representative example from the test set. Figure 9(a) shows the original laparoscopic image, Figure 9(b) shows the corresponding

ground truth segmentation mask, and Figure 9(c) shows the predicted segmentation output generated by the VGG16 model. Each class nerve, ureter, uterine artery, and background is color-coded for clarity.

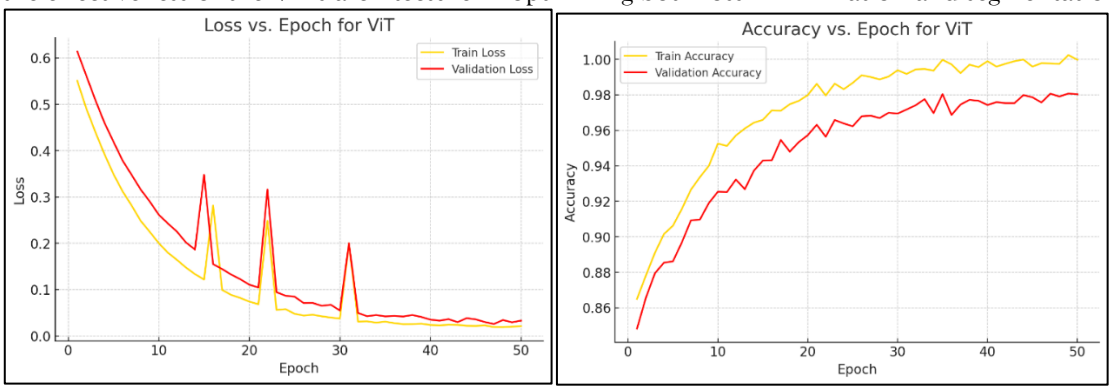


(a) (b) (c)

Figure 9. Example of VGG16 model segmentation results. (a) Original laparoscopic image; (b) Ground truth segmentation mask; (c) Predicted segmentation output by the VGG16 model.

Vision Transformer Performance Analysis

Figure 10 presents the training loss and accuracy curves of the ViT-based segmentation model across 50 epochs. The results show that the model achieves rapid and stable convergence, as indicated by a significant reduction in loss with continued training. At the same time, the accuracy curve demonstrates the model's ability to learn complex patterns, resulting in consistently high accuracy scores throughout training. These findings highlight the effectiveness of the ViTs architecture in optimizing both loss minimization and segmentation accuracy.



(a) (b)

Figure 10. (a) Training and validation loss, and (b) accuracy curves for the ViT model.

Figure 11, illustrates the evolution of the mean IoU scores for the ViTs segmentation model over the course of training. The curve reflects the model's ability to consistently learn and delineate object boundaries with high accuracy across epochs. The ViT model for segmentation achieved a mean IoU of 84% (95% CI: 81.52–85.35%). This result indicates the model's ability to capture image semantics and perform accurate object segmentation.

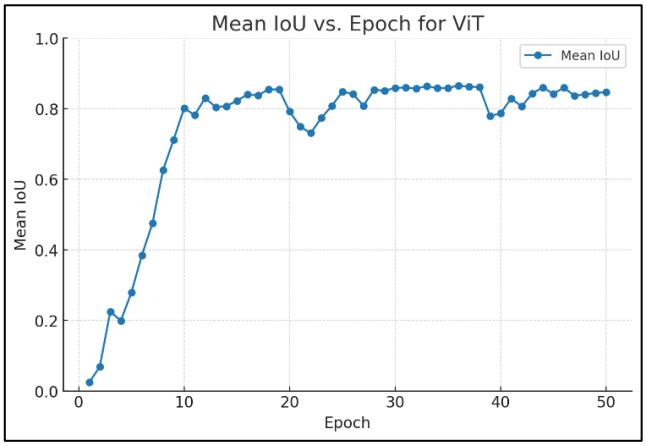


Figure 11. Mean IoU scores over training epochs for the Vision Transformer segmentation model

Figure 12 presents a visual evaluation of ViT model segmentation results, where Figure 12(a) shows the original laparoscopic images, Figure 12(b) shows the ground truth labels, and Figure 12(c) shows the predicted labels generated by our ViT segmentation model. The ground truth labels illustrate the precise delineation of anatomical structures, while the predicted labels highlight the regions identified by the model. This combined

visualization provides an intuitive assessment of the ViT model's segmentation performance relative to the manual annotations.

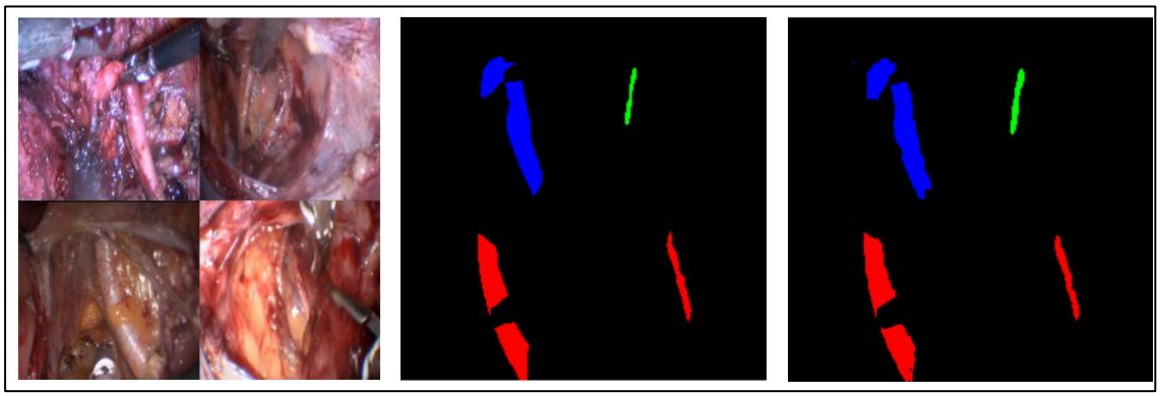


Figure 12. Visual evaluation of ViT model segmentation results. (a) Original laparoscopic images, (b) ground truth segmentation masks, and (c) predicted segmentation masks generated by the Vision Transformer model.

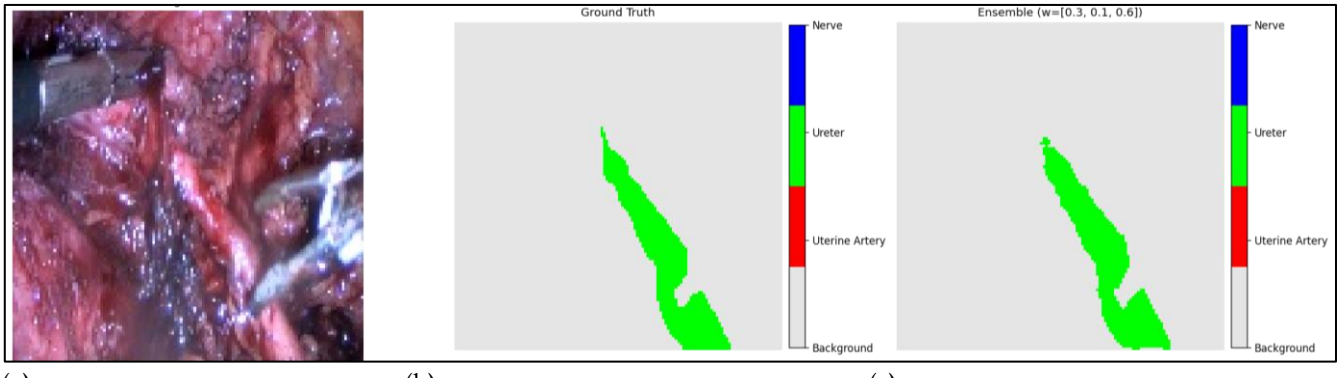
These results highlight the significant potential of ViT-based models for semantic segmentation, particularly in complex medical imaging scenarios where both fine structural details and global context are essential for accurate delineation. The consistent outperformance of ViT, along with the demonstrated effectiveness of transfer learning strategies and ensemble techniques, underscores the strength of the proposed methodology. Collectively, these approaches represent a valuable advancement in automated organ segmentation for laparoscopic hysterectomy, offering a robust foundation for both clinical application and future research.

Ensemble Optimization Result

Ensemble weights were selected on the validation split via grid search over the 3-simplex and then fixed for testing; the final weights were (0.3, 0.1, 0.6). The ensemble achieved 86.57% mean IoU (95% CI: 84.49–87.99%).

This combination of weights resulted in the highest mean IoU accuracy for our ensemble model, showcasing the effectiveness of our weighted average ensemble approach. It's a significant achievement in fine-tuning our model for optimal performance.

The impact of the optimized ensemble weights is visually demonstrated in Figure 13, which presents a representative example from the test set. Figure 13(a) shows the original laparoscopic image, Figure 13(b) shows the ground truth segmentation mask, and Figure 13(c) shows the predicted segmentation output generated by the ensemble model after applying the optimized weights. As observed, the predicted segmentation closely matches the ground truth, particularly in delineating the ureter region, demonstrating the ensemble model's ability to accurately identify anatomical structures after fine-tuning. This qualitative evidence further supports the quantitative improvement achieved by the weighted average ensemble, highlighting its effectiveness in real-world clinical scenarios.



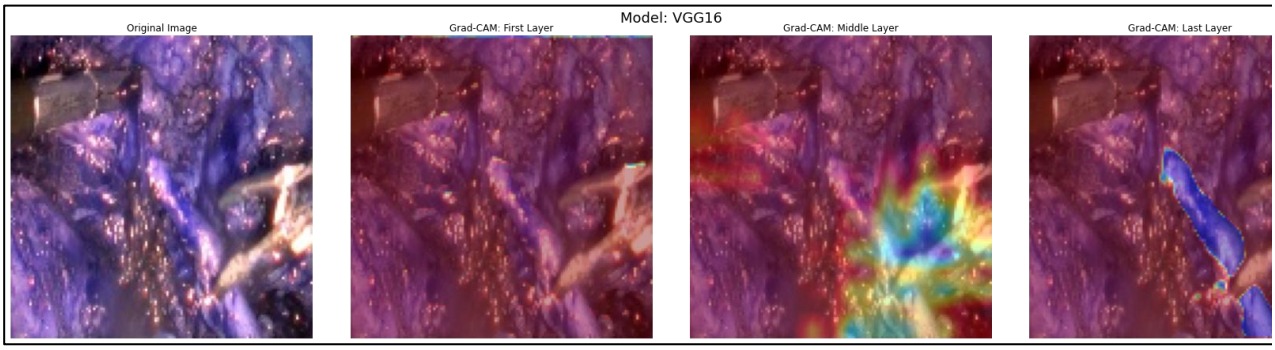
(a) (b) (c)

Figure 13. Example of ensemble prediction after applying grid search optimization. (a) Original laparoscopic image; (b) Ground truth segmentation mask; (c) Predicted segmentation mask by the optimized ensemble model.

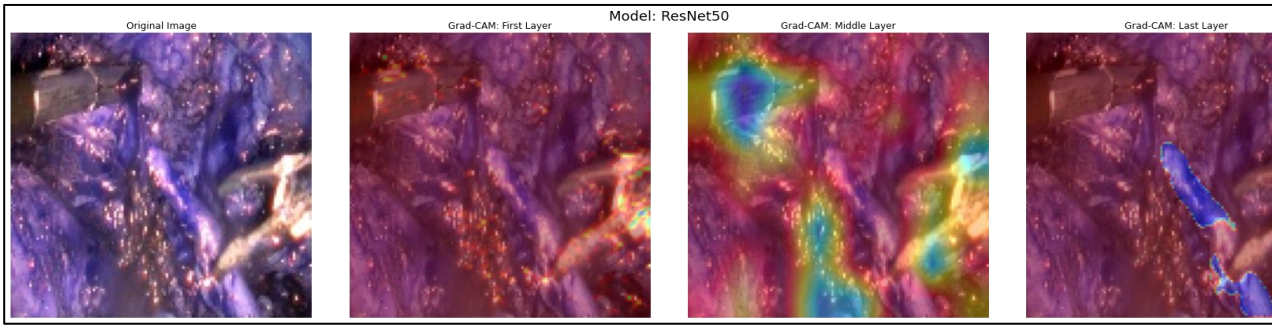
Explainable Artificial Intelligence Analysis for all Models

Grad-CAM overlays demonstrated anatomically coherent focus across models. Transfer Learning backbones exhibited progressively sharper activations in deeper layers, as shown in Figure 14, the Vit model showed hierarchical refinement, as shown in Figure 15, and the ensemble combined broad contextual focus with fine boundary cues, as shown in Figure 16.

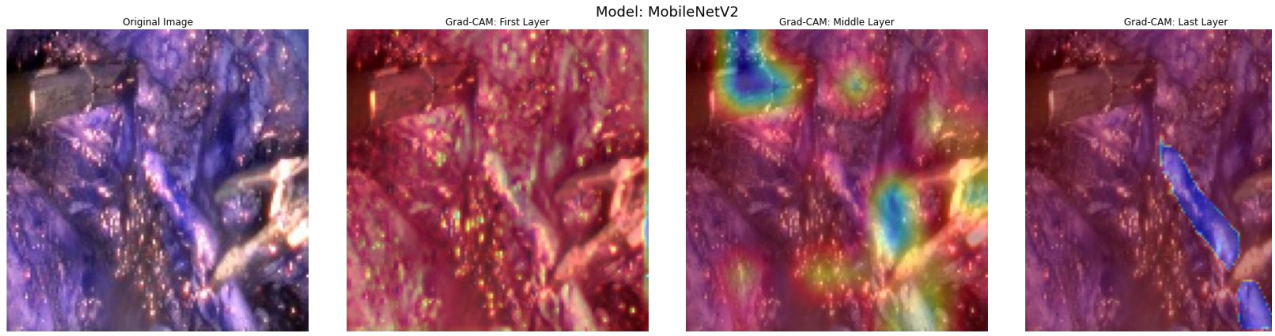
VGG16



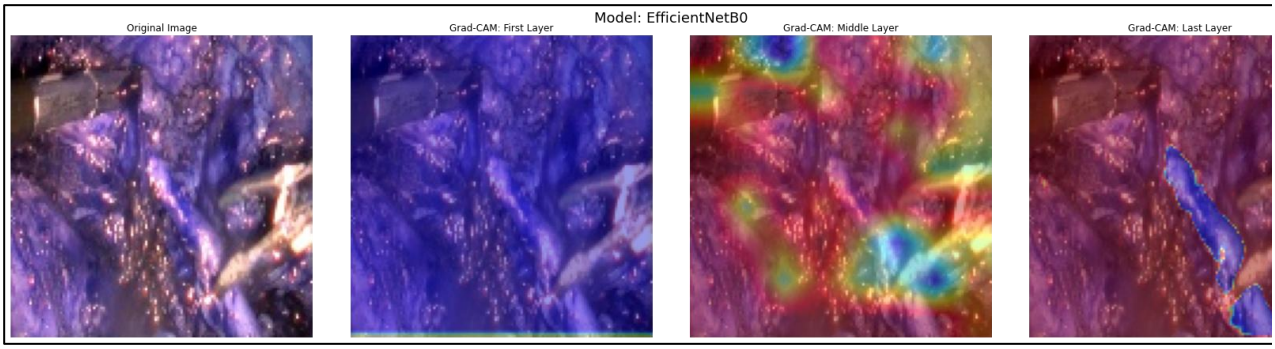
ResNet50



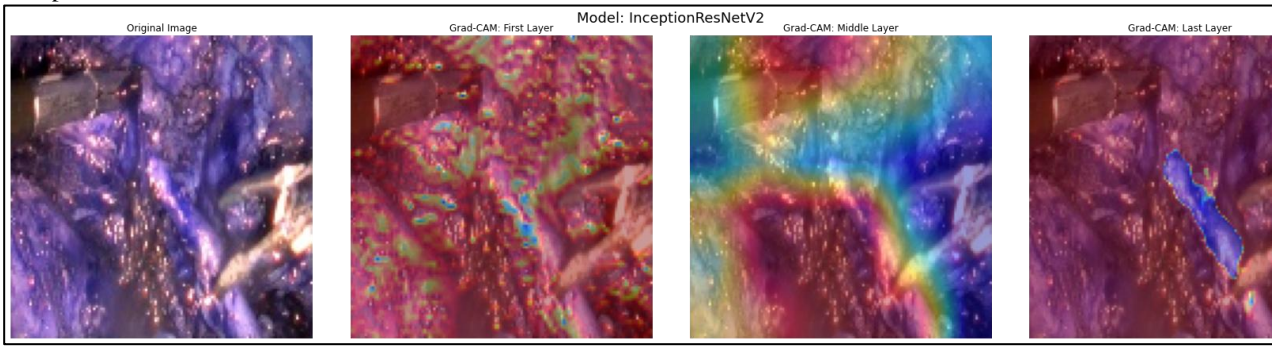
MobileNetV2



EfficientNetB0



InceptionResNetV2



MultiResUNet

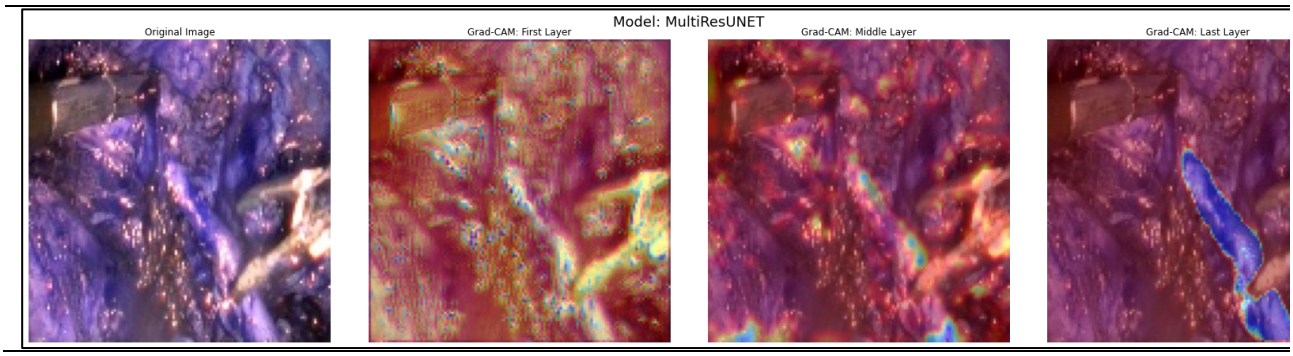


Figure 14. Grad-CAM visualizations for different network layers (first, middle, last) for TL models

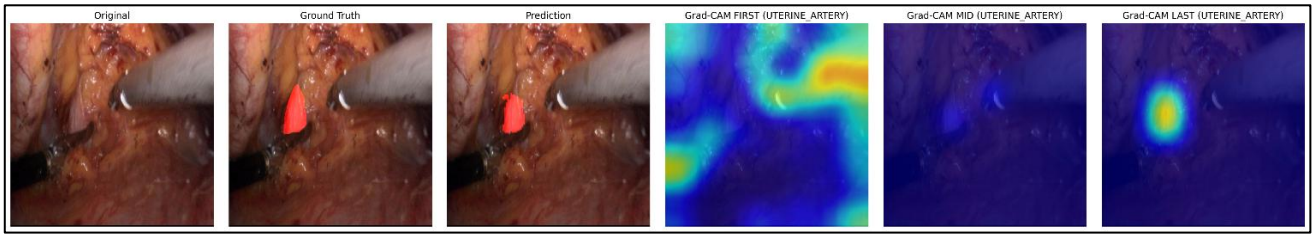


Figure 15. Grad-CAM visualizations of the Vision Transformer model

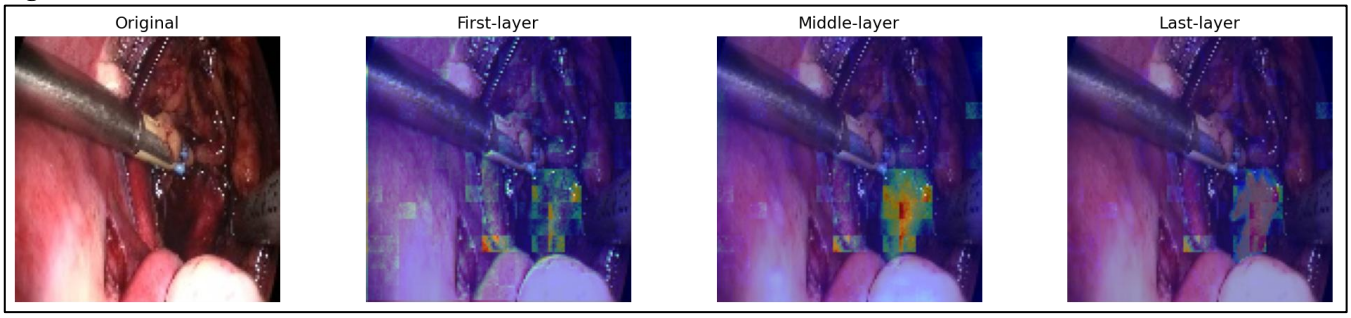


Figure 16. Grad-CAM visualizations of the ensemble model

Mean Intersection-over-Union Comparison Across All Models

For completeness, Figure 17 visualizes mean IoU across all architectures, emphasizing the improvement from the transformer and the ensemble over CNN baselines.

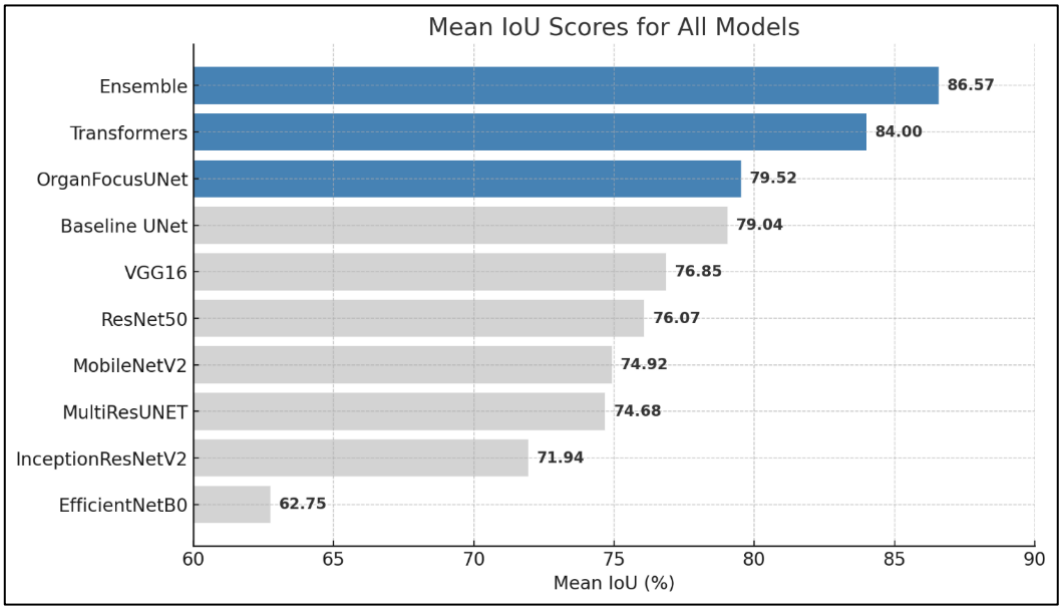


Figure 17. Mean IoU scores for all evaluated architectures

DISCUSSION

Using a clinically acquired dataset of laparoscopic hysterectomy images, this study shows that a Vision Transformer (ViT) and a weight-optimized ensemble outperform transfer-learning (TL) U-Net variants on multiclass organ segmentation. The best TL backbone (VGG16) achieved a mean IoU of 76.85% (95% CI:

73.71–79.76%), the ViT achieved 84.00% (81.52–85.35%), and the weighted ensemble reached 86.57% (84.49–87.99%). Qualitative examples indicate closer adherence to anatomical boundaries, particularly for the ureter, which is thin, tortuous, and often occluded. Grad-CAM heatmaps further corroborate anatomically coherent focus across models.

Interpretation in context. The observed hierarchy aligns with architectural inductive biases. TL U-Nets leverage locality and parameter sharing that favor crisp boundary delineation but can under-represent global dependencies in cluttered laparoscopic scenes. Transformers, by contrast, model long-range relations via self-attention, aiding disambiguation of slender structures within complex backgrounds. Fusing these complementary biases via weighted averaging reduces variance and aggregates error diversity, explaining ensemble gains over single models. These findings are consistent with prior evidence on TL under data constraints and domain shift [2–9], on the benefits of transformer or hybrid encoders for medical segmentation [10–19], and on the superiority of ensembles when constituents are both accurate and diverse [20–31].

Computational considerations and overfitting. Transformers typically incur higher memory and compute than like-for-like CNNs; combined with a modest dataset size, this can elevate overfitting and deployment risk. We mitigated these risks using post-split augmentation, early stopping, weight decay, and ensembling, and we quantified uncertainty via non-parametric bootstrap confidence intervals. Future work will profile FLOPs/throughput and investigate lightweight Transformer variants suited for real-time systems.

Strengths and limitations. Strengths include (i) use of a clinically acquired, publicly available dataset; (ii) explicit train/validation/test splits with augmentations confined to training to prevent leakage; (iii) a weight-optimized heterogeneous ensemble; (iv) bootstrap-based 95% CIs for all models; and (v) systematic Grad-CAM for case-level interpretability. Limitations include single-center data, image-level splits that do not stress patient-level independence, and a 128×128 training resolution that may constrain fine detail. Grad-CAM is post-hoc and does not guarantee causal faithfulness.

Implications. Accurate pixel-wise identification of the ureter, uterine artery, and pelvic nerves can support intraoperative decision-making and potentially reduce iatrogenic injury. The demonstrated gains of ViT and the ensemble, together with interpretable heatmaps, provide a pragmatic path toward clinically oriented assistance tools. Prospective validation on multi-institutional datasets and surgical videos, alongside temporal modeling and surgeon-in-the-loop evaluation of explanations, constitute logical next steps.

CONCLUSIONS

Combining convolutional encoders with a Vision Transformer in a weight-optimized ensemble yields state-of-the-art segmentation performance on a clinically acquired laparoscopic dataset. The ensemble achieved 86.57% mean IoU (95% CI: 84.49–87.99%), outperforming the ViT at 84.00% (81.52–85.35%) and the best transfer-learning U-Net (VGG16) at 76.85% (73.71–79.76%). Grad-CAM offered case-level interpretability, highlighting anatomically relevant regions. Future work will emphasize compute/throughput profiling, lightweight Transformer variants, patient-level and multi-institutional validation, and temporally aware models for surgical video.

List of Abbreviations: AI-Artificial Intelligence; DL-Deep Learning; CNN-Convolutional Neural Network; U-Net-U-shaped Convolutional Network for Biomedical Image Segmentation; ViT-Vision Transformer; TL-Transfer Learning; FCN-Fully Convolutional Network; RCA-Reverse Classification Accuracy; DNN-Deep Neural Network; SVM-Support Vector Machine; XAI-Explainable Artificial Intelligence; IoU-Intersection over Union; RGB-Red, Green and Blue; CT-Computed Tomography; MRI-Magnetic Resonance Imaging; 3D-Three-Dimensional; CPU-Central Processing Unit; GPU-Graphics Processing Unit; RAM-Random Access Memory; UD-University of Debrecen.

Author Contributions: Methodology: M.R.G., M.G. and, A.R.; Software: A.R. and, M.G.; Validation: M.G. and, A.R.; Formal analysis (statistics): A.H.S.S.; Investigation (experiments): A.R. and, M.G.; Resources: O.A.S. and, M.R.G.; Data curation: A.R.; Visualization: A.R.; Writing original draft: A.R. and, M.G.; Writing review and editing: all authors; Supervision: M.R.G.; Project administration: M.R.G.

Funding: This research received no external funding.

Ethics Statement: This study analyzed a public, fully de-identified dataset (UD Ureter–Uterine Artery–Nerve Dataset; curator Norbert Serban, University of Debrecen; last updated 11 Jul 2023; doi:10.21227/q2dd-yt09). No new human data were collected and no re-identifiable information was accessed; therefore, IRB review was not required. All uses complied with the dataset terms, and patient confidentiality was preserved.

Data Availability Statement: All data analyzed in this study are publicly available from IEEE Dataport as the UD Ureter–Uterine Artery–Nerve Dataset curated by Norbert Serban (University of Debrecen); DOI: 10.21227/q2dd-yt09 (last updated 11 Jul 2023). No new datasets were generated or analyzed in this study.

Acknowledgments: We would like to express our sincere gratitude to all the individuals and organizations who made this work possible. We are especially grateful to our university and its faculty members for their administrative and technical support throughout this study. We also thank the journal reviewers for their insightful comments and suggestions, which improved the quality of this manuscript. Finally, we are deeply grateful to our families for their support and encouragement.
Conflict of Interest: The authors declare no conflicts of interest.

REFERENCES

1. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering. 2010;22(10):1345–1359. doi:10.1109/TKDE.2009.191.
2. Raina R, Ng AY, Koller D. Constructing informative priors using transfer learning. In: Proceedings of the 23rd International Conference on Machine Learning (ICML). Pittsburgh (PA); 2006. p. 713–720. doi:10.1145/1143844.1143934.
3. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis. 2019;54:280–296. doi:10.1016/j.media.2019.03.009.
4. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical Image Analysis. 2020;63:101693. doi:10.1016/j.media.2020.101693.
5. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis. 2020;65:101759. doi:10.1016/j.media.2020.101759.
6. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Transactions on Medical Imaging. 2016;35(5):1299–1312. doi:10.1109/TMI.2016.2535302.
7. Valindria VV, Lavdas I, Cerrolaza JJ, et al. Domain adaptation for MRI organ segmentation using reverse classification accuracy. arXiv [Internet]. 2018. https://arxiv.org/abs/1806.00363
8. Valindria VV, Grau V, Glocker B, Rueckert D. Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth. IEEE Transactions on Medical Imaging. 2017;36(8):1597–1606. doi:10.1109/TMI.2017.2665165.
9. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv [Internet]. 2021. https://arxiv.org/abs/2010.11929
10. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. U-NETR: Transformers for 3D medical image segmentation. arXiv [Internet]. 2021. https://arxiv.org/abs/2103.10504
11. Zhou H-Y, Guo J, Zhang Y, Yu L, Wang L, Yu Y. nnFormer: Interleaved transformer for volumetric segmentation. arXiv [Internet]. 2022. https://arxiv.org/abs/2109.03201
12. Cordonnier J-B, Loukas A, Jaggi M. On the relationship between self-attention and convolutional layers. arXiv [Internet]. 2020. https://arxiv.org/abs/1911.03584
13. Tang Y, Chen W, Li Y, et al. Self-supervised pre-training of Swin Transformers for 3D medical image analysis. arXiv [Internet]. 2022. https://arxiv.org/abs/2111.14791
14. Peiris H, Hayat M, Chen Z, Egan G, Harandi M. A robust volumetric transformer for accurate 3D tumor segmentation. arXiv [Internet]. 2022. https://arxiv.org/abs/2111.13300
15. Chen B, Liu Y, Zhang Z, Lu G, Kong AWK. TransAttUnet: Multi-level attention-guided U-Net with transformer for medical image segmentation. arXiv [Internet]. 2022. https://arxiv.org/abs/2107.05274
16. Lin A, Chen B, Xu J, Zhang Z, Lu G. DS-TransUNet: Dual Swin Transformer U-Net for medical image segmentation. arXiv [Internet]. 2021. https://arxiv.org/abs/2106.06716
17. Zhang Y, Liu H, Hu Q. TransFuse: Fusing transformers and CNNs for medical image segmentation. arXiv [Internet]. 2021. https://arxiv.org/abs/2102.08005
18. Deng K, Duan J, Chen Z, et al. TransBridge: A lightweight transformer for left ventricle segmentation in echocardiography. In: Yang G, et al., editors. Lecture Notes in Computer Science. 2021. p. 63–72. doi:10.1007/978-3-030-87583-1_7.
19. Yu X, Yang Q, Zhou Y, et al. UNesT: Local spatial representation learning with hierarchical transformer for efficient medical segmentation. Medical Image Analysis. 2023;90:102939. https://doi.org/10.1016/j.media.2023.102939
20. Peng H, Fan W, Li R, Peng Y, Luo X. U-attention nested U-Transformer for medical image segmentation. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). 2024. p. 1–5. https://doi.org/10.1109/ISBI56570.2024.10635475
21. Pak S, Park SG, Park J, Choi HR, Lee JH, Lee W, Cho ST, Lee YG, Ahn H. Application of deep learning for semantic segmentation in robotic prostatectomy: Comparison of convolutional neural networks and visual transformers. *Investig Clin Urol.* 2024 Nov;65(6):551–558. doi:10.4111/icu.20240159.
22. Wang X, Zhu C, Li J. CTUnet: A novel paradigm integrating CNNs and Transformers for medical image segmentation. In: 2024 International Joint Conference on Neural Networks (IJCNN). 2024. p. 1–6. https://doi.org/10.1109/IJCNN60899.2024.10650785
23. Sun G, Zhang Y, Xu X, Chen H, Wang J. DA-TransUNet: Integrating spatial and channel dual attention with Transformer U-Net for medical image segmentation. *Front Bioeng Biotechnol.* 2024;12:1398237. doi:10.3389/fbioe.2024.1398237.
24. Idri A, Hosni M, Abran A. Improved estimation of software development effort using classical and fuzzy analogy ensembles. Applied Soft Computing. 2016;49:990–1019. doi:10.1016/j.asoc.2016.08.012.
25. Zhou Z-H. Ensemble methods: Foundations and algorithms. Boca Raton (FL): Chapman & Hall/CRC; 2012. doi:10.1201/b12207.
26. Dietterich TG. Ensemble methods in machine learning. In: Multiple Classifier Systems. 2000. p. 1–15. doi:10.1007/3-540-45014-9_1.
27. Hansen LK, Salamon P. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1990;12(10):993–1001. doi:10.1109/34.58871.

28. Schapire RE. The strength of weak learnability. Machine Learning. 1990;5(2):197–227. doi:10.1023/A:1022648800760.
29. Breiman L. Bagging predictors. Machine Learning. 1996;24(2):123–140. doi:10.1023/A:1018054314350.
30. Ho TK. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998;20(8):832–844. doi:10.1109/34.709601.
31. Seni G, Elder J. Ensemble methods in data mining: Improving accuracy through combining predictions. San Rafael (CA): Morgan & Claypool; 2010. doi:10.2200/S00240ED1V01Y200912DMK002.
32. Zhou Z-H, Feng J. Deep forest: Towards an alternative to deep neural networks. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17). 2017; p. 3553–3559. doi:10.24963/ijcai.2017/497.
33. Utkin LV, Kovalev MS, Meldo AA. A deep forest classifier with weights of class probability distribution subsets. Knowledge-Based Systems. 2019;173:15–27. doi:10.1016/j.knosys.2019.02.022.
34. Nguyen TT, Van Pham N, Dang MT, Luong AV, McCall J, Liew AWC. Multi-layer heterogeneous ensemble with classifier and feature selection. In: Proceedings of the 2020 Genetic and Evolutionary Computation Conference (GECCO). 2020. p. 725–733. doi:10.1145/3377930.3389832.
35. Qi Z, Wang B, Tian Y, Zhang P. When ensemble learning meets deep learning: A new deep support vector machine for classification. Knowledge-Based Systems. 2016;107:54–60. doi:10.1016/j.knosys.2016.05.055.
36. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A review of machine learning interpretability methods. Entropy. 2020;23(1):18. doi:10.3390/e23010018.
37. Dardouillet P, Benoit A, Amri E, Bolon P, Dubucq D, Créd A. Explainability of image semantic segmentation through SHAP values. ICPR-XAIE Workshop, 26th International Conference on Pattern Recognition. 2022. https://hal.science/hal-03719597
38. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). 2017. p. 618–626. doi:10.1109/ICCV.2017.74.
39. Gomaa M, Rabie A, Saad AHS, Sadek OA, Girgis MR. Improved semantic segmentation in medical imaging using U-Net and attention mechanisms. Journal of Information Systems Engineering and Management. 2025;10(18). doi:10.52783/jisem.v10i18s.2891.
40. Serban N. UD Ureter–Uterine Artery–Nerve Dataset [dataset on the Internet]. IEEE Dataport; 2023 Jul 11. https://ieee-dataport.org/documents/ud-ureter-uterine-artery-nerve-dataset. doi:10.21227/q2dd-yt09.
41. Cheng D, Lam EY. Transfer learning U-Net deep learning for lung ultrasound segmentation. arXiv [Internet]. 2021. https://arxiv.org/abs/2110.02196
42. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv [Internet]. 2017. https://arxiv.org/abs/1706.03762
43. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles. 1901;37:547–579.
44. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv [Internet]. 2015. https://arxiv.org/abs/1409.1556
45. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv [Internet]. 2015. https://arxiv.org/abs/1512.03385
46. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: Inverted residuals and linear bottlenecks. arXiv [Internet]. 2018. https://arxiv.org/abs/1801.04381
47. Ibtehaz N, Rahman MS. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Networks. 2020;121:74–87. doi:10.1016/j.neunet.2019.08.025.
48. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. arXiv [Internet]. 2016. https://arxiv.org/abs/1602.07261
49. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. arXiv [Internet]. 2019. https://arxiv.org/abs/1905.11946
50. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). 2015. p. 234–241. https://arxiv.org/abs/1505.04597
51. Anand V, et al. Weighted average ensemble deep learning model for stratification of brain tumor in MRI images. Diagnostics. 2023;13(7):1320. doi:10.3390/diagnostics13071320.