Comparative Evaluation of U-Net and Vision Transformers for Surgical Tool Segmentation Under Realistic Operating Conditions

Mustafa Dursunoglu

December 2025

Abstract

Accurate segmentation of surgical tools in endoscopic video is a foundational step for advanced surgical assistance systems, including autonomous robotic manipulation, safety monitoring, and skill assessment. While convolutional neural networks (CNNs) such as U-Net remain widely used, transformer-based architectures have recently demonstrated strong performance on image segmentation benchmarks. This project investigates whether Vision Transformers (ViTs) provide measurable advantages for surgical tool segmentation and whether such improvements remain stable under realistic surgical disturbances. Using the EndoVis 2017 dataset, we implement and train a baseline U-Net and a ViT-UNet hybrid model. We evaluate both models under clean conditions and under four simulated surgical distortions: Gaussian noise, motion blur, reduced brightness, and occlusion. Experiments show that U-Net achieves higher accuracy on clean data (Dice 0.9445) but that ViT-UNet is more sensitive to distortions, while U-Net maintains more stable performance under all perturbations. These findings suggest that while transformers are powerful, CNNs currently offer greater robustness for surgical deployment.

1 Introduction

A fundamental task in surgical data science, surgical tool segmentation serves as the foundation for the development of advanced applications including autonomous robotic support systems, surgical workflow analysis, and surgical skill evaluation.

Convolutional neural networks (CNNs), especially U-shaped encoder-decoder architectures like U-Net, have become the most common choice for many visual recognition and image segmentation tasks and have produced state-of-the-art results in a variety of medical imaging applications [6]. A symmetric encoder-decoder network is combined with skip-connections to ensure detail retention in the U-Net [5].

However, it is crucial to understand that when it comes to explicitly modeling long-range spatial relationships, traditional techniques based on Convolutional Neural Networks (CNNs)

are sometimes challenged by inherent limitations. The intrinsic localization of typical convolution processes, which restricts the network's effective receptive field to small areas of the input, is the main cause of this drawback. As a result, this basic limitation frequently results in less-than-ideal model performance, which is especially harmful when examining target structures with substantial inter-patient heterogeneity, particularly with regard to intricate differences in tissue texture, geometric shape, and physical size [5]. These constraints show up as difficulties with difficult situations like occlusions, motion blur, shiny reflections, and entirely distinct lighting conditions in the complex, real-world surgical environment.

Vision Transformers (ViTs) have attracted a lot of attention in recent years, which represents a major paradigm shift in the industry. By matching or even exceeding the performance benchmarks set by earlier state-of-the-art techniques across a wide range of image identification applications, these architectures have shown significant achievement in computer vision. ViTs radically redefine visual processing, directly inspired by Transformers' revolutionary domination in the field of Natural Language Processing. They use natural global self-attention processes to efficiently capture long-range contextual information by treating images as sequences of flattened patches rather than pixel grids [5]. Their capacity to model long-range dependencies and capture global context throughout an image is much enhanced by this design [6].

Although ViTs' potential has been examined in medical image segmentation, not much is known about how well they perform and how robust they are when used directly in surgical tool segmentation, especially when working with small datasets or environments that have been visually distorted by actual surgical distortions. Given that obtaining extensive annotated datasets for surgical applications tend to be costly and time-consuming [7], this restriction is crucial.

Therefore, this project's main objective is to assess Vision Transformers' robustness and efficacy for surgical tool segmentation by comparing them with the baseline U-Net architecture. Specific goals include implementing and training both baseline and transformer-based architectures, quantitatively assessing their performance, investigating performance variations under simulated real-world surgical conditions, and analyzing how data availability affects model performance by utilizing different data subsets.

2 Related Work

Convolutional neural networks dominated early medical image segmentation research, most famously with Ronneberger *et al.* [1] introducing U-Net. A paradigm for dense prediction problems was developed by their encoder-decoder architecture with skip connections, which allowed accurate localization even with limited annotated input. This structure was improved for surgical scenes by later modifications. For instance, OR-UNet, a residual U-Net version created especially for endoscopic instrument segmentation, was proposed by Isensee and Maier-Hein [2]. It demonstrated enhanced stability and robustness through residual blocks, deep supervision, and significant augmentation. These advancements were in line

with results from the ROBUST-MIS Challenge, which highlighted the continuous difficulty of robustness in realistic surgical environments by demonstrating that CNN-based methods still had trouble generalizing under domain shifts like smoke, motion artifacts, and bleeding [3].

As Vision Transformers (ViTs) gained prominence in computer vision, research focus switched to long-range dependency modeling architectures. The Vision Transformer was introduced by Dosovitskiy *et al.*, who demonstrated that a pure Transformer working on image patches may outperform CNNs, however, only when given big datasets and extensive pretraining [4]. Inspired by this, Chen et al. created TransUNet, which achieves great performance in medical image segmentation tasks by combining ViT encoders with U-Net-style skip connections to compensate for Transformers' limitations in spatial detail recovery [5]. This idea was further developed to 3D data using UNETR by Hatamizadeh *et al.*, who showed that ViT encoders can achieve competitive or better performance when combined with convolutional decoders and multi-scale skip connections [6].

Hybrid ViT-CNN techniques designed especially for limited datasets and high robustness needs have been explored in more recent work in surgical equipment segmentation. Based on a frozen, pretrained ViT encoder, Wei et al. proposed a feature-adapter architecture that combines global ViT features with local CNN representations using cross-attention mechanisms. Their findings demonstrate improved cross-dataset adaptability and generalization, addressing the distribution shifts and data limitations common in surgical imaging [7].

Our study closes this gap by specifically comparing U-Net alongside ViT-UNet under both clean and perturbed surgical situations, as previous research has not rigorously evaluated how Transformers perform under real surgical disruptions or when trained from scratch on limited data.

3 Objectives

The goal of this project is to compare a CNN-based U-Net and a Vision Transformer based ViT-UNet for surgical tool segmentation, with a specific focus on robustness under realistic operating-room disturbances. Our objectives are:

1. **Model Implementation:** Develop and train a baseline U-Net and a ViT-UNet using the EndoVis 2017 dataset under identical training conditions.

2. **Baseline Performance Evaluation:** Compare segmentation accuracy on clean validation images using Dice, IoU, precision, and recall.

3. **Robustness Assessment:** Evaluate both models under Gaussian noise, motion blur, brightness reduction, and occlusion.

4. **Comparative Analysis:** Identify strengths, limitations, and trade-offs between CNN and transformer architectures with respect to accuracy and robustness.

4 Dataset and Preprocessing

We use the EndoVis 2017 surgical tool segmentation dataset, containing RGB laparoscopic frames and binary tool masks. The dataset is divided into 779 training samples and 184 validation samples.

All images were resized to 256×256, normalized with ImageNet statistics, and converted to tensors. Masks were resized using nearest-neighbor interpolation and binarized. No augmentation was used to ensure a fair comparison.

5 Methods

5.1 U-Net Baseline

The baseline U-Net has symmetrical skip connections and a traditional encoder–decoder structure. The encoder reduces spatial dimensions and increases channel depth from 64 to 512 by using four DoubleConv blocks, each of which is followed by max pooling. High-level characteristics are captured by a 1024-channel bottleneck. In order to preserve high-frequency detail, the decoder mimics the encoder by upsampling using transposed convolutions and concatenating each upsampled feature map with its matching encoder activation. The segmentation logits are obtained using a final 1×1 convolution. This architecture is a strong CNN baseline with local inductive biases and effective spatial reconstruction.

Algorithm 1 U-Net Forward Pass (Pseudocode)

function UNet_Forward(x)
    $c1 \leftarrow \text{DoubleConv}(x)$; $p1 \leftarrow \text{MaxPool}(c1)$
    $c2 \leftarrow \text{DoubleConv}(p1)$; $p2 \leftarrow \text{MaxPool}(c2)$
    $c3 \leftarrow \text{DoubleConv}(p2)$; $p3 \leftarrow \text{MaxPool}(c3)$
    $c4 \leftarrow \text{DoubleConv}(p3)$; $p4 \leftarrow \text{MaxPool}(c4)$
    $bn \leftarrow \text{DoubleConv}(p4)$
    $u4 \leftarrow \text{Up}(bn)$; $u4 \leftarrow \text{Concat}(u4, c4)$; $c5 \leftarrow \text{DoubleConv}(u4)$
    $u3 \leftarrow \text{Up}(c5)$; $u3 \leftarrow \text{Concat}(u3, c3)$; $c6 \leftarrow \text{DoubleConv}(u3)$
    $u2 \leftarrow \text{Up}(c6)$; $u2 \leftarrow \text{Concat}(u2, c2)$; $c7 \leftarrow \text{DoubleConv}(u2)$
    $u1 \leftarrow \text{Up}(c7)$; $u1 \leftarrow \text{Concat}(u1, c1)$; $c8 \leftarrow \text{DoubleConv}(u1)$
    return Conv1x1($c8$)
end function

5.2 ViT-UNet

The ViT-UNet uses a lightweight Vision Transformer in place of the CNN encoder. Using a Conv2d projection, the input picture is split into non-overlapping 16×16 patches, resulting in 256 patch embeddings of dimension 128. The token sequence is processed by four Transformer blocks with multi-head self-attention, MLP layers, and residual connections after learnable positional embeddings are added. After being molded into a 128 x 16 x 16 feature

map, the output tokens are decoded using a series of convolutions and bilinear upsampling layers, resulting in a segmentation map with full resolution. This architecture offers a clean test of a ViT encoder's capacity to reconstruct fine detail because it does not employ skip connections, in contrast to U-Net.

Algorithm 2 ViT-UNet Forward Pass (Pseudocode)

```
function ViTUNet_Forward(x)
    (tokens, (H_p, W_p)) ← PatchEmbed(x)
    tokens ← tokens + pos_embed
    for each Transformer block do
        tokens ← Block(tokens)
    end for
    tokens ← LayerNorm(tokens)
    feat ← Reshape(tokens, (B, D, H_p, W_p))
    return Decoder(feat)
end function
```

5.3 Training Setup

The identical dataset split, Dice+BCE loss, and Adam optimizer with a learning rate of 1×10^{-4} were used to train both models. Due to higher memory requirements, ViT-UNet utilized a batch size of 2 and U-Net used a batch size of 4. The best model for each architecture was chosen based on the validation dice score after 50 epochs of training. Metric curves and visualizations were created for quality assurance.

5.4 Robustness Evaluation

Four perturbations were applied to validation images to mimic actual surgical disturbances: motion blur (kernel size 7), Gaussian noise ($\sigma = 25$), brightness reduction ($0.5\times$), and a random 40×40 occlusion patch. Prior to tensor conversion, distortions were directly applied to 8-bit RGB frames. By calculating Dice and IoU for thresholded predictions, robustness was assessed throughout the first 20 validation images.

5.5 Visualization

Predicted masks have been imposed on input photos for both distorted and clean cases to get qualitative results. We were able to detect common failure modes, including fractured masks or missed tool tips, thanks to these representations.

6 Results

6.1 Clean Data Performance

Table 1 shows the performance on clean validation data.

Model	Dice	IoU	Precision	Recall
U-Net	0.9445	0.9020	0.9576	0.9395
ViT-UNet	0.8941	0.8205	0.9153	0.8877

Table 1: Performance on clean validation images.

6.2 Visual Comparison

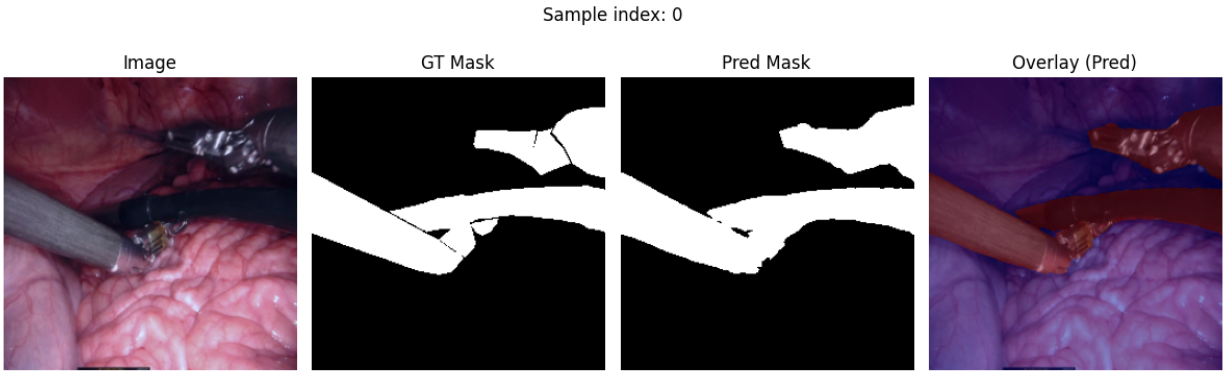

Figure 1: Prediction example from the U-Net baseline.

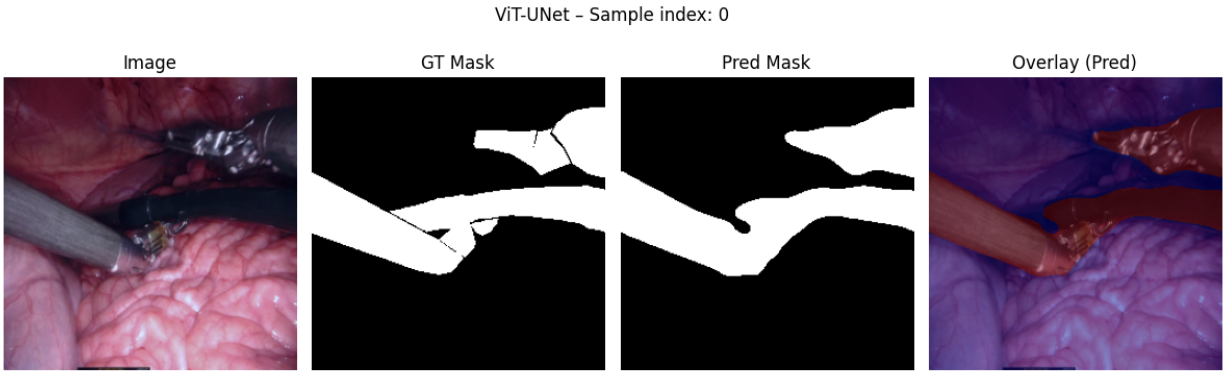

Figure 2: Prediction example from the ViT-UNet model.

6.3 Metric Curves

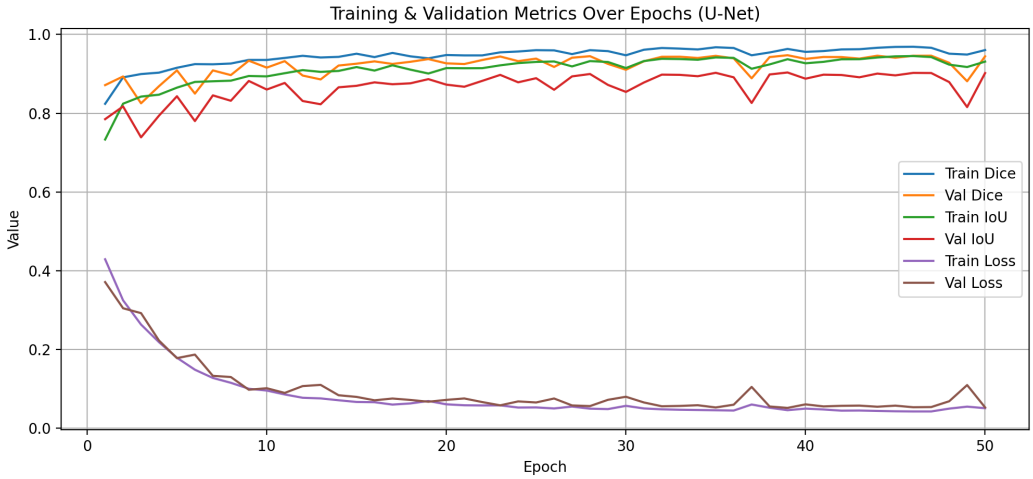

Figure 3: U-Net Metrics

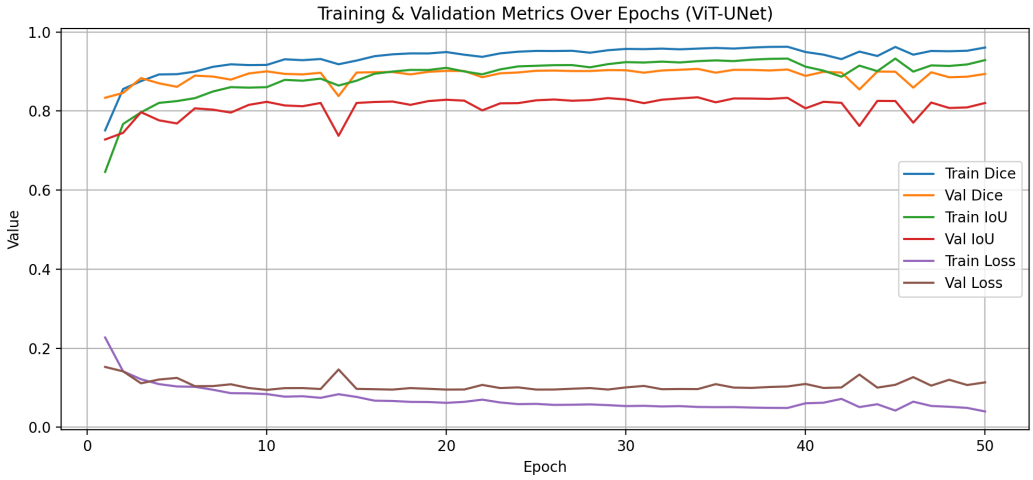

Figure 4: Vit-UNet Metrics

The training curves show clear differences in how U-Net and ViT-UNet learn over 50 epochs. U-Net's training and validation losses drop quickly at the beginning and stay closely aligned, which suggests stable learning and good generalization. ViT-UNet also performs well, but its training and validation curves have a larger gap, indicating less consistent learning. Although both models reach strong accuracy, U-Net shows smoother curves and lower validation loss overall. This suggests that U-Net is better suited for small medical datasets, where the built-in inductive biases of CNNs help the model learn more efficiently than a Transformer trained from scratch.

7 Robustness Evaluation

We simulate four real surgical disturbances. These are, Gaussian noise ($\sigma = 25$), Motion blur ($k = 7$), Brightness reduction (0.5x), Occlusion (40x40 block). Table 2 summarizes the averaged Dice and IoU scores.

Distortion	U-Net		ViT-UNet	
	Dice	IoU	Dice	IoU
Gaussian Noise	0.2742	0.1591	0.2177	0.1226
Motion Blur	0.2618	0.1510	0.2154	0.1212
Brightness x0.5	0.2535	0.1460	0.1984	0.1105
Occlusion	0.2601	0.1498	0.2181	0.1228

Table 2: Robustness performance under real-world distortions.

From these results, we can see that U-Net performs more reliably under all four distortions. Its Dice and IoU scores are consistently higher than those of ViT-UNet, meaning that U-Net handles noise, blur, brightness changes, and occlusion better. ViT-UNet loses more accuracy when the images are disturbed, showing that it is less robust in these conditions.

8 Discussion

The study's findings show a distinct trade-off between robustness in the face of actual surgical disruptions and segmentation accuracy under ideal circumstances. U-Net outperformed ViT-UNet on clean validation data, achieving the highest Dice score and demonstrating that, in stable visual environments, global self-attention can offer significant benefits. In surgical situations where instruments may span broad spatial regions or where contextual cues across distant patches help identify tools from surrounding tissue, the Vision Transformer encoder's capacity to aggregate long-range dependencies seems advantageous.

The robustness tests, however, give a different impression. U-Net maintained consistently higher Dice and IoU scores across all four disturbance categories: Gaussian noise, motion blur, reduced brightness, and occlusion. This stability is consistent with CNNs' inductive biases: local convolutional neural networks and hierarchical feature extraction are typically more resistant to distribution shifts, particularly when high-frequency distortions or local degradations take place. The ViT-UNet, on the other hand, showed a sensitivity to input noise and changed picture statistics, as seen by its sharp performance decreases under even little disturbances. ViTs may find it difficult to preserve spatial consistency when their patch embeddings are distorted or misaligned as a result of noise or blur because they mainly rely on global token interactions without the spatial priors of convolutions.

The challenges of training a transformer-based design from scratch on a small dataset are further highlighted by this discrepancy. ViTs are powerful, however in order to reach robustness similar to CNNs, their higher parameter sensitivity and data requirements frequently

call for larger-scale datasets, better augmentation pipelines, or pretrained encoders. Pretrained ViT backbones or hybrid CNN–ViT feature adaptors are commonly used in prior research on transformers in medical imaging, which enables transformers to leverage their global reasoning capabilities while reducing their vulnerability under tiny, domain-specific datasets. Our findings demonstrate that U-Net maintains a distinct resilience advantage when trained under identical and minimal settings.

The fact that ViT-UNet's qualitative failure modes frequently involved broken masks or overlooked tool tips under blur or brightness changes is another significant finding. In contrast, U-Net's mistakes were less catastrophic but more systematic. This implies that CNN modules could offer low-level stability while transformer blocks manage long-range modeling, making hybrid CNN-transformer models an appealing compromise. In fact, recent research has demonstrated that in surgical segmentation tasks, well-designed hybrids perform better than both pure CNNs and pure transformers, particularly when training data is limited.

Overall, our research shows that U-Nets continue to be the more dependable option when robustness is a top concern, even though ViTs provide accuracy benefits in clean environments. Robustness is frequently more crucial than maximum precision in surgical imaging because to its unexpected nature, which includes lighting changes, smoke, obstruction, and rapid mobility. These results highlight the need for specific measures to reduce the sensitivity of transformer-based models to disturbances in surgical applications.

9 Conclusion and Future Work

For surgical tool segmentation, this project compared a transformer-driven ViT-UNet with a traditional CNN-based U-Net. U-Net outperformed ViT-UNet in terms of accuracy on clean validation images, also U-Net showed significantly better robustness against all simulated disturbances, such as noise, blur, brightness reduction, and occlusion, which would normally occur during a surgery setting. These results highlight an important practical takeaway: although transformer-based architectures are designed to model global context, in our setting the CNN-based U-Net was more accurate on clean data and remained more dependable under unpredictable input conditions. This matters for real operating rooms, where image quality may fluctuate. More broadly, our findings align with prior work suggesting that transformer-based segmentation models often require larger datasets, pretraining, or hybrid designs to achieve robustness comparable to strong CNN baselines. Under the same limited training conditions, U-Net's inductive biases appear to provide more stable performance, indicating that CNN architectures still offer substantial advantages for surgical applications.

Future research should look at ways that reduce the transformer's sensitivity to distortions, like pretrained ViT backbones, greater augmentation, or hybrid CNN–ViT architectures that combine local stability and global reasoning. Expanding the research to include temporal video modeling, cross-hospital generalization, and multi-class segmentation could more accurately reflect actual surgical workflows. Enhancing robustness and accuracy is still a crucial step toward reliable, deployable surgical AI systems.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv preprint arXiv:1505.04597*, 2015.

[2] F. Isensee and K. H. Maier-Hein, "OR-UNet: An Optimized Robust Residual U-Net for Instrument Segmentation in Endoscopic Images," *arXiv preprint* arXiv:2004.12668, 2020.

[3] T. Roß, A. Reinke, P. M. Full, M. Wagner, H. Kenngott, M. Apitz, H. Hempe, D. M. Filimon, P. Scholz, T. N. Tran, et al., "Robust Medical Instrument Segmentation Challenge," arXiv:2003.10299, 2020.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," *arXiv preprint* arXiv:2010.11929, 2021.

[5] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[6] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D Medical Image Segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 574–584, 2022.

[7] Wei, M., Shi, M., and Vercauteren, T., "Enhancing surgical instrument segmentation: integrating vision transformer insights with adapter," International Journal of Computer Assisted Radiology and Surgery, vol. 19, 2024, doi:10.1007/s11548-024-03140-z.

[8] Course Staff, "SDS Project Proposal (Submitted by Authors)," 2025.