# Surg-SegFormer: A Dual Transformer-Based Model for Holistic Surgical Scene Segmentation

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Holistic surgical scene segmentation in robot-assisted surgery (RAS) enables surgical residents to identify various anatomical tissues, articulated tools, and critical structures, such as veins and vessels. Given the firm intraoperative time constraints, it is challenging for surgeons to provide detailed real-time explanations of the operative field for trainees. This challenge is compounded by the scarcity of expert surgeons relative to trainees, making the unambiguous delineation of go- and no-go zones inconvenient. Therefore, high-performance semantic segmentation models offer a solution by providing clear postoperative analyses of surgical procedures. However, recent advanced segmentation models rely on user-generated prompts, rendering them impractical for lengthy surgical videos that commonly exceed an hour. To address this challenge, we introduce Surg-SegFormer, a novel prompt-free model that outperforms current state-of-the-art techniques. Surg-SegFormer attained a mean Intersection over Union (mIoU) of 0.80 on the EndoVis2018 dataset and 0.54 on the EndoVis2017 dataset. By providing robust and automated surgical scene comprehension, this model significantly reduces the tutoring burden on expert surgeons, empowering residents to independently and effectively understand complex surgical environments.

## 1 Introduction

Accurate decision-making in robot-assisted surgery (RAS) requires a thorough understanding using computer vision models (1). These models need to identify and segment anatomical structures and articulated tools to interpret and understand the relation between objects within the scene (2). Nevertheless, accurate and comprehensive surgical scene segmentation remains a significant challenge due to the complexity of anatomical structures and the dynamic nature of the surgical environment.

Entry-level surgeons can benefit from using these models to convert surgical scenes into self-explanatory videos, as they are not accustomed to how these structures appear in live surgery settings (3). Simply, the output video highlights critical zones and detects various articulated tools within the frame. Moreover, such automation frees expert surgeons from suspending the operation to answer the trainees' questions (4). Once objects within the surgical scene are accurately identified, understanding the procedure becomes significantly easier.

Cutting-edge segmentation models, e.g., AdaptiveSAM (5), exhibit excellent performance; however, their dependence on manual prompts restricts their autonomy and scalability in real surgical practice. This limitation is particularly significant in post-operative analysis, as surgical videos often exceed three hours, making manual prompting infeasible. In comparison, models like ISINet (6), SegNet (7), and Ternaus (8) are more efficient and better suited for large-scale automated surgical analysis. Despite their autonomy, the promptable model still outperforms.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

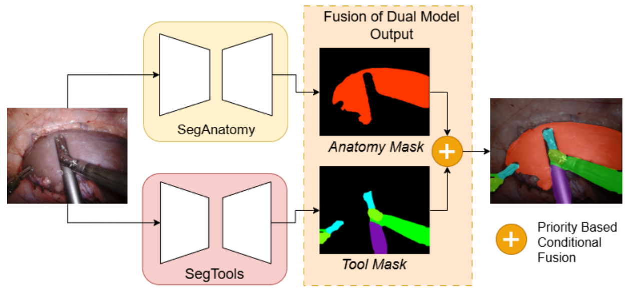


Figure 1: Surg-SegFormer Architecture

To overcome these limitations, we extend SegFormer (9) by developing a dual-instance pipeline. The first instance employs the SegFormer B2 variant, fine-tuned exclusively for anatomical structure segmentation—referred to as SegAnatomy. The second instance uses B5 variant encoder and incorporates a custom-designed, lightweight decoder optimized for segmenting articulated surgical tools, which we refer to as SegTool. In the end, the outputs of the two instances are fused using a priority-weighted conditional fusion strategy, offering comprehensive and consistent segmentation of surgical frames. We call this complete pipeline Surg-SegFormer.

This paper has three main contributions:

1. *Dual-Model Segmentation Framework:* A framework for robotic-assisted surgery (RAS) that uses two distinct models specialized in segmenting anatomical structures and surgical instruments.

2. *Priority-Weighted Conditional Fusion Strategy:* An advanced fusion strategy that combines both model outputs, prioritizing valuable segmentation cues to enhance overall accuracy and robustness.

3. *Comprehensive Evaluation on Benchmark Datasets:* Validation of our framework on two benchmark datasets, demonstrating superior segmentation performance compared to current state-of-the-art (SOTA) methods.

# 2 Methods

## 2.1 Model Overview

We propose **Surg-SegFormer**, a dual-model framework that leverages two SegFormer instances and fuses their outputs. Figure 1 shows the model's architecture and the connection between the two instances. The first instance, **SegAnatomy**, is fine-tuned specifically on anatomical structures. The second instance, **SegTool**, uses a SegFormer encoder fine-tuned for tool segmentation and replaces the original decoder with a lightweight design that incorporates skip connections. This modification enhances the retention of spatial information—especially for smaller objects like surgical tool tips, which are prone to information loss during down-sampling. We introduce a priority-weighted conditional fusion strategy to merge the outputs from both instances, ensuring that critical features are preserved in the final segmentation. We evaluate performance using mean intersection over union (IoU) and Dice scores, demonstrating the model's efficacy in both anatomical and tool segmentation tasks.

As surgical data usually suffers from class imbalance, we implemented a combined loss function (Eq. 6) that integrates Tversky loss (Eq. 4) with cross-entropy loss (Eq. 5). We applied geometric augmentations—flips, cropping, and rotations—that preserved color distribution and maintained segmentation precision.

We set $\alpha = 0.7$ and $\beta = 0.3$ to penalize false negatives, enhancing the segmentation of delicate structures such as suturing needles and instrument shafts. This configuration enabled consistent improvement in Dice and mIoU scores while avoiding overfitting.

**Equation 1:** Tversky Index formula (14).

$$\text{Tversky Index} = \frac{TP}{TP + \alpha \cdot FP + \beta \cdot FN} \tag{1}$$

**Equation 2:** Cross Entropy formula (15).

$$\text{Cross Entropy} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{2}$$

**Equation 3:** Combined loss.

$$\text{Combined\_Loss} = \alpha \cdot \text{Tversky\_Loss} + (1 - \alpha) \cdot \text{CE\_Loss} \tag{3}$$

## 3 Results and Discussion

We thoroughly assessed Surg-SegFormer on two publicly available benchmarks for robot-assisted surgery—EndoVis2017 (17) and EndoVis2018 (18) —and compared it with SOTA models. The model demonstrated notable performance on classes with subtle structures. EndoVis2017 focuses on segmenting seven instruments: Bipolar Forceps, Prograsp Forceps, Large Needle Driver, Vessel Sealer, Grasping Retractor, Monopolar Curved Scissors, and Ultrasound Probe. On the other hand, EndoVis2018 is divided into two tasks: Task 1 (Holistic scene segmentation) originally contains 12 labels spanning anatomy and instrument parts; following common practice, we merge the three fine-grained part labels—instrument shaft, wrist, and clasper—into a single Robotic Instrument Part class, yielding seven labels for per-class analysis: Background Tissue, RI, Kidney Parenchyma, Covered Kidney, Small Intestine (SI), Suturing Needle (SN), and UP. Task 2 (instrument-type segmentation) likewise comprises seven categories—BF, PF, LND, MCS, UP, Suction Instrument (SI), and Clip Applier (CA). Across both datasets and tasks, Surg-SegFormer achieved SOTA performance, with particular increase over SOTA in SN and UP classes. The per-class performance of Surg-SegFormer on EndoVis 2017 in table 2 and for EndoVis 2018 in table 3 in the appendix section.

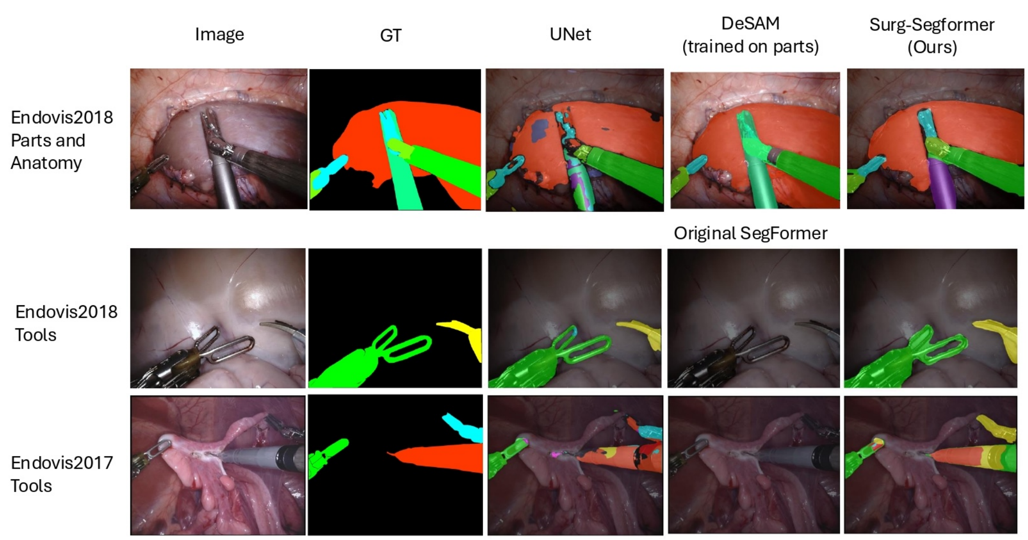


Figure 2: Models' Performance on Different Segmentation Tasks

### 3.1 Performance Analysis

Table 1 reveals that several baseline architectures excel on a single benchmark yet underperform on the other. S3Net and AdaptiveSAM, for instance, lead the instrument–only EndoVis2017 task (mIoU 0.72) but fall to 0.74 and 0.65, respectively, when anatomy and part labels are introduced in EndoVis2018. A complementary pattern appears for MATIS, which ranks near the top for EndoVis2018 instrument–type segmentation (0.77 mIoU) yet drops to 0.63 on EndoVis2017.

Table 1: Models' Overall Performance

| Model | Parts 2018 | | Type 2018 | | Type 2017 | |
|---|---|---|---|---|---|---|
| | mIoU | Dice | mIoU | Dice | mIoU | Dice |
| UNet* | 0.53 | 0.58 | 0.57 | 0.60 | 0.49 | 0.51 |
| SurgicalSAM (16) | - | - | **0.80** | - | 0.70 | - |
| TernausNet (16) | - | - | 0.40 | - | 0.13 | - |
| ISI-Net (16) | - | - | 0.71 | - | 0.52 | - |
| S3Net (16) | - | - | 0.74 | - | **0.72** | - |
| MATIS (16) | - | - | 0.77 | - | 0.63 | - |
| MedT (5) | 0.64 | 0.68 | - | - | 0.29 | 0.31 |
| AdaptiveSAM (5) | 0.65 | 0.69 | - | - | **0.72** | **0.74** |
| SAM-ZS (5) | 0.06 | 0.10 | - | - | 0.03 | 0.06 |
| SegFormer* (Single Model) | 0.57 | 0.59 | 0.46 | 0.47 | 0.41 | 0.42 |
| Surg-SegFormer* | **0.80** | **0.89** | 0.64 | **0.66** | 0.54 | 0.56 |

Parts 2018: EndoVis2018, Type 2018: EndoVis2018 dataset tools type only. Type 2017: EndoVis2017 dataset tools type only. The * means that we trained the models from our side and reported the results. The rest of the results were taken from the models' papers.

Surg-SegFormer presents a more balanced profile. Although its 0.54 mIoU on EndoVis2017 trails the prompt-tuned leaders by roughly eighteen percentage points, it remains well ahead of classical U-Net (0.49) and the re-trained SegFormer backbone (0.41). The same architecture rises to the top of EndoVis2018 Task 1 with 0.80 mIoU and 0.89 Dice, outperforming the strongest transformer-based baseline, MedT, by sixteen percentage points. Qualitative examples in Fig. 2 confirm the numerical trend: in scenes with overlapping tools and ambiguous tissues, baseline outputs either smooth away fine structures or miss entire parts, whereas Surg-SegFormer retains complete masks and sharp boundaries.

The consistency across the two datasets is attributed to three design elements: a dual-branch encoder that specialises separately in tissue and metallic cues; a priority-weighted fusion rule that reduces false negatives in crowded frames; and a hybrid Tversky–cross-entropy loss that counteracts background dominance while preserving sub-pixel detail.

# 4 CONCLUSION

In this work, we presented Surg-SegFormer, a unified and lightweight transformer-based architecture tailored for surgical scene understanding. Unlike many existing models that specialize in either anatomical or tool segmentation, Surg-SegFormer addresses both tasks simultaneously, demonstrating strong performance in multi-class and single-class surgical segmentation. Through extensive experiments on the EndoVis2017 and EndoVis2018 datasets, our model consistently outperformed classical and recent SoTA approaches, including prompt-based methods, particularly in anatomically complex or visually challenging scenes. This suggests that Surg-SegFormer can serve as a robust backbone for real-time, intraoperative surgical assistance systems, providing precise segmentation of both instruments and critical anatomy.

The high segmentation accuracy—achieved without reliance on handcrafted prompts, large models, or heavy post-processing—emphasizes the efficiency and scalability of our approach. The incorporation of a hybrid loss function (Tversky + Cross-Entropy) proved particularly effective in handling class imbalance, contributing to more stable training and better performance across underrepresented categories.

# References


[1] Andrea Moglia, Konstantinos Georgiou, Evangelos Georgiou, Richard M. Satava, and Alfred Cuschieri, "A systematic review on artificial intelligence in robot-assisted surgery, *International Journal of Surgery*, vol. 95, article 106151, 2021.

[2] Fatimaelzahraa Ali Ahmed, Mahmoud Yousef, Mariam Ali Ahmed, Hasan Omar Ali, Anns Mahboob, Hazrat Ali, Zubair Shah, Omar Aboumarzouk, Abdulla Al Ansari, and Shidin Balakrishnan, "Deep learning

for surgical instrument recognition and segmentation in robotic-assisted surgeries: a systematic review," *Artificial Intelligence Review*, vol. 58, no. 1, pp. 1, 2024, doi: 10.1007/s10462-024-10979-w.

[3] D.Kiyasseh, R.Ma, T.F. Haque, B.J.Miles, C.Wagner, D.A.Donoho, A.Anandkumar, and A.J.Hung, 'A vision transformer for decoding surgeon activity from surgical videos,' *Nature Biomedical Engineering*, vol.7, no.6, pp.780–796, June 2023.

[4] L. Sadati, S. Yazdani, and P. Heidarpoor, "Surgical residents' challenges with the acquisition of surgical skills in operating rooms: A qualitative study," *Journal of Advances in Medical Education & Professionalism*, vol. 9, no. 1, pp. 34–43, 2021, doi: 10.30476/jamp.2020.87464.1308.

[5] J. N. Paranjape, N. G. Nair, S. Sikder, S. S. Vedula, and V. M. Patel, "AdaptiveSAM: Towards Efficient Tuning of SAM for Surgical Scene Segmentation," 2023. [Online]. Available: https://arxiv.org/abs/2308.03726.

[6] C. González, L. Bravo-Sánchez, and P. Arbelaez, "ISINet: An Instance-Based Approach for Surgical Instrument Segmentation," 2020. [Online]. Available: https://arxiv.org/abs/2007.05533.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *arXiv preprint arXiv:1511.00561*, 2016. Available: https://arxiv.org/abs/1511.00561.

[8] V. Iglovikov and A. Shvets, "TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation," *arXiv preprint arXiv:1801.05746*, 2018. Available: https://arxiv.org/abs/1801.05746.

[9] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, 2021, pp. 12077–12090.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," *arXiv preprint arXiv:1802.02611*, 2018.

[13] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," *arXiv preprint arXiv:2304.02643*, 2023.

[14] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," *arXiv preprint arXiv:1706.05721*, 2017. Available: https://arxiv.org/abs/1706.05721.

[15] A. Mao, M. Mohri, and Y. Zhong, "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," *arXiv preprint arXiv:2304.07288*, 2023. Available: https://arxiv.org/abs/2304.07288.

[16] W. Yue, J. Zhang, K. Hu, Y. Xia, J. Luo, and Z. Wang, "SurgicalSAM: Efficient class promptable surgical instrument segmentation," *arXiv preprint arXiv:2308.08746*, 2023. Available: https://arxiv.org/abs/2308.08746.

[17] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt, L. Herrera, W. Li, V. Iglovikov, H. Luo, J. Yang, D. Stoyanov, L. Maier-Hein, S. Speidel, and M. Azizian, "2017 Robotic Instrument Segmentation Challenge," *arXiv preprint arXiv:1902.06426*, 2019. Available: https://arxiv.org/abs/1902.06426.

[18] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, and I. Luengo, "2018 Robotic Scene Segmentation Challenge," *arXiv preprint* arXiv:2001.11190, 2020.

## .1 Experimental Setup

In this section we present the full training pipeline, hyperparameters, and the specifications of the used GPU. In this work we two GPUs were used to evaluate the model's performance: a local NVIDIA RTX4090, 24GB and a cloud-based NVIDIA V100-32G. Larger models were trained on the cloud to reduce runtimes. The code was implemented in PyTorch, and hyperparameters were selected empirically. We used the Adam optimizer with weight decay of $10^{-4}$ and a learning rate of $5 \times 10^{-6}$, enabling the model to learn fine details while avoiding early plateaus. A cyclic learning rate scheduler and a batch size of 4 were employed to help the model escape local minima across 100 training epochs.

To address class imbalance between extensive background regions and smaller, complex instrument areas, we implemented a combined loss function (Eq. 6) that integrates Tversky loss (Eq. 4) with cross-entropy loss (Eq. 5). We applied geometric augmentations—flips, cropping, and rotations—that preserved color distribution and maintained segmentation precision.

We set $\alpha = 0.7$ and $\beta = 0.3$ to penalize false negatives, enhancing the segmentation of delicate structures such as suturing needles and instrument shafts. This configuration enabled consistent improvement in Dice and mIoU scores while avoiding overfitting.

**Equation 1:** Tversky Index formula (14).

$$\text{Tversky Index} = \frac{TP}{TP + \alpha \cdot FP + \beta \cdot FN} \tag{4}$$

**Equation 2:** Cross Entropy formula (15).

$$\text{Cross Entropy} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{5}$$

**Equation 3:** Combined loss.

$$\text{Combined\_Loss} = \alpha \cdot \text{Tversky\_Loss} + (1 - \alpha) \cdot \text{CE\_Loss} \tag{6}$$

# A Detailed Results

In this section we show more detailed results of Surg-SegFormer against SOTA models.

## A.1 EndoVis2017

On the seven-instrument *EndoVis2017* benchmark (see Table 1), Surg-SegFormer achieved an overall **mIoU of 0.54** and **Dice of 0.56**. Recent prompt-driven models such as AdaptiveSAM reported higher mIoU (0.72); Surg-SegFormer clearly outperforms the canonical U-Net (0.49/0.51), the re-trained SegFormer backbone (0.41/0.42), and the task-specific ISI-Net (0.52). Class-wise inspection underscores the model's strength on fine tools. Results in table 2 shows the model's best IoUs on three classes: *Ultrasound Probe* (**0.87**) and *Monopolar Curved Scissors* (**0.69**). Lower scores for *Bipolar Forceps* (0.24) and *Prograsp Forceps* (0.16) likely stem from limited visual diversity and inter-class ambiguity among graspers, yet overall the method delivers balanced segmentation without task-specific tuning.

Table 2: mIoU Values on EndoVis2017 - Tools Type

| Model | BF | PF | LND | VS | GR | MCS | UP |
|---|---|---|---|---|---|---|---|
| UNet* | 0.27 | 0.20 | 0.39 | 0.35 | <u>0.41</u> | <u>0.65</u> | 0.65 |
| TernausNet (16) | 0.13 | 0.12 | 0.21 | 0.06 | 0.01 | 0.01 | 0.17 |
| ISI-Net (16) | 0.39 | 0.39 | 0.50 | 0.27 | 0.02 | 0.29 | 0.13 |
| S3Net (16) | **0.75** | **0.54** | **0.62** | <u>0.36</u> | 0.27 | 0.43 | 0.28 |
| MATIS (16) | <u>0.66</u> | <u>0.51</u> | <u>0.52</u> | 0.33 | 0.16 | 0.19 | 0.24 |
| SegFormer* | 0.00 | 0.0 | 0.003 | **0.45** | **0.47** | 0.49 | **0.97** |
| **Surg-SegFormer*** | 0.24 | 0.16 | 0.47 | **0.45** | **0.47** | **0.69** | <u>0.87</u> |

BF: Bipolar Forceps, PF: Prograsp Forceps, LND: Large Needle Driver, VS: Vessel Sealer, GR: Grasping Retractor, MCS: Monopolar Curved Scissors, UP: Ultrasound Probe

For the instrument-type task Table 3 illustrates the strong performance of Surg-SegFormer, which remains among the top three in five of seven tools, leading on *Suction Instrument* (0.83) and nearly matching the best on *Clip Applier* (0.93). Lower IoUs for *Prograsp Forceps* (0.13) and *Large Needle Driver* (0.09) echo trends already seen in EndoVis2017 and can be traced to visually similar end-effectors and a paucity of examples in the training split. These fine-grained insights confirm that Surg-SegFormer's hybrid scale-aware design excels when subtle structural cues differentiate classes, while leaving room for future work on grasper-type instruments with high intra-class variance.

Table 3: mIoU Values on EndoVis2018 - Tools Type

| Model | BF | PF | LND | MCS | UP | SI | CA |
|---|---|---|---|---|---|---|---|
| UNet* | 0.64 | 0.12 | 0.12 | 0.66 | 0.54 | 0.73 | 0.8 |
| TernausNet (8) | 0.44 | 0.05 | 0.00 | 0.50 | 0.00 | 0.00 | 0.00 |
| ISI-Net (16) | 0.74 | 0.49 | 0.31 | 0.88 | 0.02 | 0.38 | 0.00 |
| S3Net (16) | 0.77 | **0.50** | 0.20 | 0.92 | 0.07 | 0.51 | 0.00 |
| MATIS (16) | **0.83** | 0.39 | **0.40** | **0.93** | 0.16 | 0.64 | 0.04 |
| SegFormer* | 0.04 | 0.05 | 0.08 | 0.20 | **0.76** | 0.80 | **0.95** |
| **Surg-SegFormer*** | 0.67 | 0.13 | 0.086 | 0.82 | 0.70 | **0.83** | 0.93 |

BF: Bipolar Forceps, PF: Prograsp Forceps, LND: Large Needle Driver, MCS: Monopolar Curved Scissors, UP: Ultrasound Probe, SI: Suction Instrument, CA: Clip Applier.

# B Ablation Study

Our ablation study was conducted to evaluate the impact of various configurations on the performance of the Surg-SegFormer model. The model went through different configurations for different aspects, such as loss functions, and the data fusion operation of the dual model output.

*Loss Functions* We chose the model's loss function through examining various single-loss-function methodologies versus composite loss methodologies. The combined loss function in our model integrates Tversky Loss and Cross-Entropy Loss, addressing the inherent class imbalance present in surgical datasets dominated by background pixels. Tversky loss excels at managing class imbalance and highlighting the boundaries of segmented objects, whereas multi-class cross-entropy is superior in achieving overall classification accuracy across multiple classes. The synergistic Tversky loss and multi-class cross-entropy use their strengths to improve training. This approach effectively penalizes false negatives, particularly for small and intricate objects like suturing needles and tool-tips, ensuring better delineation against complex surgical backgrounds.

First, we tested the single-loss functions approach—Tversky loss function and multi-class cross-entropy loss, then compared it with the combination of the two. To achieve the best balance between class imbalance and classification accuracy, the parameters $\alpha$ and $\beta$ are optimized through testing. We found that the configuration, with $\alpha = 0.7$ and $\beta = 0.3$, prioritizes the penalization of false negatives to enhance segmentation accuracy for challenging regions. As seen in Table 4, our combined strategy had the highest mIoU of 85.7% and Dice coefficient of 89.21%, outperforming the single-loss approaches, and demonstrating that combining Tversky loss with multi-class cross-entropy effectively optimizes segmentation accuracy and addresses class imbalance.

| **Loss Function** | **mIoU** | **Dice** |
|---|---|---|
| Tversky | 64.79 | 65.54 |
| Cross-entropy | 74.34 | 76.18 |
| Combined loss | **85.70** | **89.21** |

Table 4: Loss Functions Comparisons

Future refinements could explore dynamic loss weighting schemes, where weights adjust adaptively based on the proportion of each class within a frame. Additionally, focal Tversky Loss could be incorporated to down-weight well-classified regions while emphasizing hard-to-segment examples. These enhancements would further improve segmentation robustness in highly imbalanced datasets, paving the way for more accurate and generalizable models.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract describes the problem that our model is trying to solve, while the introduction presents existing solutions with their limitations. Both sections highlight the contribution and findings of our work.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In our paper we discussed our main findings and the limitation of one of the model's elements, which we elaborated on in the appendix section under "Ablation." We believe that the model can be enhanced for smoother use by integrating the tools and anatomical instance together, which is one of our current work motivations. Additionally, the existence of two instances led to the use of static priority-based operation, which can be converted to a dynamic loss weighting scheme.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not include any theoretical assumptions.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: This work has a detailed explanation of the methodology and experimental setup, which is added to the appendix. As we believe that research is about transparency to give a chance for researchers to build on our approach and improve the field instead of redundant repetition of experiments.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: At this stage we are not releasing the code, as our methodology is well-explained and enables reproducibility. However, once our final modified model is ready, all codes will be released.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, our paper mentions all model specifications and training pipeline to ensure reproducipility.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: The paper does not report any error bars, as they are not applicable.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper includes the GPU specification, which was the model trained on as part of the experimental setup.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer:[Yes]

    Justification: Our paper represents the issue that the surgical society is facing here in the country and suggests a solution, which is Surg-SegFormer. As the model showed great results, it is in the process of the clinical validation phase.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: **[TODO]**

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: **[TODO]**

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: This paper does not release new assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: No crowdsourcing was used.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA] .

Justification: No involvement of crowdsourcing.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer:[No]

    Justification: This paper does not use LLMs in either ideation or writing. loss weighting schemes,