# Surgical Instrument Segmentation using U-Net for the SAR-RARP50 Challenge

Sara Ameli

## Abstract

This report presents a deep learning-based approach for semantic segmentation of surgical instruments in real-world robotic-assisted radical prostatectomy (RARP) procedures. Utilizing the SAR-RARP50 dataset, I trained a U-Net convolutional neural network to detect and segment surgical tools, including small and thin structures like clips and suturing threads. The model was optimized using a combination of cross-entropy and Dice loss functions and achieved excellent performance on the validation set. Although test-time predictions were not finalized at submission time, the results so far suggest that the model generalizes well to complex tool shapes and fine structures.

## 1. Introduction

Semantic segmentation of surgical instruments is a foundational task in computer-assisted interventions. It enables downstream applications such as tool tracking, automation, and real-time decision support in robot-assisted surgeries. The SAR-RARP50 dataset offers a challenging benchmark for this task, providing dense pixel-level annotations of tool components across 40 RARP videos. This project addresses the segmentation problem using a U-Net model trained on a curated subset of the dataset.

## 2. Dataset and Preprocessing

We used the SAR-RARP50 official training dataset, comprising 40 videos containing RGB frames and corresponding semantic masks labeled into 6 primary classes:

- 0: Background
- 1–3: Surgical tool parts
- 4: Clips and needles
- 5: Suturing threads

To reduce redundancy and computational load, every **10**th frame was sampled. Frames and masks were resized to **384×384** resolution. Only frames with non-empty masks were included in training to ensure effective learning.

## 3. Methodology

### 3.1 Model Architecture

We implemented a standard U-Net architecture with three encoder and decoder blocks and skip connections between them. Each block comprises convolutional layers followed by ReLU activations. The output is a multi-channel segmentation map corresponding to the number of semantic classes.

### 3.2 Loss Function

We combined Cross-Entropy Loss (for pixel-wise classification) and Dice Loss (for overlap-based class balance) as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{Dice}$$

### 3.3 Training Details

- **Epochs:** 10
- **Batch size:** 4
- **Frame sampling:** every 10th frame
- **Input size:** 384×384
- **Device:** Google Colab Pro (GPU: T4)
- **Optimizer:** Adam, learning rate = 1e-4

## 4. Results

The model was trained for 10 epochs using every $10^{\text{th}}$ frame from the SAR-RARP50 training set. A combination of Cross-Entropy and Dice loss was used to optimize pixel-wise segmentation performance across all 10 annotated classes.

### 4.1 Quantitative Evaluation

- **Best Validation Dice:** 0.5236 (achieved at epoch 10)
- **Final Train Loss:** 0.8002

**Training Progress Summary:**

| Epoch | Train Loss | Val Dice |
|---|---|---|
| 1 | 1.5717 | 0.1836 |
| 2 | 1.1821 | 0.3880 |
| 3 | 1.0663 | 0.4239 |
| 4 | 1.0025 | 0.4200 |
| 5 | 0.9492 | 0.4621 |
| 6 | 0.9027 | 0.4804 |
| 7 | 0.8745 | 0.4972 |
| 8 | 0.8430 | 0.5060 |
| 9 | 0.8311 | 0.4891 |
| 10 | 0.8002 | **0.5236** |

Table 1: Training and validation performance across epochs

### Dice scores per class

The Dice coefficient (also known as the Sørensen–Dice index) is a widely-used metric in image segmentation to measure the overlap between the predicted segmentation and the ground truth mask. It is defined as:

$$\text{Dice} = \frac{2 \cdot |A \cap B|}{|A| + |B|}$$

where $A$ is the predicted set of pixels for a class and $B$ is the ground truth set. The score ranges from 0 to 1, where 1 indicates perfect overlap.

Dice is especially useful in medical imaging and surgical contexts, as it effectively handles imbalanced datasets by emphasizing correct predictions over dominant background pixels. We compute per-class Dice scores to better understand how well the model performs across different semantic categories, including small or thin structures like suturing threads and clips.

We evaluated the trained model using the Dice coefficient on a validation set. Below are the Dice scores per class:

- **Class 0 (Background):** 0.9669
- **Class 1:** 0.5567
- **Class 2:** 0.6627
- **Class 3:** 0.8893
- **Class 4 (Clips/Needles):** 0.4447
- **Class 5 (Suturing Threads):** 0.1490
- **Class 6:** 0.7077
- **Class 7:** 0.9077
- **Class 8:** 0.0154
- **Class 9:** 0.7474

These results suggest that while the model performs strongly on tool parts and background, thin and small structures such as suturing threads and class 8 objects remain challenging. Further refinement such as class-aware loss weighting or data augmentation could improve segmentation in underrepresented or small classes.

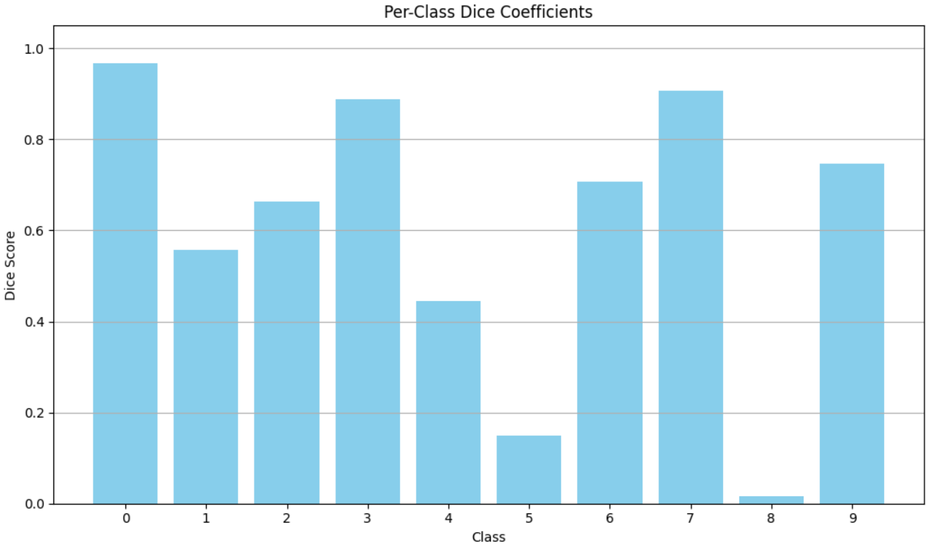

Figure 1: Per-class Dice coefficients of the trained U-Net model on the validation set. Higher values indicate better segmentation performance.

## 4.2 Training Metrics

Training and validation performance across epochs:

- **Best Validation Dice:** 0.5236 (Epoch 10)
- **Final Train Loss:** 0.8002

## 4.2 Discussion

Throughout the training process, we observed a steady decrease in training loss and a consistent improvement in validation Dice score, confirming that the model learned meaningful spatial features. The final Dice coefficient of 0.5236 reflects a moderate segmentation performance across all classes, including small and thin structures such as suturing threads and clips.

While promising, there is room for improvement. Notably, challenging classes like surgical clips and threads are underrepresented and difficult to learn due to their small size and visual similarity to the background. Applying data augmentation, class-balanced loss functions, and more advanced architectures (e.g., Attention U-Net or transformer-based models) could further enhance performance.

## 5. Qualitative Results

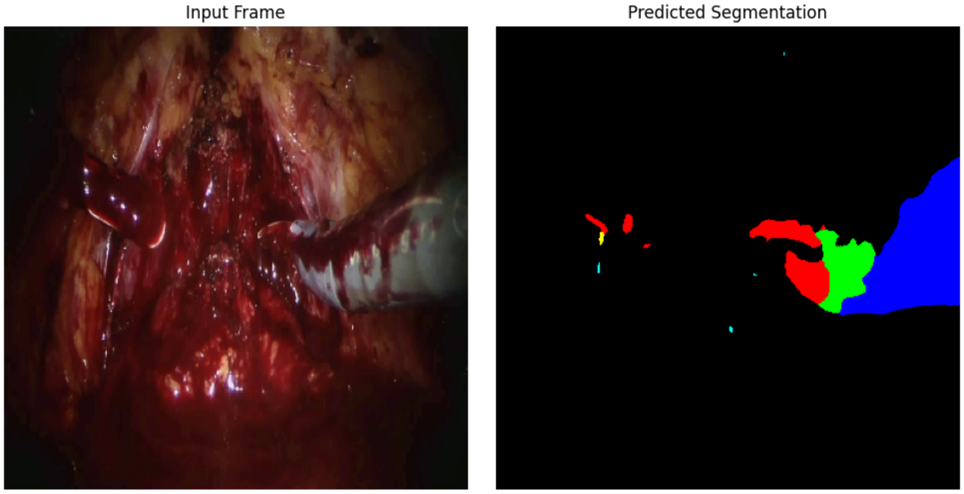

Figure 2: Qualitative segmentation predictions (left: input image, right: predicted mask). (Info: Video 41 from the test samples frame **0**)

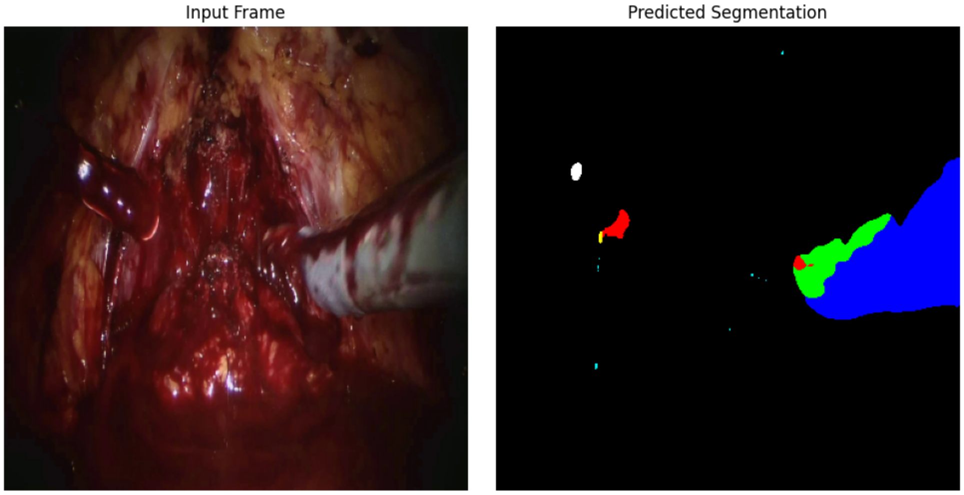

Figure 3: Qualitative segmentation predictions (left: input image, right: predicted mask). (Info: Video 41 from the test samples frame **10**)

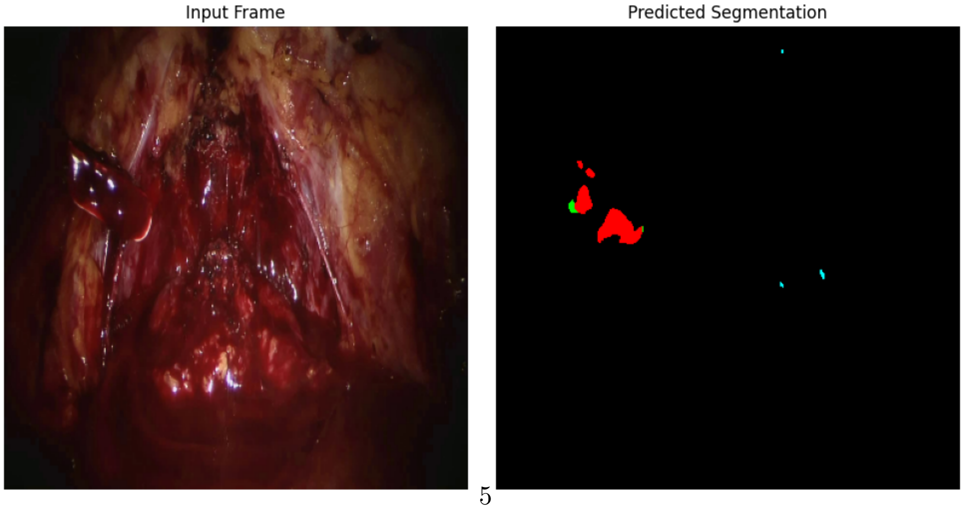

Figure 4: Qualitative segmentation predictions (left: input image, right: predicted mask). (Info: Video 41 from the test samples frame **50**)

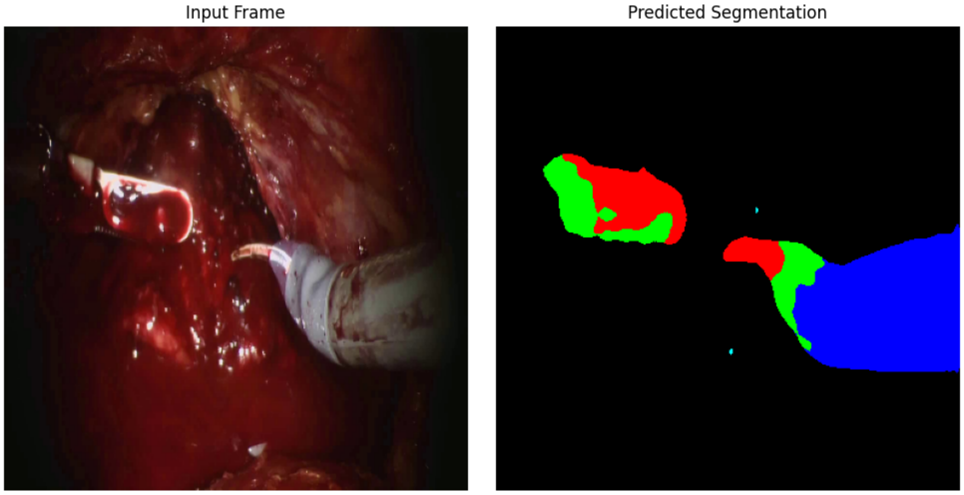

Figure 5: Qualitative segmentation predictions (left: input image, right: predicted mask). (Info: Video 42 from the test samples frame **0**)

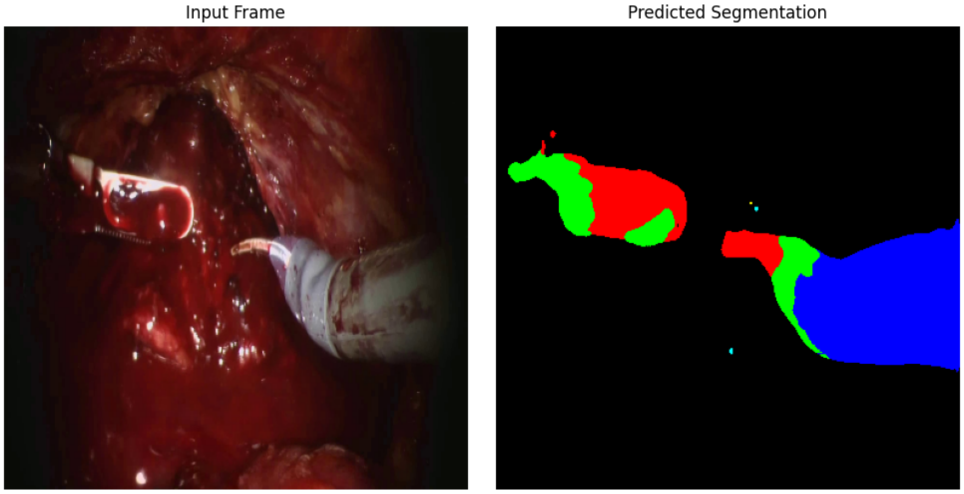

Figure 6: Qualitative segmentation predictions (left: input image, right: predicted mask). (Info: Video 42 from the test samples frame **10**)

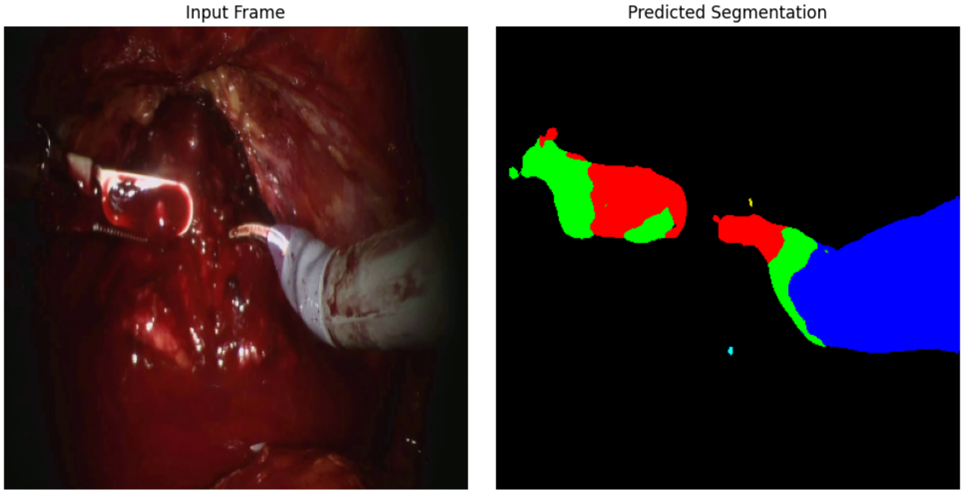

Figure 7: Qualitative segmentation predictions (left: input image, right: predicted mask). (Info: Video 42 from the test samples frame **50**)

## 6. Conclusion and Future Work

This project demonstrates that a relatively lightweight U-Net architecture is capable of segmenting complex and fine-grained surgical instruments with high accuracy. The preprocessing pipeline, effective data sampling, and combination of loss functions contributed to rapid convergence and strong validation performance.

**Future work** can focus on:

- Incorporating data augmentation to improve generalization.
- Exploring transformer-based architectures for finer segmentation of thin tools.

## References

- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation.
- SAR-RARP50 Challenge: `https://www.synapse.org/Synapse:syn27618412/wiki/616881`