EVALUATION OF DATA AUGMENTATIONS FOR SURGICAL TOOL SEGMENTATION

Shubhanjay Varma

April 20, 2025

Abstract

This study evaluates the impact of data augmentation using the Albumentations library on the task of surgical tool segmentation, a key challenge in robotic-assisted surgery analysis. Leveraging a deep learning-based semantic segmentation model, we conducted two experiments: one with standard training on unaugmented data, and another with applying the 'Horizontal Flip' augmentation. The observed results have been analyzed for discussion and future work.

1 Introduction

Surgical tool segmentation plays a critical role in the analysis of robotic-assisted surgeries, particularly in procedures like Robot-Assisted Radical Prostatectomy (RARP)[1]. Accurate segmentation of surgical tools from video data captured during such procedures is essential for tasks such as tool tracking, automated assistance, and performance analysis, which ultimately contribute to improving the precision and efficiency of robotic surgeries.

Data-driven approaches, particularly deep learning-based models, have gained significant attention in medical image analysis due to their potential to automatically learn complex features from raw data. In particular, semantic segmentation models such as UNet[2] have demonstrated great success in tasks involving pixel-level classification, where each pixel in an image is assigned a class label corresponding to specific anatomical structures or surgical tools.

However, training deep learning models, particularly in medical domains, requires large amounts of annotated data. Given the limited availability of labeled datasets, data augmentation techniques have emerged as a vital tool to artificially expand training datasets. These techniques allow models to generalize better by simulating various transformations on the existing data, thereby improving robustness to unseen conditions. One such augmentation technique is the application of random transformations like horizontal flips, which have been shown to be effective in enhancing the diversity of training samples and preventing overfitting.

This study aims to evaluate the effectiveness of data augmentation using the Albumentations library on the performance of a deep learning-based UNet model for surgical tool segmentation. Specifically, we compare the performance of the model trained on unaugmented data against one trained with the 'Horizontal Flip' augmentation. The findings from this study provide insights into the potential benefits of data augmentation in improving model generalization for surgical tool segmentation tasks.

2 Methodology

2.1 Dataset

The dataset is the SAR-RARP50 dataset[1]. The data comprises of 50 real-world RARP operations which have been recorded with an endoscope and the data from the left endoscopic channel has been stored for each. Along with the video, labelled data in the form of png files of the shape (1080, 1920) is provided. Each 'position' of the labelled data corresponds to the 60th frame from the RARP video data which is available.
More information is present in the dataset, but it is not relevant for performing the surgical tool segmentation task. Each value of the grayscale png files is a value between 0 and 9, inclusive, with 0 corresponding to the 'background' class, and 1 corresponding to the 'tool clasper'. Information regarding the other classes can be found in the SAR-RARP50 Dataset's readme file.

2.2 Training, Validation, and Test Datasets

The first 40 videos have been provided as the training dataset, and the last 10 videos (video 41 to video 50) have been provided as the test dataset. The training dataset was further divided into the first 32 videos (operations) as the actual training dataset, and the next 8 videos as the validation dataset for testing our model architecture for further improvements before a final check on the testing dataset.
While splitting the training dataset, random frames (and their corresponding masks or labelled data) were not chosen across different videos as that could result in frames from the same operation being present in both the training dataset, and the validation dataset, leading the model to learn the styles and patterns of that particular surgery, causing artificially high validation accuracy, while demonstrating poor generalization on new, unseen surgeries.

2.3 Model Architecture

For the surgical tool segmentation task, we employed a **UNet** architecture implemented using the `segmentation_models.pytorch` (SMP) library. UNet is a popular encoder-decoder architecture widely used in biomedical image segmentation due to its ability to capture both global context and fine-grained spatial details.

We used `ResNet-34`[3] as the encoder backbone, pre-trained on the ImageNet dataset. This encoder extracts multi-scale hierarchical features from the input RGB frames, which are then passed through a symmetric decoder that progressively upsamples the feature maps to produce dense pixel-wise predictions. Skip connections between corresponding encoder and decoder layers preserve spatial information, improving segmentation accuracy, especially for small or fine structures.

The model was configured with:

- `in_channels` = 3 (to match the RGB input),
- `classes` = 10 (to segment 10 distinct semantic categories in the surgical tool dataset).

For training, we used the Cross Entropy Loss function, which is suitable for multi-class segmentation problems. The model was optimized using the Adam optimizer with a learning rate of 1×10^{-4}. Training was conducted for a single epoch due to time and resource constraints, and model performance was evaluated using the mean Intersection over Union (mIoU) metric on the validation set.

2.4 Metrics and Evaluation

To assess the performance of the segmentation model, we conducted evaluations exclusively on the validation set after each training epoch. Two key metrics were used:

- **Cross Entropy Loss**: This loss function was used during training and validation to quantify the discrepancy between the predicted segmentation map and the ground truth labels. It is well-suited for multi-class pixel-wise classification tasks and penalizes incorrect class predictions at the pixel level.
- **Mean Intersection over Union (mIoU)**: This metric is widely adopted in semantic segmentation tasks. For each class, the IoU is computed as the ratio of the intersection to the union of the predicted and ground truth regions. The mIoU is then calculated as the mean across all segmentation classes. It provides an interpretable and class-sensitive measure of segmentation quality.

All reported values in this study — including comparisons between the augmented and unaugmented training conditions — were obtained from the validation set. A separate test set was not used in this experiment, and therefore results should be interpreted as preliminary. Nonetheless, these validation metrics offer meaningful insights into the impact of data augmentation on model generalization during training.

3 Results

Table 1 presents the performance of the UNet model on the validation set under two training conditions: with and without data augmentation using horizontal flipping.

Training Condition	**Train Loss**	**Validation Loss**	**Validation mIoU**
No Augmentation	0.1591	0.1323	0.4663
Horizontal Flip Augmentation	0.6507	0.1967	0.3985

Table 1: Performance comparison of models trained with and without data augmentation on the validation set.

The model trained without data augmentation achieved a higher mean Intersection over Union (mIoU) of 0.4663 and lower validation loss compared to the model trained with horizontal flip augmentation. Additionally, the model without augmentation exhibited significantly lower training loss, suggesting faster convergence.

4 Discussion

4.1 Impact of Data Augmentation

The results indicate that the model trained without augmentation outperformed the model trained with horizontal flip augmentation in terms of both validation loss and mIoU. At first glance, this may suggest that augmentation had a detrimental effect on model performance. However, this difference is better interpreted in light of the experimental constraints and the dynamics of learning under data augmentation.
The higher validation loss and lower mIoU for the augmented model may be attributed to the introduction of greater input variability through flipping. Data augmentation, by design, exposes the model to more diverse transformations of the training data, which can initially increase the difficulty of learning. Especially in early training stages, such as the single epoch used in this study, augmented inputs may hinder convergence as the model has not yet seen enough examples to generalize robustly. Moreover, horizontal flips may not introduce sufficiently novel or informative variations in a surgical context. Given the structured and often symmetrical nature of laparoscopic video frames, especially in prostatectomy procedures, such augmentations might yield marginal gains compared to other transformations. There is also the possibility that certain flipped inputs misalign spatial priors implicitly learned from the unflipped images, leading to confusion in class boundaries during training.

Nonetheless, data augmentation is a well-established technique for improving generalization, and its benefits often become more apparent over multiple epochs[4]. The observed higher training loss and validation loss under augmentation is consistent with findings in other domains, where augmented models initially converge more slowly but yield improved robustness with extended training.

4.2 Model Performance and Limitations

The UNet architecture with a ResNet-34 encoder demonstrated promising performance even with minimal training. The relatively high mIoU of 0.4663 achieved without augmentation underscores the capacity of the model to learn meaningful spatial representations in a single pass through the training data.
However, the use of Cross Entropy Loss, while standard, may not be optimal for segmentation tasks where class imbalance is prevalent — a likely scenario in surgical tool segmentation where background pixels dominate[5]. Additionally, training for just one epoch limits the ability to draw strong conclusions about long-term trends in generalization or overfitting. With more extensive training, both models — particularly the one with augmentation — may show different performance characteristics. The evaluation was also constrained to the validation set, meaning the generalization of these models to the test data remains unverified. The absence of normalization and other preprocessing steps may also affect model robustness, particularly in varying lighting conditions or across different patients and surgical setups.

4.3 Future Work

Future iterations of this experiment should include training over multiple epochs to allow the effects of augmentation to properly manifest. It would also be beneficial to experiment with a broader set of augmentation strategies, including brightness changes, elastic deformations, and random cropping, which may introduce more relevant variability for surgical tool segmentation tasks.
In addition, incorporating advanced loss functions such as Dice Loss or Focal Loss may address the potential class imbalance problem and improve segmentation quality, especially for underrepresented tool classes such as suturing threads or clips. Evaluating per-class mIoU could provide further insights into which tools are better segmented and where the model struggles. Lastly, the inclusion of a separate test set for final evaluation, as well as qualitative visualizations of the predicted segmentation masks, will be crucial for a more comprehensive assessment of model performance and to guide future improvements.

References

[1] Dimitrios Psychogyios et al. "SAR-RARP50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge". In: *arXiv preprint arXiv:2401.00496* (2023).

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. Vol. 9351. Lecture Notes in Computer Science. Munich, Germany: Springer, 2015, pp. 234–241.

[3] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[4] Prabhjot Kaur, Balwinder Singh Khehra, and Emanual B.S. Mavi. "Data augmentation for object detection: A review". In: *2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE. 2021, pp. 537–543. DOI: 10.1109/MWSCAS47672.2021.9531802.

[5] Michael Yeung et al. "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation". In: *Computerized Medical Imaging and Graphics* 95 (2022), p. 102026. DOI: 10.1016/j.compmedimag.2021.102026.