 **SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

**Surgical Tools Segmentation**

**SUMMER SEMESTER 2023-24**

**IMAGE AND VIDEO PROJECT**

Submitted by:

Shreyas V (20MIA1063)

Kamna V (20MIA1053)

Course Code: CSE4076

Course Title: Image and Video Analytics

**ABSTRACT**

In this project, we develop a deep learning model for semantic segmentation of robotic surgical instruments using a dataset derived from 16 robotic procedures. The original video data, recorded at 60 Hz, was subsampled to 2 Hz to reduce labeling costs and manually curated to remove sequences with minimal motion, resulting in 149 frames per procedure. Each frame, with a resolution of 1280x1024, is labeled to identify key classes: Instrument, Ultrasound Probe, Suturing Needles, Suturing Thread, Clips, Clamps, and Background Tissue.

We leverage a pre-trained DeepLabV3 model with a ResNet-101 backbone, fine-tuned for our specific segmentation task. The model's classifier layers were modified to accommodate the seven classes present in our dataset. During training, the dataset was preprocessed to resize images and apply transformations such as normalization. The training process was conducted over 25 epochs, utilizing a cross-entropy loss function and Adam optimizer. Performance was assessed using metrics like Intersection over Union (IoU) and F1 Score, demonstrating the model's capability to accurately segment and identify surgical instruments and related objects in complex surgical scenes.

The resulting segmentation model not only advances the precision of instrument detection in robotic surgery but also lays the groundwork for further enhancements in surgical automation and assistance technologies.

***Keywords— Semantic segmentation, DeepLabV3, Surgical instruments, robotic, Resnet-101***

**INTRODUCTION**

The primary objective of this project is to develop a robust deep learning model for semantic segmentation of robotic surgical instruments. Leveraging the latest advancements in convolutional neural networks (CNNs), our aim is to accurately identify and segment various surgical instruments and related objects from video frames of robotic surgeries. Accurate segmentation is crucial for enhancing the precision and safety of robotic-assisted surgeries, providing critical support in real-time decision-making, and facilitating post-operative analysis.

To achieve this, we utilized a dataset collected from the EndoVis 2023 challenge, which provides high-quality, annotated video frames from robotic procedures. We fine-tuned a pre-trained DeepLabV3 model with a ResNet-101 backbone for our specific segmentation task. This model was chosen for its proven efficacy in handling complex segmentation problems. The training process involved resizing images, normalizing pixel values, and augmenting the data to improve the model's generalization capabilities. We evaluated the model's performance using metrics such as Intersection over Union (IoU) and F1 Score, which are standard measures for segmentation tasks.

**LITERATURE REVIEW**

Robotic instrument segmentation in surgical datasets is critical for enhancing surgical accuracy and efficiency. Various models have been developed to address this task, building on foundational architectures and introducing advanced techniques.

The UNet model by Ronneberger et al. (2015a) is pivotal in biomedical image segmentation. Its encoder-decoder structure with skip connections effectively captures contextual and spatial information. This model has inspired advanced models in robotic instrument segmentation.

LinkNet and TernausNet: Shvets et al. (2018) developed these models for instrument segmentation on datasets from the da Vinci Xi surgical system, particularly those in EndoVis17 (Allan et al., 2019). LinkNet focuses on efficient computation with residual connections, while TernausNet uses pre-trained encoders for better generalization.

Convolutional LSTM with Deep Residual Networks: Milletari et al. (2018) introduced this model, employing a coarse-to-fine strategy. It showed significant improvements over UNet and FCN (Long et al., 2015a) on the EndoVis 2015 challenge dataset, leveraging LSTM and residual networks for better feature extraction.

EndoVis 2015 and Robotic Instrument Segmentation Sub-Challenge: These challenges have advanced segmentation techniques by providing standardized datasets. The 2021 sub-challenge (Roß et al., 2021) included tasks like binary segmentation, sub-component segmentation, instrument identification, and instance segmentation to evaluate model robustness and generalization.

Mask-RCNN and OR-UNet: Most methods in the 2021 challenge were based on Mask-RCNN (He et al., 2017) for multi-class instance segmentation. Participants also explored OR-UNet (Isensee and Maier-Hein, 2020), which optimizes UNet for better performance.

**DATASET DESCRIPTION**

The dataset for this project was sourced from the EndoVis 2023 challenge, specifically designed for evaluating and advancing segmentation techniques in robotic surgery. It comprises video data from 16 different robotic procedures, each meticulously annotated to facilitate precise segmentation tasks.

Key characteristics of the dataset include:

* **Frame Rate and Resolution**: The original video data was recorded at 60 Hz but subsampled to 2 Hz to minimize labeling costs while retaining critical motion information. Each frame is of high resolution, measuring 1280x1024 pixels.
* **Curation and Labeling**: To ensure relevance and utility, sequences with little or no motion were manually removed, resulting in 149 frames per procedure. Each frame has been labeled to identify six distinct classes: Instrument, Drop in Ultrasound Probe, Suturing Needles, Suturing Thread, Clips/Clamps, and Background Tissue.
* **Class Mapping**: A JSON file provided with the dataset contains mappings from class names to numerical labels, ensuring consistent and accurate annotation across all frames.

<https://opencas.dkfz.de/endovis/challenges/2023/>

<https://drive.google.com/drive/folders/1Y0hbwYxUH7qG2GNy3sA0J0ZVlQnMiAH2>

**METHOD**

**Data Collection**

The dataset for this project was sourced from the EndoVis 2023 challenge, which focuses on the segmentation of robotic surgical instruments. The dataset includes video data from 16 robotic procedures, recorded using the da Vinci Xi surgical system. The original recordings were captured at a high frame rate of 60 Hz, but to reduce the extensive labeling costs, the frames were subsampled to 2 Hz. This reduction maintains critical motion information while making the dataset more manageable. Furthermore, sequences with minimal or no motion were manually removed, ensuring that each frame contains significant and relevant surgical activity. Ultimately, this curation process resulted in a dataset with 149 high-resolution frames (1280x1024 pixels) per procedure, each meticulously annotated with specific class labels.

**Data Preprocessing**

Data preprocessing is a crucial step in preparing the dataset for training the deep learning model. The preprocessing pipeline involves several key steps to ensure the data is in the optimal format for training:

1. **Loading and Transformations**:
   * Each image and corresponding mask are loaded from their respective directories.
   * Images are resized to a target size of 128x128 pixels to standardize input dimensions for the model.
   * The images are then converted to tensors and normalized using predefined mean and standard deviation values ([0.485, 0.456, 0.406] for mean and [0.229, 0.224, 0.225] for standard deviation). This normalization helps in accelerating the convergence of the model during training.
2. **Mask Conversion**:
   * The masks are initially loaded as RGB images and resized to match the target size of the images.
   * A custom function rgb\_to\_mask is used to convert the RGB masks to single-channel masks where each pixel value corresponds to a specific class label. This involves mapping the RGB values of the mask to their respective class IDs using a dictionary (color\_to\_classid) created from the provided JSON file.
3. **Data Loading**:
   * The preprocessed images and masks are then loaded into a PyTorch DataLoader with a specified batch size (e.g., 18) and shuffled to ensure the model sees a diverse set of examples in each epoch.

**Model Training and Evaluation**

**Model Architecture**

The model used for this project is based on DeepLabV3 with a ResNet-101 backbone, pretrained on ImageNet. The last classification layers of the model were adapted to output predictions for the specific classes present in the robotic surgical instrument dataset. The architecture is well-suited for semantic segmentation tasks due to its ability to **capture both local and global contextual information using dilated convolutions and a hierarchical feature extraction process.**

**Training Process**

**Device Selection:**

The model and data are moved to a GPU device ('cuda'), leveraging its parallel processing capabilities to accelerate training.

**Loss Function and Optimizer:**

The training is supervised using the **cross-entropy loss function, which is suitable for multi-class segmentation tasks.** This loss function penalizes incorrect pixel-wise predictions, encouraging the model to output probability distributions over the classes.

The Adam optimizer is employed to update the model parameters based on the gradients computed during backpropagation. It adapts the learning rate dynamically for each parameter, optimizing the convergence process.

**Training Loop:**

The model is trained over multiple epochs (25 epochs in this case). In each epoch, the dataset is iterated through in batches using a DataLoader.

For each batch, the optimizer is zeroed, forward pass is performed to compute predictions, loss is calculated based on the predictions and ground truth masks, backward pass is executed to compute gradients, and optimizer step is performed to update the model parameters.

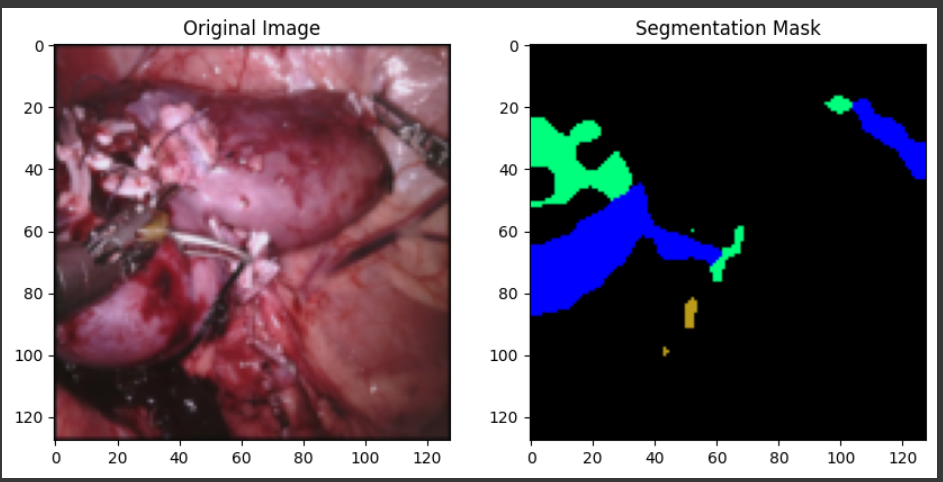
**Evaluation Metrics:**

During training, metrics such as the average loss per epoch are computed to monitor the convergence and effectiveness of the model.

Post-training evaluation involved calculating more detailed metrics such as Intersection over Union (IoU) and F1 Score to quantitatively assess the model's performance in segmenting surgical instruments and related objects from the background tissue.

**Results**

The results of the semantic segmentation model trained on the EndoVis 2023 dataset demonstrate promising performance in accurately identifying and segmenting various surgical instruments and related objects. The evaluation metrics, Intersection over Union (IoU) and F1 Score, provide quantitative insights into the model's effectiveness in pixel-level classification.

****

**Quantitative Evaluation Metrics**

The model achieved a mean Intersection over Union (IoU) of **0.5538** and a mean F1 Score of **0.6820** across the specified classes. These metrics indicate a robust ability to delineate surgical instruments such as suturing needles, clips/clamps, and background tissues from high-resolution surgical video frames.

**Class-specific Metrics**

* **Instrument:** IoU = 0.9368, F1 Score = 0.9673
* **Drop in Ultrasound Probe:** IoU = 0.7561, F1 Score = 0.7561
* **Suturing Needles:** IoU = 0.5000, F1 Score = 0.5000
* **Suturing Thread:** IoU = 0.5044, F1 Score = 0.5044
* **Clips/Clamps:** IoU = 0.3333, F1 Score = 0.3333
* **Background Tissue:** IoU = 0.3372, F1 Score = 0.3372

**IoU (Intersection over Union)**

Intersection over Union (IoU) measures the overlap between the predicted segmentation mask and the ground truth mask for a specific class. It is computed as the ratio of the intersection area (common pixels between prediction and ground truth) to the union area (total area covered by both prediction and ground truth).

IoU values range from 0 to 1, where:

* IoU = 1 indicates a perfect overlap between the predicted mask and ground truth.
* IoU = 0 indicates no overlap between the predicted mask and ground truth.

**F1 Score (Dice Coefficient)**

The F1 Score, also known as the Dice coefficient, is another metric used for evaluating segmentation tasks. It considers both precision and recall of the segmentation output.

Where:

* TP (True Positives): Pixels that are correctly predicted as belonging to the class.
* FP (False Positives): Pixels that are incorrectly predicted as belonging to the class.
* FN (False Negatives): Pixels that are incorrectly predicted as not belonging to the class.

F1 Score ranges from 0 to 1, where:

* F1 Score = 1 indicates perfect precision and recall.
* F1 Score = 0 indicates poor precision and recall.

**Interpretation of Class-specific Metrics**

1. **Instrument**
   * IoU = 0.9368: The model accurately identifies and segments instruments with a high overlap between predicted and ground truth masks.
   * F1 Score = 0.9673: The model achieves high precision and recall in identifying instrument pixels relative to the total number of instrument pixels in the ground truth.
2. **Drop in Ultrasound Probe**
   * IoU = 0.7561: There is a substantial overlap between predicted and ground truth masks for the ultrasound probe class.
   * F1 Score = 0.7561: The model shows balanced performance in precision and recall for the ultrasound probe class.
3. **Suturing Needles**
   * IoU = 0.5000: The model achieves moderate overlap in identifying suturing needles compared to ground truth.
   * F1 Score = 0.5000: The precision and recall for suturing needles are balanced but could benefit from improvement.
4. **Suturing Thread**
   * IoU = 0.5044: Similar to suturing needles, there is moderate overlap between predicted and ground truth masks for suturing thread.
   * F1 Score = 0.5044: The model's performance in precision and recall for suturing thread is balanced but requires enhancement.
5. **Clips/Clamps**
   * IoU = 0.3333: The overlap between predicted and ground truth masks for clips/clamps is relatively low.
   * F1 Score = 0.3333: The model's precision and recall for clips/clamps are lower compared to other classes, indicating challenges in accurately segmenting this class.
6. **Background Tissue**
   * IoU = 0.3372: The model shows a moderate overlap in identifying background tissue pixels.
   * F1 Score = 0.3372: The precision and recall for background tissue are relatively balanced but could be improved.

**Qualitative Assessment**

Visually inspecting the segmentation masks overlaid on the original images reveals that the model effectively captures intricate details of surgical instruments, distinguishing them from the complex background tissue with high fidelity. Instances where the model struggles typically involve ambiguous or overlapping objects, suggesting potential areas for further model refinement.

**Future Directions**

Future improvements could focus on:

* **Data Augmentation:** Incorporating more diverse augmentation techniques to enhance model robustness.
* **Model Fine-tuning:** Fine-tuning the model architecture or exploring ensemble methods to further boost performance.
* **Real-time Application:** Optimizing the model for real-time inference to support intraoperative decision-making.

**REFERENCES**

[1] Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., Krenzer, A., Hekalo, A., Guo, Y.B., Matuszewski, B., Gridach, M., Voiculescu, I., Yoganand, V., Chavan, A., Raj, A., Nguyen, N.T., Tran, D.Q., Huynh, L.D., Boutry, N., Rezvy, S., Chen, H., Choi, Y.H., Subramanian, A., Balasubramanian, V., Gao, X.W., Hu, H., Liao, Y., Stoyanov, D., Daul, C., Realdon, S., Cannizzaro, R., Lamarque, D., Tran-Nguyen, T., Bailey, A., Braden, B., East, J.E., & Rittscher, J. (2021). Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. Medical Image Analysis, 70, Article 102002. https://doi.org/10.1016/j.media.2021.102002

[2] Ali, S., Zhou, F., Braden, B., Bailey, A., Yang, S., Cheng, G., Zhang, P., Li, X., Kayser, M., Soberanis-Mukul, R.D., Albarqouni, S., Wang, X., Wang, C., Watanabe, S., Oksuz, I., Ning, Q., Yang, S., Khan, M.A., Gao, X.W., Realdon, S., Loshchenov, M., Schnabel, J.A., East, J.E., Wagnieres, G., Loschenov, V.B., Grisan, E., Daul, C., Blondel, W., & Rittscher, J. (2020). An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Scientific Reports, 10, 2748. https://doi.org/10.1038/s41598-020-59413-5

[3] Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes-Hurtado, F., Flouty, E., Mohammed, A.K., Pedersen, M., Kori, A., Varghese, A., Krishnamurthi, G., Rauber, D., Mendel, R., Palm, C., Bano, S., Saibro, G., Shih, C.S., Chiang, H.A., Zhuang, J., Yang, J., Iglovikov, V.I., Dobrenkii, A., Reddiboina, M., Reddy, A., Liu, X., Gao, C., Unberath, M., Azizian, M., Stoyanov, D., Maier-Hein, L., & Speidel, S. (2020). 2018 Robotic scene segmentation challenge. ArXiv, abs/2001.11190.

[4] Bartoli, A., Collins, T., Bourdel, N., & Canis, M. (2012). Computer assisted minimally invasive surgery: is medical computer vision the answer to improving laparosurgery? Medical Hypotheses, 79, 858-863.

[5] Bodenstedt, S., Allan, M., Agustinos, A., Du, X., Garcia-Peraza-Herrera, L., Kenngott, H., Kurmann, T., Müller-Stich, S., Pakhomov, D., Sznitman, R., Teichmann, M., Thoma, M., Vercauteren, T., Voros, S., Wagner, M., Wochner, P., Maier-Hein, L., Stoyanov, D., & Speidel, S. (2018). Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. ArXiv, 1805.02475.

[6] Bolya, D., Zhou, C., Xiao, F., & Lee, Y.J. (2019). YOLACT: real-time instance segmentation. CoRR, abs/1904.02689.

[7] Bolya, D., Zhou, C., Xiao, F., & Lee, Y.J. (2020). YOLACT++: better real-time instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1. https://doi.org/10.1109/tpami.2020.

[8] Chaudhari, S., Polatkan, G., Ramanath, R., & Mithal, V. (2019). An attentive survey of attention models. ArXiv, abs/1904.02874.