RibRib Segmentation in Surgical Images for Video-Assisted Thoracoscopic Surgery

Wiley Tam, Jean-Louis Dillenseger, Paul Babyn, Javad Alirezaie

RibRib Segmentation in Surgical Images for Video-Assisted Thoracoscopic Surgery

Wiley Tam
Department of Elect Comp & Biomed Eng
Toronto Metropolitan University
Toronto, Canada
wiley.tam@torontomu.ca

Jean-Louis Dillenseger
Laboratoire Traitement du Signal et Image
INSERM, U1099; Université de Rennes
Rennes, France
jean-louis.dillenseger@univ-rennes.fr

Paul Babyn
Department of Medical Imaging
University of Saskatchewan
Saskatoon, Canada
paul.babyn@gmail.com

Javad Alirezaie
Department of Elect Comp & Biomed Eng
Toronto Metropolitan University
Toronto, Canada
javad@torontomu.ca

***Abstract*—Lung cancer is the leading cause of cancer deaths worldwide. A potential early indicator of lung cancer is the presence of lung nodules that can be detected through screening. Open thoracotomy, a surgical approach for nodule resection, carries inherent risk which can be minimized with use of Video-Assisted Thoracoscopic Surgery (VATS); a minimally invasive alternative, reducing risks and recovery time. Precise nodule localization is crucial for efficient navigation during VATS. The utilization of intraoperative Cone-Beam Computed Tomography (CBCT), an imaging modality, can improve localization of the nodules. However, this poses a challenge when attempting to accurately align the nodule position from the CBCT to the surgical view. To address this, we propose a novel approach that segments corresponding features visible in both modalities, specifically the rib cages and Alexis O Wound Protector/Retractor (Alexis). The segmentation of these features is performed using YOLOv8 allowing image registration and alignment of the CBCT data with the surgical view. With the established correspondence, we can gauge possible camera locations and create an augmented reality overlay of the surgical site to provide real-time guidance in VATS.**

I. Introduction

According to the American Cancer Society, lung cancer is the leading cause of cancer deaths in the United States (US) [1]. Based on their predictions, the US is anticipated to see over 230,000 new lung cancer diagnoses in 2024, with a projected death toll exceeding 125,000. One early indicator of lung cancer is the presence of lung nodules. Lung nodule are small, abnormal growths typically identified with lung Computed Tomography (CT). Routine lung CT screening is highly recommended for high risk individuals as earlier detection and treatment of lung cancer can lead to improved outcomes. Surgical interventions such as nodule resection procedures may be necessary. Video-Assisted Thoracoscopic Surgery (VATS) offers a more minimally invasive approach to traditional open thoracotomy for nodule resection reducing potential risks and recovery time. However, the intraoperative localization of pulmonary lung nodules during VATS can be challenging. There are two main approaches to nodule localization techniques: preoperative, performed before surgery, and intraoperative, conducted during surgery. An example of preoperative localization is the use of a hook wire placed during pre-operative CT. While feasible, there are potential complications including pneumothorax and localization failure. Hence, intraoperative approaches have been proposed utilizing a Cone-Beam Computed Tomography (CBCT) generated by a C-Arm machine during surgery. Benefits include the avoidance of preoperative procedures leading to reduced radiation exposure and time efficiency.

The localization of the lung nodule in the surgical camera view during VATS is problematic. Despite being clearly visible in both the preoperative CT scans and intraoperative CBCT scans, lung deformations can obscure the nodule during surgery, making it difficult to visualize with the camera. Lung deformations may arise from changes in patient position from supine to lateral decubitus and the change in lung density due to the induced lung deflation (pneumothorax). Since the nodule can be difficult to locate, several registration methods have been tested to map the nodule location from the CT scan to the CBCT scan and ultimately to the operative field for surgeons. In order to achieve this final registration between CBCT and endoscopic view, this paper uses common features shared by both modalities: the rib cages and Alexis O Wound-Retractors (Alexis). By segmenting the ribs and Alexis in the camera view, we can identify the surgical field of view in the CBCT view by translating the features relative positions and possibly determining the camera location in the CBCT.

Several prior efforts have been made into segmenting anatomical regions of interests in surgical videos, but it still remains challenging. While traditional methods such as thresholding and colour segmentation are viable with individual frames, they struggle with the dynamic nature of surgical videos. The unpredictable changes in environment and image quality during surgery limits their generalizability. Thus, we focus on deep learning approaches that utilize Convolutional Neural Networks (CNNs). CNNs excel at automatically

identifying and analyzing image features, enabling them to effectively separate foreground objects from the background. Kadomatsu et al. [2] developed a novel system that applies AI to identify pulmonary air leak sites in thoracic surgeries. Specifi cally, they perform intraopertaive leak site detection using YOLO to identify these complications. Their model was trained using still images of deflated lung tissue after pulmonary resection obtained from a robotic or thoracoscopic camera. A surgeon would identify and label the true leak site to use as the ground truth images. Bilodeau et al. [3] presented a graph-based segmentation method using multiple criteria in successive stages to segment thoracoscopic images acquired during a diskectomy procedure used for thoracoscopic anterior release and fusion for scoliosis treatment. Before applying their coarse graph-based segmentation, they performed pre-processing such as Gaussian smoothing, brightness and contrast enhancement, and histogram thresholding to enhance discriminating features. Lastly, post-processing is used to remove regions considered as spurious areas. Then regions are merged in a multistage graph-based process with higher-level criteria such as grey-level similarity, region size, and common edge length to generate a segmented region map. Bamba et al. [4] leverages a CNN to perform object and anatomical feature recognition in abdominal endoscopic images from surgical videos. Their focus was to detect the GI tract, blood, vessels, uterus, forceps, ports, gauze, and clips in the images. Their model was trained and implemented through IBM Visual Insights which contained multiple open-source deep learning frame- works including GoogLeNet, Faster R-CNN, and YOLOv3. Ivantsites et al. [5] applied deep learning to automatically segment relevant anatomical structures and instruments in endoscopic images for mitral valve repairs. They tested and cross-validated the performance of three deep learning architectures: U-Net, DeepLabV3, and Obelisk-Net. Overall, the DeepLab model achieved superior results with respect to all the evaluation metrics they used. Brizuela et al. [6] performed gauze detection and segmentation in laparoscopic video images using two CNN-based models: YOLOv3 and U-Net. They created a segmentation dataset by hand-labelling 4003 frames from laparoscopic videos to efficiently train the CNN models through supervised learning. Naturally, the U-Net baseline using the MobileNetV2 architecture resulted in ideal results with a good compromise between inference speed and prediction quality. Tanzi et al. [7] proposed a deep learning and augmented reality based solution for an in-vivo robot-assisted radical prostatectomy. By using an ensemble consisting of the MobileNet as the base network and U-Net as the segmentation network, they were able to effectively segment the catheter. Scheikl et al. [8] tested numerous combinations of neural networks, loss functions, and trainability in performing semantic segmentation of organs and tissues in laparoscopic surgery. Ultimately, they executed the task by using the TernausNet-11 trained on soft-Jaccard loss with a pre-trained, trainable encoder.

The most notable work that is similar to the proposed approach is the work presented by Noblet et al. [9]. They proposed to register 2D monocular endoscopic views into the 3D CBCT space to create an augmented endoscopy guiding system. Their approach was to segment both ribs and Alexis in both imaging modalities, then perform registration using an image-to-cloud Iterative Closest Point variant. They performed rib and Alexis segmentation by using classical image processing methods and morphological operations. Specifically, in 3D segmentation, CBCT images are smoothed and operations are applied to remove noise using a Gaussian kernel. The ribs are then segmented using thresholding and closing morphological operations, focusing on the part of the rib that is close to the parietal pleurae. For 2D segmentation, the main difficulty encountered during the feature segmentation process was the disappearance of the rib outline under intercostal muscle and adipose tissue (body fat) covering parts of the ribs. To overcome this, manual segmentation of the ribs and Alexis was performed using CVAT, an online open data annotation platform. Overall, the general pattern seen in the reviewed approaches is the utilization of manual annotations to generate a training dataset with labelled data. Our proposed idea is to utilize the YOLOv8-seg model to perform automatic instance segmentation on surgical videos. Here, the YOLOv8-seg model will be custom trained using manually annotated images highlighting the individual ribs and Alexis. Ultimately, this paper presents:

Ultimately, this paper presents:

- Creation of a dataset of ribs and Alexis in surgical videos by manually annotating the targets
- Automatically segment the ribs and Alexis using the curated dataset and YOLOv8-seg
- Determine the possibility of registration and camera localization in the CBCT based on chosen results

II. Methodology

A. Data Collection

Sample surgical videos of a thoracoscopy were provided by surgeons at the University of Rennes hospital situated in France. Left lung images were extracted from these videos frame by frame. Manual filtration was performed by removing images that showed no target objects, immense occlusion of the targets, or areas outside the surgical field. After filtering, the dataset consisted of 222 images. Each image was manually labeled using LabelMe, a Python-based image annotation tool. The images were labeled with eight classes: Rib2 through Rib8, and Alexis. In a thoracoscopy, Chang et al. [10] recommends entering the chest cavity through the fourth or fifth intercostal space, after the division of intercostal muscles above the rib to preserve the neurovascular bundle. Based on this principle, we estimated that the ribs surrounding a particular Alexis were likely to be Rib4 or Rib5. Resultant annotated sample images are shown in Fig. 1 and labels are shown in Fig. 2.

Given the nature of the task, the image quality and information can be impacted significantly by subtle changes in camera movement, focus, and orientation. This often results in poor

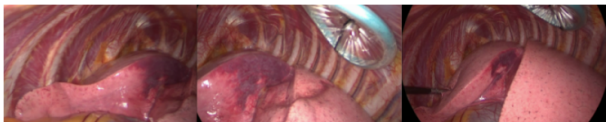

Fig. 1. Sample images of the left lung taken from the thoracoscopic camera view showing ribs and Alexis

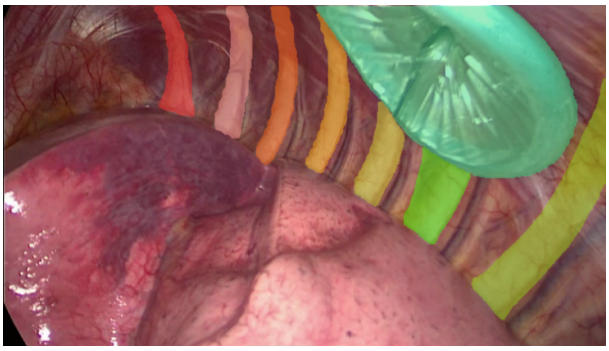

Fig. 2. Sample labeled image: Rib2 (Red), Rib3 (Pink), Rib4 (Orange), Rib5 (Mustard), Rib6 (Yellow), Rib7 (Green), Rib8 (Moss), Alexis (Emerald)

model training. Hence, data augmentation is used to introduce these permutations to the model by artificially generating more images with the flaws so that the model can be trained accordingly. Numerous data augmentation techniques were employed to address this. Firstly, rotation by ± 15 degrees was used to help the model become more resilient to camera rolls. Secondly, flip in both the horizontal and vertical direction was used to help the model become more insensitive to different camera orientations. Thirdly, a $\pm 20\%$ brightness adjustment was implemented to ensure that the model is more resilient to different lighting situations caused by the camera or occluding organs. Lastly, a Gaussian blur of 2.5 pixels was added to improve the model's resilience to camera focus. Additionally, to address the limitations of the small dataset, employing data augmentation effectively increased the dataset, tripling it from 222 images to 615 images. The final augmented dataset was broken into training, validation, and testing datasets where the percentage break down was 80%, 10%, and 10% respectively. Sample augmented images are presented in Fig. 3.

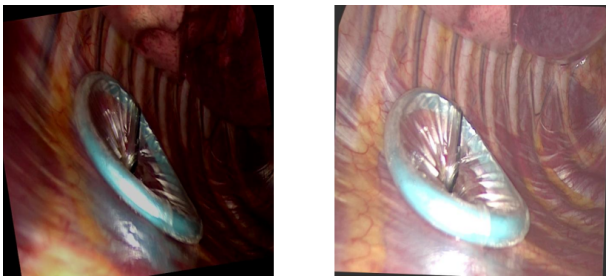

Fig. 3. Data augmentation samples

B. Model Architecture

YOLOv8, released by Ultralytics, surpasses its predecessor YOLOv5 in object detection, instance segmentation, and classification [11]. YOLOv8 [11] boasts significant architectural and developer-focused improvements. These advancements translate to enhanced segmentation and classification performance, flexibility, and efficiency for the model. Fig. 4 provides an overview of the YOLOv8-seg architecture.

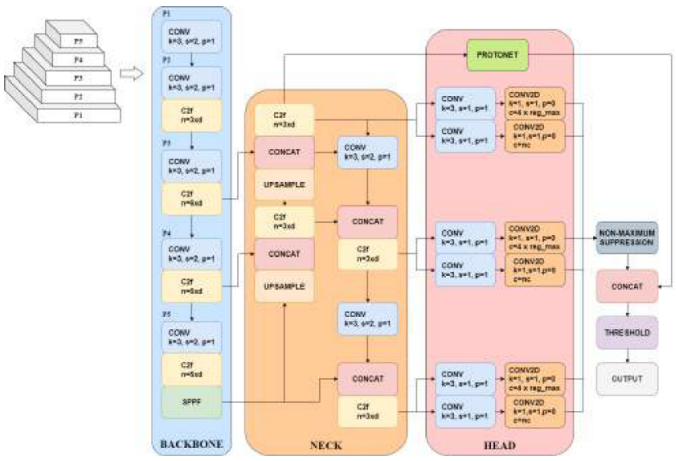

Fig. 4. YOLOv8-seg architecture visualization created based on diagram drawn by GitHub user RangeKing [12] and YOLACT [13]

According to its documentation [14], YOLOv8 builds on a standard convolutional neural network architecture with two main components: a feature extraction backbone and a prediction head. The YOLOv8 backbone is a modified version of the CSP-Darknet53 [15] architecture used in YOLOv5 where the cross stage partial (CSP) modules, referred to as C3, are replaced with C2f modules. The C2f module utilizes the number of features, a CBS (Conv, BatchNorm, SiLU) block, and concatenations to concatenate all the outputs from all bottleneck layers. In contrast, the C3 block only uses the output from the final bottleneck layer. In this instance, a bottleneck module is a sequence of two 3×3 convolutions with residual connections. YOLOv8's prediction head adopts a decoupled head approach. By separating classification and bounding box regression tasks, it allows for more focused optimization of each task. This streamlined design simplifies the model architecture, reducing computational complexity and improving inference speed without compromising detection performance. Additional improvements made include the use of anchor-free detection to improve generalization, eliminating the need for predefined anchor boxes. With anchor-free detection, the model directly predicts an object's center instead of the offset from a known anchor box, ultimately reducing the number of bounding box predictions. This makes the model more robust and adaptable to various object sizes and shapes. This also speeds up non-max suppression, a post processing step that filters through potential predictions after inference. Other notable improvements include YOLOv8's easy implementation through CLI (Command Line Interface) or Python IDEs, the

use of Yet Another Markup Language (YAML) files to define the model configurations, and using mosaic data augmentation that mixes four images to force the model to learn objects in different locations, occlusions, and surrounding pixels.

This paper utilizes YOLOv8-seg, a variant of the YOLOv8 model specifically designed for instance segmentation tasks [14]. While its architectural foundation shares similarities with the general YOLOv8 model discussed previously, YOLOv8-seg incorporates a couple key modifications. These include the addition of an output module in the head for generating mask coefficients and an additional ProtoNet module comprised of Fully Connected Network (FCN) layers to output the masks. These additions draw inspiration from principles established in the YOLACT model [13] for instance segmentation. The YOLACT influence is built in Fig. 4 by incorporating the ProtoNet and modified output process.

The YOLACT [13] architecture performs instance segmentation with a three-step approach. Firstly, it relies on a standard pre-trained CNN backbone network like ResNet to extract high-level features from the input image. A Feature Pyramid Network (FPN) then processes the feature maps, combining the features at different resolutions. This ensures that the model captures both fine-grained details and broader semantic information crucial for segmentation. The feature pyramid is then used as the input for the ProtoNet module. The ProtoNet module is implemented as a Fully Connected Network (FCN) and conists of two main components: Prototype Masks and Mask Coefficients Predictor. For the prototype masks, the ProtoNet module learns a set of pre-defined prototype masks which represent generic shapes. The Mask Coefficient Predictor then predicts a set of coefficients for each prototype mask at each location in the feature pyramid. The coefficients indicate how much each prototype mask contributes to the final segmentation mask for that specific location. Finally, the coefficients are multiplied element-wise with their corresponding prototype masks and summed together to produce the final instance segmentation mask.

C. Training

The YOLOv8-seg model was trained on a Windows operating system using Python 3.10.0 and PyTorch 2.2.0 with CUDA 11.8. The hardware used to execute the training consisted of an Intel(R) Core(TM) i7-8550U CPU @ 1.80 GHz and a NVIDIA GeForce RTX 2070 Super GPU with 8GB VRAM. For best performance, the model was trained for 200 epochs with a batch size of 4 and default parameters. Specifically, to optimize the learning process, Stochastic Gradient Descent (SGD) was employed with a fixed learning rate of 0.01. Additionally, a momentum factor of 0.937 was used to influence the direction of weight updates, and a weight decay of 0.0005 was included to prevent overfitting by penalizing large weights. The box loss ('box'), classification loss ('cls'), and distribute focal loss ('dfl') were assigned weights of 7.5, 0.5, and 1.5 respectively. Patience value of 50 was used to avoid early stopping. The addition of close mosaic was set to 10 which disables mosaic data augmentation in the last 10 epochs to stabilize training before completion. Finally, an Intersection over Union (IoU) threshold of 0.7 was used to determine if a prediction is considered correct.

D. Loss Function

Machine learning models rely on loss functions to quantify the discrepancy between their predictions and the actual targets (labels). During training, the model minimizes these loss functions by adjusting its internal parameters (weights and biases), ultimately leading to improved prediction accuracy. Developers of YOLOv8 employ multiple loss functions for its various parts [16]. For example, the branch responsible for classifying objects uses binary cross-entropy (BCE) loss function, as shown in Eq. 1. Here, w represents the weight, y_n is the label value, and x_n is the predicted value.

$$BCE = -w[y_n \log(x_n) + (1 - y_n) \log(1 - x_n)] \quad (1)$$

For the regression branch, which is responsible for bounding box predictions, a combination of two losses was used: distribute focal loss (DFL) and Complete Intersection over Union (CIoU) loss. DFL is applied to improve the model's ability to assess less predictable objects by broadening the probability distribution around the object's position. Its equation is expressed in Eq. 2 where S_n and S_{n+1} are given in Eq. 3.

$$DFL_{(S_n, S_{n+1})} = -[(y_{n+1} - y) \log(S_n) + (y - y_n) \log(S_{n+1})] \quad (2)$$

$$S_n = \frac{y_{n+1} - y}{y_{n+1} - y_n}, S_{n+1} = \frac{y - y_n}{y_{n+1} - y_n} \quad (3)$$

CIoU is similar to the distance IoU loss but introduces an influential factor which considers the aspect ratio of both the prediction and ground truth bounding box. Its equation is expressed in Eq. 4 where v represents the parameter that measures the consistency of the aspect ratio. Lastly, the final loss for YOLOv8 is a weighted sum of the aforementioned three individual losses.

$$CIoU = 1 - IoU + \frac{Distance_2^2}{Distance_C^2} + \frac{v^2}{(1 - IoU + v)} \quad (4)$$

E. Metrics

This paper evaluates the performance of the model in segmenting ribs and Alexis using three metrics. The first metric used was Precision. Precision is a statistic measuring the proportion of true positive predictions among all positive predictions. The formula for precision is shown in Eq. 5.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Here, TP refers to true positives, FP refers to false positives, and FN refers to false negatives. The second metric used was Recall, also known as sensitivity. Recall measures the proportion of true positives that are correctly identified out of

all the actual positive predictions. The formula for recall is shown in Eq. 6.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

The last metric used was Mean Average Precision (mAP) measured at two different IoU thresholds: 0.5 (mAP50) and a range of 0.5 to 0.95 (mAP50-95). IoU reflects the overlap between predicted and ground truth bounding boxes. By considering both precision and recall, mAP provides a singular score that evaluates the model's ability to accurately find relevant objects while minimizing false positives. It achieves this by averaging the precision obtained at various IoU thresholds across all object classes. The formula for mAP is shown in Eq. 7.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{7}$$

III. Experiments and results

The YOLOv8-seg network has numerous pre-trained models of varying sizes labelled as n, s, m, l, and x. Each distinct model varies based on size, primarily the channel depth and filter numbers, resulting in trade-offs between accuracy and processing speed. The YOLOv8-seg model that was tested was the medium sized pre-trained model called YOLOv8m-seg with 27.3M parameters and 110.2B floating-point operations (FLOPs).

To assess the impact of data augmentation, we compared the performance of two models: one trained on the original dataset and another trained on the augmented dataset. Both models were trained for 10 epochs. A comparison was made in Table I between the two model's performance.

TABLE I
RESULTS RECORDED FROM THE MODEL WITH DIFFERENT DATA AUGMENTATION TECHNIQUES

		Metrics			
Augment.	**Object**	**Precision**	**Recall**	**mAP50**	**mAP50-95**
None	Overall	87.4	82.8	88.7	59.1
	Rib 2	91.6	84.1	92.7	53.8
	Rib 3	70.0	71.4	80.2	45.3
	Rib 4	80.5	65.9	75.6	34.8
	Rib 5	84.9	70.4	78.4	53.6
	Rib 6	84.8	78.3	84.9	57.6
	Rib 7	92.8	94.7	98.8	66.7
	Rib 8	95.2	100	99.5	68.2
	Alexis	99.1	97.4	99.4	93.0
Flip	Overall	87.8	87.2	91.7	60.6
Rotate	Rib 2	94.6	97.3	99.2	55.4
Brightness	Rib 3	86.4	78.6	91.4	60.2
Gaussian Blur	Rib 4	86.0	84.0	91.6	59.6
	Rib 5	83.3	71.3	81.6	52.5
	Rib 6	77.9	76.5	80.6	42.4
	Rib 7	87.8	100	99.5	58.4
	Rib 8	88.1	90.0	89.9	60.6
	Alexis	98.2	100	99.5	96.1

Data augmentation evidently benefits the model's robustness for surgical images and videos. As shown in Table I, the vast majority of all metrics for each object class exhibited improvements. This suggests that the broader range of data provided by augmentation enhances the model's ability to handle the inherent variability seen within surgical images and videos. Thus, the augmented dataset was used to train the model. To illustrate the model's performance qualitatively, Fig. 5 showcases successful segmentation examples from the test images.

Referring to Fig. 5, it can be seen that the predicted labels accurately highlight the correct regions of interest, indicating the model's effectiveness for these specific cases. In particular, the last two samples were segmented correctly despite the surgical instrument occluding portions of the ribs. However, Fig. 6 presents instances where predictions failed. These failures can be attributed to factors like the presence of large amounts of adipose tissue in patients with a higher body mass index (BMI) or significant occlusions obscuring major portions of the ribs. This is inevitable due to the abundant variability observed in surgical scenarios. Therefore, accurate and detailed annotations during dataset creation are crucial. By highlighting extensive adipose tissue around the ribs and potential occlusions in the images, we can substantially decrease the failed predictions.

The final quantitative results are recorded and displayed in Table II. After 200 epochs, the developed model was able to achieve high scores across all object classes for all the metrics used. Achieving an overall precision of 94.6%, a recall of 95.2%, a mAP50 of 95.0%, and a mAP50-95 of 71.5%. Alexis segmentation achieved the best performance with a precision of 98.7%, a perfect recall of 100%, a mAP50 or 99.5%, and a mAP50-95 of 94.3%. This is likely attributed to the Alexis's unique shape and color, making them visually distinct from all other elements in the scene. Rib6 segmentation yielded the lowest performance metrics with a precision of 88.1%, a recall of 94.1%, a mAP50 of 86.3%, and a mAP50-95 of 53.1%. This may stem from the vast variations in Rib6's shape and size throughout the different surgical videos.

TABLE II
FINAL QUANTITATIVE RESULTS OBTAINED

	Metrics			
Object	**Precision**	**Recall**	**mAP50**	**mAP50-95**
Overall	94.6	95.2	95.0	71.5
Rib 2	98.7	100	99.5	72.8
Rib 3	95.6	92.9	93.3	69.3
Rib 4	95.9	94.7	96.0	75.5
Rib 5	100	96.2	99.5	72.7
Rib 6	88.1	94.1	86.3	53.1
Rib 7	91.8	93.8	91.7	62.3
Rib 8	88.2	90.0	94.0	72.2
Alexis	98.7	100	99.5	94.3

IV. Conclusion

In conclusion, this paper presents a novel rib and Alexis dataset, enabling the utilization of YOLOv8-seg for automatic segmentation of relevant anatomical structures in surgical videos. By leveraging data augmentation techniques, the

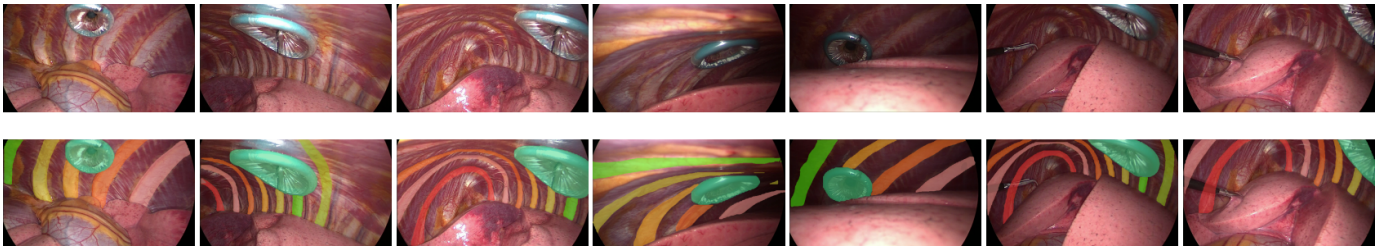

Fig. 5. Sample test images displayed on top row and corresponding resultant predictions displayed on the bottom row

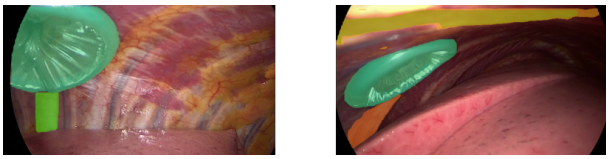

Fig. 6. Sample failed predictions due to adipose tissue and occlusions

model achieved a substantial boost in all the performance metrics. Specifically rotation, flip, brightness, and Gaussian blur. Qualitative evaluation demonstrated promising results, with occasional errors observed in a limited number of cases. These errors primarily stemmed from the presence of adipose tissue or vast occlusions, resulting in small portions of the ribs being visible. Accurate and detailed annotations in the data creation process can drastically reduce these errors. Additional training samples with the aforementioned scenarios can also be integrated to further improve the model's robustness. Quantitatively, the model achieved impressive results, demonstrating high precision, recall, mAP50, and mAP50-95 for rib segmentation (Rib2 through Rib8) and Alexis segmentation. The low performance metrics recorded for ribs like Rib6 likely stem from the variance in rib shape and size caused by adipose tissue and occlusion. Hence, should also be resolvable through enriching the training dataset and improving annotations. Following the successful segmentation of the relevant features, cross-modality translation can now be performed through registration, enabling feature and camera localization in the CBCT and potentially open doors for future applications.

V. Acknowledgments

This project has also been supported by the French National Research Agency (ANR) as part of the VATSop project (ANR-20-CE19-0015).

References

[1] "American cancer society. lung cancer." Mar. 2024. [Online]. Available: https://www.cancer.org/cancer/types/lung-cancer.html.

[2] Y. Kadomatsu, M. Nakao, H. Ueno, S. Nakamura, and T. F. Chen-Yoshikawa, "A novel system applying artificial intelligence in the identification of air leak sites," *JTCVS Techniques*, vol. 15, pp. 181–191, Oct. 2022.

[3] G.-A. Bilodeau, Y. Shu, and F. Cheriet, "Multistage graph-based segmentation of thoracoscopic images," *Computerized Medical Imaging and Graphics*, vol. 30, no. 8, pp. 437–446, Dec. 2006.

[4] Y. Bamba, S. Ogawa, M. Itabashi, H. Shindo, S. Kameoka, T. Okamoto, and M. Yamamoto, "Object and anatomical feature recognition in surgical video images based on a convolutional neural network," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, no. 11, pp. 2045–2054, Jun. 2021.

[5] M. Ivantsits, L. Tautz, S. Sündermann, I. Wamala, J. Kempfert, T. Kuehne, V. Falk, and A. Hennemuth, "DL-based segmentation of endoscopic scenes for mitral valve repair," *Current Directions in Biomedical Engineering*, vol. 6, no. 1, May 2020.

[6] G. Sánchez-Brizuela, F.-J. Santos-Criado, D. Sanz-Gobernado, E. de la Fuente-López, J.-C. Fraile, J. Pérez-Turiel, and A. Cisnal, "Gauze detection and segmentation in minimally invasive surgery video using convolutional neural networks," *Sensors*, vol. 22, no. 14, p. 5180, Jul. 2022.

[7] L. Tanzi, P. Piazzolla, F. Porpiglia, and E. Vezzetti, "Real-time deep learning semantic segmentation during intra-operative surgery for 3D augmented reality assistance," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, no. 9, pp. 1435–1445, Jun. 2021.

[8] P. M. Scheikl, S. Laschewski, A. Kisilenko, T. Davitashvili, B. Müller, M. Capek, B. P. Müller-Stich, M. Wagner, and F. Mathis-Ullrich, "Deep learning for semantic segmentation of organs and tissues in laparoscopic surgery," *Current Directions in Biomedical Engineering*, vol. 6, no. 1, May 2020.

[9] B. Noblet, M. Chabanas, S. Rouzé, and S. Voros, "Registration of 2D monocular endoscopy to 3D CBCT for video-assisted thoracoscopic surgery," in *SPIE Medical Imaging 2023: Image-Guided Procedures, Robotic Interventions, and Modeling*, San Diego, Apr. 2023.

[10] B. Chang, W. D. Tucker, and B. Burns., *Thoracotomy*. StatPearls, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK557600

[11] G. Joche, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023.

[12] "Rangeking Ultralytics. Brief summary of yolov8 model structure." Mar. 2023. [Online]. Available: https://docs.ultralytics.com/

[13] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: real-time instance segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.

[14] "Ultralytics yolov8 docs." [Online]. Available: https://docs.ultralytics.com/

[15] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: a new backbone that can enhance learning capability of CNN," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2020.

[16] A. Sahafi, A. Koulaouzidis, and M. Lalinia, "Polypoid lesion segmentation using YOLO-V8 network in wireless video capsule endoscopy images," *Diagnostics*, vol. 14, no. 5, p. 474, Feb. 2024.